On the Empty Miles of Ride-Sourcing Services: Theory, Observation and Countermeasures

by

Zhengtian Xu

A dissertation submitted in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
(Civil Engineering)
in the University of Michigan
2020

Doctoral Committee:

Professor Yafeng Yin, Chair
Professor Xiuli Chao
Assistant Professor Neda Masoud
Professor Romesh Saigal

Zhengtian Xu

ORCID iD: 0000-0001-5626-285X

To my family

ACKNOWLEDGMENTS

First and foremost, I am sincerely indebted to my advisor, Professor Yafeng Yin, for his continued education, inspiration, and encouragement. It would be impossible for me to accomplish what has been achieved so far without his attentive mentorship. I can still recall the email he sent when admitting me as a doctoral student, where he hoped/committed that "we work on something interesting together". The many years afterward have seen the persistent fulfillment of such a commitment as we initiated and expanded the research on ride-sourcing services. It was quite a journey but full of excitement and enthusiasm. I enjoyed those moments when some troublesome math eventually got worked out, felt thrilled when finding simple relationships that fitted "magically" well on some complex systems, and rejoiced when my tricks led complicated algorithms to convergence, etc. Despite the inevitable obstacles during the process, I am fortunate to have experienced all these sweet moments working with Dr. Yin, which drive me to keep going, exploring, and accomplishing. I am grateful to Dr. Yin for endowing me the maximum freedom in research pursues while offering the timely and pivotal discussions and suggestions as I need. His wisdom, expertise, and professionalism in teaching, mentoring, and research set up a lifetime model for me in the future career path.

I would also like to thank my committee members, Professors Neda Masoud, Xiuli Chao, and Romesh Saigal, for their accommodating guidance and constructive suggestions throughout my doctoral study. Particular thanks go to Dr. Masoud for always believing in me and my research, and Dr. Chao for the lasting but highly rewarding discussions on queueing models. Their experiences and references helped me appreciably to establish confidence in my academic job search.

I am deeply indebted to Professors Jieping Ye and Hongtu Zhu at DiDi AI Labs for their generous endorsement of my research by making available the abundant resources from Didi Chuxing. The scope and quality of this dissertation are vastly strengthened with the comprehensive empirical studies conducted using real-world data from DiDi. I would like to thank Ms. Lulu Zhang, Drs. Pinghua Gong and Shikai Luo, as well as their team members for their professional and considerate support on empirical data preparation and model calibrations. My unique and enduring experience with DiDi is indeed more than a journey. Through the interactions, I was able to experience the working styles in the industry, work directly on real-world problems, and understand better the gaps between the theories on books and the implementations in reality. These cherishable memories continuously shape my research objectives and philosophies, and will undoubtedly become valuable assets for my future career development.

I want to extend my thanks to all my dear friends both in the U.S. and China, including but not limited to Zhibin Chen, Liteng Zha, Xiaotong Sun, Eleftheria Kontou, Nima Shirmohammadi, Daniel Vignon, Alex Sundt, Mojtaba Abdolmaleki, and Qi Luo at the Lab for Innovative Mobility Systems; Shenyang Chen, Yinan Zheng, Ruizhi Zou, Shanyue Guan, Like Liu, Zilin Bian, Wei Sun, and Hao Jun at the University of Florida; Yan Zhao, Xintao Yan, Xinmin Wang, Wai Wong, Shuo Feng, Shengyin Shen, Zhen Yang, Shihong Huang, Yiyang Wang, and Shurui Zhang at the University of Michigan; as well as Sen Wang, Jiabin Chen, Jintao Ke, Kaixiang Lin, and my three

housemates Qiang Wang, Shiba Dandpat, and Mingjie Bi. I am incredibly grateful to have them by my side to share my laughter and tears and to enrich my life with the many joys other than research.

Last but not least, I owe my deep gratitude to my parents, Qingwei Xu and Dan Wang, for their unswerving accompany and support. Even though being overseas, they are always the firm backing that I can freely turn to for the ups and downs. Thank them for always holding faith in me, respecting my decisions, and rooting for my growth. I can never imagine how challenging it would be to reach where I am without their selfless love.

TABLE OF CONTENTS

LIST OF FIGURES

Figure

LIST OF TABLES

Table

LIST OF APPENDICES

ABSTRACT

The proliferation of smartphones in recent years has catalyzed the rapid growth of ride-sourcing services such as Uber, Lyft, and Didi Chuxing. Such on-demand e-hailing services significantly reduce the meeting frictions between drivers and riders and provide the platform with unprecedented flexibility and challenges in system management. A big issue that arises with service expansion is the empty miles produced by ride-sourcing vehicles. To overcome the physical and temporal frictions that separate drivers from customers and effectively reposition themselves towards desired destinations, ride-sourcing vehicles generate a significant number of vacant trips. These empty miles traveled result in inefficient use of the available fleet and increase traffic demand, posing substantial impacts on system operations. To tackle the issues, my dissertation is dedicated to deepening our understanding of the formation and the externalities of empty miles, and then proposing countermeasures to bolster system performance.

There are two essential and interdependent contributors to empty miles generated by ride-sourcing vehicles: cruising in search of customers and deadheading to pick them up, which are markedly dictated by forces from riders, drivers, the platform, and policies imposed by regulators. In this dissertation, we structure our study of this complex process along three primary axes, respectively centered on the strategies of a platform, the behaviors of drivers, and the concerns of government agencies. In each axis, theoretical models are established to help understand the underlying physics and identify the trade-offs and potential issues that drive behind the empty miles. Massive data from Didi Chuxing, a dominant ride-sourcing company in China, are leveraged to evidence the presence of matters discussed in reality. Countermeasures are then investigated to strengthen management upon the empty miles, balance the interests of different stakeholders, and

improve the system performance. Although this dissertation scopes out ride-sourcing services, the models, analyses, and solutions can be readily adapted to address related issues in other types of shared-use mobility services.

CHAPTER 1

Introduction

1.1 Motivation

Recent advancements in information and vehicular technologies drive an unprecedented wave of innovations in mobility services. Specifically, the number of smart mobile devices in the U.S. has been rising steadily, and studies suggest that over eighty percent of Americans now own at least one such device (e.g. Anderson, 2019). These devices retrieve users' geolocations, enable ubiquitous communications, and allow instant peer-to-peer interaction, giving rise to various on-demand mobility services for goods and people, which bring together suppliers of resources and services and their customers with very low transaction costs (Greer et al., 2018). As a typical example of emerging mobility services, ride sourcing has attracted much attention and achieved substantial growth in recent years. In particular, between the two largest ride-sourcing companies (also known as transportation network companies or TNCs), Uber has cumulatively served over ten billion customer trips and extended its business to 65 countries on six continents, while Didi Chuxing provides 30 million daily rides and has covered more than 400 cities globally by May 2019 (Iqbal, 2019). These ride-sourcing services are expected to play an increasingly important role in meeting the travel needs in urban areas. Looking into the foreseeable future, we envision that the connected and automated vehicle (CAV) technologies will further favor the shared-use mobility and substantially reshape the mode structure of urban transportation.

The great success of ride-sourcing services continuously generates heated discussions and has aroused extensive research interest on their impacts, operations and regulation. Previous studies have focused on the demographics of ride-sourcing riders and drivers (Rayle et al., 2016; Hughes and MacKenzie, 2016; Kooti et al., 2017; Hall and Krueger, 2018), economic efficiency and equity characteristics of the service (Smart et al., 2015; Ge et al., 2016; Hahn and Metcalfe, 2017), pricing strategies of the platform (Chen and Sheldon, 2016; Banerjee et al., 2016; Taylor, 2018; Cachon et al., 2017; Zha et al., 2016, 2018b; He et al., 2018; Castillo et al., 2018; Bimpikis et al., 2019), impacts on the traditional taxi market (Wallsten, 2015; Cramer and Krueger, 2016; Rayle et al.,

2016; Nie, 2017; Lam and Liu, 2017), as well as policies and regulations (Koopman et al., 2014; Ranchordás, 2015; Zha et al., 2016, 2018a,b; Wang et al., 2018). While there are fewer doubts on the comparative efficiency of ride-sourcing service, controversies have repeatedly arisen for its competition with regulated taxi service, surge pricing by the platform as well as equity issues in matching drivers and riders. Recently, more and more attentions from both the public and academic sectors have been drawn to the deteriorating traffic conditions in major cities partially due to the expansion of ride-sourcing markets (Schaller, 2018; Castiglione et al., 2018; Erhardt et al., 2019).

In fact, other than serving induced travel demand and substituting for public transit, biking, and walking, ride-sourcing services generate massive empty miles (also referred as out-of-service/ vacant/ unproductive/ dead/ ghost miles or trips by other literatures) in addition to the trips with customers on board. These empty miles originate from the end of one customer trip to the start of the next, conquering the meeting frictions between drivers and customers. It is estimated that the amount of empty miles comprise 20% of total mileage traveled by ride-sourcing vehicles (RVs) in San Francisco (Castiglione et al., 2017) and 45% in New York City (Schaller, 2017). They not only occupy the precious labor supply, but also contribute additional traffic demand and continuously scale up as the adoption rises for ride-sourcing services. Thus, there is a pressing need of a profound understanding on the production of empty miles as well as its externalities to the urban mobility systems to facilitate management and regulations.

Centering on the empty miles of RVs, this dissertation aims to understand the behavior of ride-sourcing systems and then put forward countermeasures to help enhance and sustain the system efficiency. As a typical two-sided market, the performance of a ride-sourcing system is largely dependent on the interactions between the four parties – riders, drivers, the platform and sometimes the public-sector regulators. Riders as a group are substantially abstracted in this work, as their participation is relatively short-lived following regular demand profiles. We thus concentrate on the rest three participants by looking into their intricate forces driving behind the empty miles. To facilitate understanding, the interpretation separates by parties into three axes, respectively associated with platform strategies, driver behaviors, and regulatory concerns. In each axis, theoretical models are established to help reveal the critical trade-offs underlying the empty miles, and identify the weaknesses in real-world operations that may undermine the system performance. Comprehensive data from Didi Chuxing are leveraged to evidence the presence and magnitude of issues discovered. Strategic remedies are then proposed to strengthen controls over the empty miles, balance the needs of different parties, and optimize the system efficiency in various management circumstances.

1.2 Challenge

Our investigations on the empty miles of ride-sourcing services manifest the following challenges:

- *Intricate interdependencies*. Ride-sourcing vehicles in service switch between three stages of movements, i.e., either cruising in search of riders, deadheading to pick them up, or delivering them to their destinations. The cruising stage reflects drivers' choice behaviors, while the platform's matching strategies dictate the latter two stages; Notwithstanding, the three phases are closely inter-related. For example, the cruising and matching of drivers can influence the distribution of supply availability over the space, giving rise to changes in pickups. Conversely, by capping the matching distance between riders and drivers, it impacts not only the deadheading phase but also the chances of riders and drivers getting matched, which further yield variations to delivering and cruising trips. In-depth knowledge of mechanisms behind the empty miles thus requires skillful disentanglement of the complexities in dependencies.

- *Changeable environment*. On the one hand, the number of trip requests fluctuates considerably in both time and space, casting great challenges for demand forecasting. On the other hand, unlike traditional firm-employee relationship, the ride-sourcing market is featured by freelance drivers, who are free to determine where, when and how long they want to work. Such fluctuations in demand and the flexibility enjoyed by ride-sourcing drivers create highly changeable market condition, which troubles the platform's management and adds up to our difficulties for modeling and empirical analyses. As models are more or less simplified, the question worthy of raising is to what extent the simplifications are valid. To accommodate the spatial and temporal variability, proper modeling granularity is needed to stay consistent with the context of discussions and to grasp the crux of major concerns.

- *Conflicts of interests*. Although we frequently claim to "optimize the system performance", there seldom is such a well-defined measure off-the-shelf to work with. In fact, each major players hold their own list of concerns. For instance, riders use the service typically by its affordability and accessibility, while drivers who also favor the matching efficiency probably care more about their profitability. The platform may be eager to expand the service; meanwhile, the government pays more attention to the congestion effects brought with the service growth. As we notice, the demands of different parties are not necessarily well-aligned and often contradicting with each other, under which favoring one party may leave behind the other and eventually jeopardize the operations. Resolving the conflicts of interests thus requires a thorough understanding of the operational mechanisms in a ride-sourcing system along with techniques to balance between multidimensional trade-offs.

By virtue of methodologies from extensive disciplines, this dissertation tackles the above challenges and methodically examine the key properties and issues upon the empty miles. It scopes from the microscopic fundamentals of individual driver's choice behaviors to the macroscopic mechanisms behind the system operations, and integrates an abundant series of analytical models, empirical validations and practical countermeasures.

1.3 Contribution and Dissertation Outline

To disentangle the complexities and facilitate understanding, our discussions of empty miles are structured into three major axes, respectively laying over the perspectives of a platform, drivers, and the government as a system regulator. We highlight the contributions made by this dissertation via an exhaustive summary below.

The first axis focuses on the impacts of a platform's matching strategies on the empty miles. Within the matching process, once a vehicle become available, the platform gets to decide when and to whom the driver will be matched. The pace of matching primarily determines how long drivers shall wait in matching and how far drivers should travel to pick riders up. In addition, by pairing drivers to requests heading for different destinations, the platform essentially redistribute the supply over time and space. Following these considerations, Chapter 2 and 3 investigate the macroscopic trade-offs that the platforms encountered in strategic matching.

Specifically, the two chapters study the supply curve of ride-hailing systems under different market conditions. The curve defines a relationship between the throughput of trips of the system and the cost of riders it serves. By assuming isotropic market conditions, Chapter 2 first revisit a matching failure, named "Wild Goose Chases", identified recently that matches a requesting rider with an idle driver very far away. The failure will cause the supply curve of an e-hailing market to be backward bending, but it is proved that the backward bend does not arise in the street-hailing market. By constructing a double-ended queuing model, we prove that the supply curve of an e-hailing system with finite matching radius is always backward bending, but a smaller matching radius leads to a weaker bend. We further reveal the possibility of completely avoiding the bend by adaptively adjusting the matching radius. Chapter 3 then turns to the anisotropic markets and identify another type of matching failure, named as "Countryside Delivery", due to indiscriminate matching between drivers and riders, which again causes a backward bending supply. Given the prevalence of such a matching failure in real-world operations, we discuss how to avoid it using price or matching discrimination over customer demand. A generalized fluid-based queueing model is developed for quick and effective state evaluations of realistic, large-scale ride-hailing systems. An empirical study is conducted to showcase the capability of spatial pricing in expanding the system throughput and reducing the supply wastes in empty miles.

The next axis starts with drivers' cruising behaviors in service provision. Given the spatial variations in market conditions, drivers constantly make short-term decisions of where and how to search for customers when their cars become vacant. Spurring predominantly by drivers' behaviors, however, the cruising movements may also be impacted by interventions from the platform. For example, the platform can provide heatmaps, customized messages, or even incentives to drivers to assist their customer search. Therefore, the seemingly purposeless idle movements can indeed embody rich considerations and management. To better serve the purpose, we in this axis distinguish two categories of search behaviors that widely exist in ride-sourcing markets. The first category happens in urban areas with service opportunities scattered continuously over the space, where drivers after becoming idle usually zigzag in the neighborhood or simply stay motionless. We call this category of behaviors as "local search" and study it in Chapter 4. In contrasts, there is another type of service areas islanded from the rest space carrying condensed demand of rides, such as airports, train stations, and stadiums etc. For drivers either head towards or abandon those areas, they may experience long distances of cruising. Such far-distance repositioning of RVs is categorized as "zonal choice", implying drivers' selections of their favorite service regions, which is examined in Chapter 5. These two categories of behaviors are sometimes tangled together and cannot be easily separated.

Our investigations in both chapters are initiated by the calibration of structural models, to enlighten the effects such as the profitability and transitional costs in affecting drivers' local search and zonal choices. In Chapter 4, we materialize drivers' local search process as sequential Markovian choices, and model it using a dynamic discrete choice model proposed to rationalize the time-dependent search movements and impact factors. Specifically, the model differentiates unobservable intentions behind drivers' search movements: remaining motionless, cruising around without target areas, or repositioning toward specific destinations. Calibration results uncover the time-varying and heterogeneous concerns of drivers in customer search. While Chapter 4 caters to drivers' local search behavior where particular attention is paid to the search directions in decision makings, Chapter 5 explores drivers' cross-regional choices at the service islands by focusing on the temporal dimension in decision-making. We apply survival models to study drivers' dwelling behaviors in those areas under random and sequential matching strategies, respectively. Results, insights as well as suggestions for less miscommunication and better management at the service islands are discussed.

The last axis addresses the government's concerns. Facing the rapidly growing empty miles, the government in response can act in the two ways. First, there is a necessity to accurately estimate the scope and scale of the influences from the empty miles produced by RVs, a missing ingredient in the current traffic assignment methodology. Chapter 6 thus proposes to explicitly model those vacant trips including cruising and deadheading. Due to the similarity between taxi and ride-

sourcing services, we firstly extend previous taxi network models to construct a base model, which assumes intra-zonal matching between customers and idle RVs and thus only considers cruising vacant trips. Considering spatial matching among multiple zones commonly practiced by ride-sourcing platforms, we further enhance the base model by encapsulating inter-zonal matching and considering the empty miles from both cruising and deadheading. The effectiveness of the inter-zonal matching framework is validated using large-scale empirical data from DiDi, and a solution algorithm is proposed to efficaciously solve the enhanced equilibrium model. The extended model describes the equilibrium state that results from the interactions between background regular traffic, and occupied, idle and deadheading RVs. Numerical examples are additionally presented to demonstrate the model and solution algorithm.

On top of the precise estimate, the government could seek countermeasures to mitigate the negative externalities caused by the empty miles. Chapter 7 at first briefly covers the commonly discussed traffic management approaches, such as congestion road pricing and strategic vehicle routing. Then, in opposition particularly to the congestion incurred by the empty miles, we investigate the allocation of a certain portion of road space to on-street parking for idle RVs. A macroscopic conceptual framework is developed to capture the trade-off between capacity loss and the reduction of cruising. Considering a hypothetical matching mechanism adopted by the platform, we further materialize the framework and then apply it to study the interactions between the ride-sourcing system and parking provision under various market structures.

In Chapter 8 we conclude the dissertation and point out possible directions for future research.

Platform Strategies

A ride-sourcing system itself can be characterized as a congestible system, where fleet resources are shared among users. For a given number of idle vehicles, newly arrival riders can shed negative/congestion externalities on other users already waiting online, causing them to wait for longer time. Meanwhile, an increasing number of users usually brings about positive network effects, thereby reducing the per capita service costs and benefiting the service provider. Thus, the prevailing goal of managing a congestible system is to balance between the negative congestion and positive network effects (Johari and Kumar, 2010). The supply of trips, defined as the number of trips a ride-sourcing system accommodates per unit of time, can manifestly signify the efficiency of a system and intimate the level of the service for both riders and drivers. Chapter 2 and 3 thus propose to comprehend and model the internal relationships that associate the supply of trips with other basic macro-measures of a ride-sourcing system. By virtue of their theoretical conciseness and practical descriptiveness, queueing models are prescribed as basic setups in modeling the system, where two spatial granularities are explicitly considered. A double-ended queueing model is constructed to mimic the system operations of matching riders and drivers in each local subregion, whereas a fluid queueing network is built up to investigate the global effects of spatial anisotropic among different subregions. These queueing models disentangle the complexities of interactions in space and states, and enable us to identify the issues that aggravate the supply congestion and drag down the system efficiency.

CHAPTER 2

"Wild Goose Chases"

Tragedy of Quick Commitment

2.1 Introduction

This chapter contributes to this ongoing quest by revisiting "wild goose chase (WGC)", a matching failure recently identified by Castillo et al. (2018) in the ride-sourcing market, which may lead to collapse of the system but can be solved by using surge pricing.

More specifically, WGC refers to a situation when an idle driver is matched with a requesting customer very far away and has to waste substantial time on picking him or her up. As Castillo et al. (2018) pointed out, WGCs arise when there are fewer idle vehicles than requesting customers; ride-hailing platforms are forced to match idle vehicles with distant customers, thereby yielding massive waste of effective labor supply on the deadheading trips to pick up customers. It is shown that given a fixed number of working drivers in the system, WGCs will cause the trip supply curve to be backward bending, similar to the "hypercongestion" phenomenon in highway traffic flow (see, e.g., Small and Chu, 2003). In this regime, the system is inefficient. Specifically, compared with the counterpart in the upward-sloping regime, to serve the same number of trips, more vacant vehicles are dedicated to deadheading trips without paying customers on board; longer waiting time implies a lower price (for maintaining the same level of generalized costs for customers), thereby reducing drivers' income (Zha et al., 2018b). If drivers exit the market in response to reduction in their earnings, the system may collapse.

In a broad sense, a ride-hailing system is an input/output system where riders "flow into the system, spend some time in it and then flow out" (Daganzo, 2007). Similar to many other examples given by Daganzo (2007), the output rate of a ride-hailing system can decline with the accumulation of riders. Consequently, control needs to be implemented to restore it. Given the significant implication of WGCs or more generally, the backward bend in the supply curve of a ride-sourcing system, this chapter is devoted to investigating how and when the backward bend arises and discussing strategies to avoid it or to maintain the system to operate at efficient states. A

better understanding of the supply curve will help service providers answer operational questions such as what is the capacity of the ride-hailing system, i.e., the maximum number of passengers the system can accommodate per unit of time? Is the system currently operating at an efficient or inefficient state? If inefficient, how to intervene to make it more efficient?

Castillo et al. (2018) demonstrated the existence of WGCs by considering a first-come-first-serve (FCFS) matching scheme in a homogeneous spatial region or a spatially isotropic market, and also used data from Uber to show that WGCs are present in the Manhattan market. The authors then suggested that WGCs can be avoided if a small matching radius is imposed. Rather, by constructing a double-ended queuing model, we show in this chapter that the supply curve of a ride-hailing system with finite matching radius is still backward bending. A smaller matching radius does lead to a weaker bend. In a spatially isotropic market, the bend can be avoided if the matching radius can be adjusted adaptively in a certain way. We further show that the backward bend also emerges in a spatially heterogeneous market where some zones enjoy excessive vehicle supply while others experience a shortage. After pointing out that zone-based surge pricing commonly implemented by ride-sourcing platforms would fail to address such a matching failure, we discuss various strategies to resolve the issue.

The remainder of this chapter is organized as follows. Section 2.2 firstly defines the supply curve of a ride-hailing market by specifying supply and cost of the service. Section 2.3 replicates the backward bending supply curve derived by Castillo et al. (2018) in an isotropic e-hailing market, and then employs a bilateral matching model to prove that similar bend will not be present in the street-hailing market. Subsequently, a double-ended queuing model is constructed in Section 2.4 to investigate the impact of controlling the matching radius on the shape of the supply curve and various system performance measures. A two-node network is developed to conceptualize the cause and reveal the resulting inefficiency. Potential solutions are then discussed to address the issue. Finally, Section 2.5 concludes this chapter and points out the future research directions.

2.2 Supply of Trips and Service Cost

To set the stage for our discussion, this section firstly specifies the supply and cost of a ride-hailing market. By treating rides as a service product, the system supply S under a given market condition is defined as the number of trips that a system can accommodate or serve per unit of time. To simplify the model derivation and conceptualize the system analyses, we assume in most parts of this chapter the market to be isotropic (see, e.g. Arnott and Inci, 2010; Xu et al., 2017). In these cases, without loss of generality, we refer to supply as the number of trips accommodated per unit of time and area, i.e., the flux of trips.

A rider who uses e-hailing services typically goes through three consecutive stages: sending

out a request of ride, receiving a match with an idle vehicle, and boarding the vehicle on its arrival. For each request, define w^c, w^m, and t respectively as the average matching time (the waiting time for being matched online), pickup time (the waiting time for the matched vehicle to show up) and delivery time (in-vehicle time), and F as the average fare. As traffic congestion is not of major concern in this study, we set t as a constant. Then, the cost of riders C can be described as the following function W_e, which accounts for both monetary and time costs associated with each stage, i.e.,

$$C = W_e(w^c, w^m, F) \tag{2.1}$$

Note that the partial derivative of W_e with respect to each independent variable is positive, representing customers' monotonic discomfort over the expense of time or money.

Different from e-hailing, street hailing does not yield deadheading or pickup trips. The cost function W_s is thus modified to be:

$$C = W_s(w^c, F) \tag{2.2}$$

where both derivatives $\frac{\partial W_s}{\partial w^c}$ and $\frac{\partial W_s}{\partial F}$ are again positive.

In the following sections, we derive supply curve, i.e., the relationship between supply S and cost C, frequently referred to as supply function $S(C)$ for convenience, for systems under different settings, to help reveal the efficiency property of a ride-hailing system. A complete notational glossary is provided in Appendix A.1 for reference.

2.3 "Wild Goose Chases" – Backward Bending Supply Curve

In this section, we firstly replicate the theoretical model by Castillo et al. (2018) to show the presence of a backward bending supply curve for e-hailing services in an isotropic market. For comparison, we then adopt the same setting and prove that the backward bend does not arise in traditional street-hailing services.

2.3.1 The model by Castillo et al. (2018) for e-hailing services

Vehicles operating in an e-hailing system can be at one of the following three states: idle (cruising to search for customers or waiting to be matched), deadheading (en route to pick up matched customers), or occupied (delivering customers to their destinations). Denote N and N^v respectively as the densities of all vehicles and those in the idle state. As this section assumes the market to be isotropic, we further specify Q as the trip flux. Assuming the fleet size is fixed, the

conservation gives rise to

$$N = N^v + w^m Q + tQ \tag{2.3}$$

where the three terms on the right hand side represent the density of vehicles at idle, deadheading and occupied states, respectively.

Castillo et al. (2018) considered a continuous-time matching process whereby a requesting customer will be matched right away to the closest idle vehicle on a sufficiently large plane where idle vehicles are distributed uniformly. Two assumptions are implicitly made: the matching is FCFS and perfect efficiency in the matching process (i.e. $w^c = 0$). These assumptions prescribe Eq. (2.1) and (2.3) to be of the following form:

$$\begin{aligned} C &= W_e(w^m) \\ N &= N^v(w^m) + w^m Q + tQ \end{aligned}$$

where F is assumed to be fixed and removed from the variable space of W_e; and N^v is now a convex and decreasing function[1] on w^m, with $\lim_{w^m \to 0} N^v(w^m) = +\infty$ and $\lim_{w^m \to +\infty} N^v(w^m) = 0$. Solving Q from the above functions on C yields the supply function of the e-hailing system:

$$S_e^{\infty}(C) = \frac{N - N^v(W_e^{-1}(C))}{W_e^{-1}(C) + t}$$

where the superscript symbol ∞ in S_e^{∞} labels that the function is derived without restricting the matching radius; $W_e^{-1}(C)$ is the inverse function of W_e and thus monotonically increases on C. Consequently, we obtain the proposition below (see Appendix A.1 in Castillo et al., 2018 for the proof):

Proposition 2.1. *Given $N > 0$, there exists a $\hat{C}$ such that $S_e^{\infty\prime}(C) > 0$ for $C < \hat{C}$ and $S_e^{\infty\prime}(C) < 0$ for $C > \hat{C}$.*

The solid line in Fig. 2.1 depicts such a supply curve for the e-hailing system. It can be seen that for the same level of trip supply, there are two instances: the one with the waiting cost below $\hat{C}$ is called a good outcome by Castillo et al. (2018), while the other with the customer waiting costs above $\hat{C}$ is named as WGC. Compared to the good state, WGC makes a substantial portion of vacant vehicles dedicated to deadheading trips, thereby diluting the density of idle vehicles and yielding an inefficient system performance.

[1] View it as an inverse function of w^m on N^v. As customers are matched sequentially without competition, higher density of idle vehicles yields shorter deadheading distance. Specifically, for a two-dimensional space, $w^m \propto (N^v)^{-\frac{1}{2}}$.

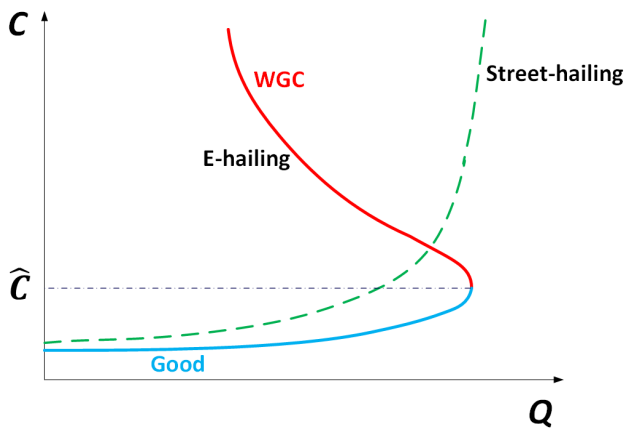


Figure 2.1: Supply curves of ride-hailing systems in isotropic market

To facilitate the understanding of the phenomenon, we further make an analogy between the supply curve herein and the fundamental diagram well known to traffic engineers. The process of matching riders and transporting them to their destinations in a ride-hailing system can be viewed as if riders traveled through "a unit-length highway segment". Then, the flux Q, density $N - N^v$ and speed $(w^m + t)^{-1}$ of riders getting served respectively resemble the flow, density and speed on the segment. Similar to the fundamental diagram in the traffic flow theory, the relationships between the flux Q and the other two correspondences are non-monotone. Given the fleet size of a ride-hailing system, the service speed $(w^m + t)^{-1}$ monotonically decreases with the density of riders accumulating in the system $N - N^v$, while the flux Q increases under low densities of accumulations and then decreases after reaching its maximum. Therefore, the backward-bending supply of an e-hailing system essentially mirrors the "congested" regime in a non-monotonic speed-flow diagram, whose presence hints the existence of inefficient system states.

As WGCs largely appear when a system experiences high customer demands[2], Castillo et al. (2018) thus suggested the implementation of surge pricing to suppress the excessive demand and sustain the system at the good states. While such a strategy appears beneficial, particularly to the platform and drivers, it tends to create controversy. The platform may be criticized by riders for price gouging, because there is "sufficient" supply around, while drivers can also blame the platform for causing their unreasonable idleness. A less controversial solution mentioned by Castillo et al. (2018) to avoid WGCs is for the platform to shrink the matching radius[3]. As we shall show in Section 2.4, the backward bend still exists with a fixed matching radius imposed.

[2]High customer demand refers to the situation where the demand curve stays far above the Q axis and thus likely intersects with the supply curve at WGC.

[3]The FCFS matching is implemented by Uber but with a capped matching radius, as mentioned by Castillo et al. (2018). Consequently, its aggregate effect may deviate from the one predicted by the theoretical model developed by Castillo et al. (2018).

2.3.2 The bilateral searching model for street-hailing service

We now turn to the traditional street-hailing taxi system, and prove that the backward bend will not arise in isotropic markets. As there are no deadheading trips in street hailing, the fleet conservation equation is then modified as

$$N = N^v + tQ \tag{2.4}$$

Combining Eq. (2.4) with Eq. (2.2) yields

$$S_s(C) = \frac{N - N^v(C)}{t}$$

Castillo et al. (2018) obtained the monotonicity by assuming another decreasing function of $\tilde{N}^v$ on w^c or $W_s^{-1}(C|F)$. Since $W_s^{-1}(\cdot)$ characterizes an increasing function, S_s is thus declared to be monotonically increasing on C. However, trivially specifying $\tilde{N}^v$ as a univariate function on w^c can be problematic, especially when the market accumulates a thick density of customers waiting to be served. The meeting of customers and drivers under street hailing is the outcome of bilateral search, under which the waiting time of customers will not only depend on the density of idle drivers but also the competition from other customers (Yang and Yang, 2011). The above treatment fails to capture the competition among customers in finding rides.

In what follows, we adopt a bilateral meeting function proposed by Yang and Yang (2011) to delineate the search frictions under street hailing, and prove the monotonicity of the resultant supply function. Let N^r denote the density of customers waiting/searching for rides. Then, A bilateral meeting function M can be presented as:

$$Q = M(N^v, N^r) \tag{2.5}$$

where $\frac{\partial M}{\partial N^v}, \frac{\partial M}{\partial N^r} > 0$. Denote α_v and α_r as the elasticity of Q with respect to N^v and N^r, respectively. In general, we have $0 < \alpha_v, \alpha_r \leqslant 1$. Integrating Eq. (2.5) with the general relationships (2.2) and (2.4) on street hailing, the supply function $S_s(C)$ solves Q from the following implicit function on C:

$$Q = M(N - tQ,\ QW_s^{-1}(C)) \tag{2.6}$$

Proposition 2.2. *The supply function $S_s(C)$ is monotonically increasing on C.*

Proof. By taking the derivative of both sides of Eq. (2.6) with respect to C, we obtain

$$\left(1+\alpha_v\frac{tQ}{N-tQ}-\alpha_r\right)\cdot\frac{\mathrm{d}Q}{\mathrm{d}C}=\alpha_r\cdot\frac{Q}{W_s^{-1}(C)}\cdot\frac{\mathrm{d}W_s^{-1}}{\mathrm{d}C} \tag{2.7}$$

Since $0<\alpha_v,\alpha_r\leqslant 1$ and $\frac{\mathrm{d}W_s^{-1}}{\mathrm{d}C}>0$, Eq. (2.7) suggests that $\frac{\mathrm{d}Q}{\mathrm{d}C}>0$ on all feasible (Q,C). Thus, the supply function $S_s(C)$ of a street-hailing system is monotonically increasing on C. □

Proposition 2.2 proves the monotonicity of the supply curve of a street-hailing system under the consideration of bilateral search and meeting. To facilitate the comparison, we present it graphically as the green dashed line in Fig. 2.1.

2.4 Supply Curves under Finite Matching Radius

Since WGC arises due to distant pickups, an intuitive solution is to restrict the matching distance. To our knowledge, ride-sourcing companies like Uber and Didi Chuxing typically cap matching radii in their matching algorithms to avoid devastating pickup distance. By virtue of a double-ended queuing model, we demonstrate in this section that the backward bend in the supply curve for isotropic e-hailing markets can be mitigated or even avoided by regulating the matching radius. It is shown that such regulation indeed thickens the accumulation of idle drivers, thereby improving the system efficiency via a thick market effect.

2.4.1 Double-ended queuing model for an isotropic market

This section confines to the scenarios where the backward bend in the supply curve arises. Specifically, we consider an isotropic space with sparse distributions of idle drivers, and the arrival rate of idle drivers is lower than that of incoming requests. In practice, a rider requesting a trip during supply shortage can often be held in the matching process for minutes before the platform finds a suitable match or the rider cancels her request. A practice by ride-sourcing platforms such as Didi Chuxing is to segment the space into bordered matching blocks; within each block, the platforms organize requests into a virtual queue and then match them in sequence with idle drivers arriving in the block. By virtue of its theoretical conciseness and practical descriptiveness, a double-ended queuing model (see, e.g., Matsushima and Kobayashi, 2006; Shi and Lian, 2016) is introduced below to assist us understanding the working mechanism of the e-hailing system under a finite matching radius.

Suppose an isotropic space is partitioned into bordered matching blocks, each with an area of A (i.e. the matching radius is roughly of the scale of $\sqrt{A}$). For simplicity, the matching process

within each block is abstracted as the platform matching requests of rides to idle drivers in a FCFS fashion:

- If a request arrives with idle drivers waiting in the block, it will get a match with the closest driver available;
- If there is no available driver nearby, requests will be held into the virtual queue and matched with arriving idle drivers in sequence;
- If the expected waiting time at the queue end exceeds a threshold, the newly arriving requests will abandon the queue[4].

We then propose a double-ended queue to deductively replicate the matching process within each block.

Denote λ^c and λ^t respectively as the Poisson arrival fluxes of requests and idle drivers. Let w_0^c and w_0^t respectively denote customers' and idle drivers' tolerance for the expected waiting time on matching. As previously mentioned, the newly arrival customers will abandon the queue when their tolerance is exceeded. On the driver side, they are not guaranteed to be matched in a FCFS fashion. Therefore, if drivers' expected waiting time in a block exceeds their tolerance, some drivers therein (not necessarily the newly entering ones) will cruise to other blocks, lowering the expected waiting time down to the tolerance.

The process within each block can thus be modeled as a double-ended queuing system, characterizing a one-dimensional birth-and-death process shown in Fig. 2.2. The positive and negative side respectively represents the states where requests or idle drivers queue in the block, with the number labeled indicating the corresponding queue length. The state space of the process Ω is thus given as

$$\Omega = \{-N^c, -N^c + 1, \cdots, -1, 0, 1, \cdots, N^t - 1, N^t\}$$

[4]In practice, a rider will be notified of the real-time queue length and her estimated queuing time before she sends out the request to join the queue. Consequently, we assume in this study that customers' abandonment happens before they join the queue. Once joined, they will wait until the system finds them suitable matches.

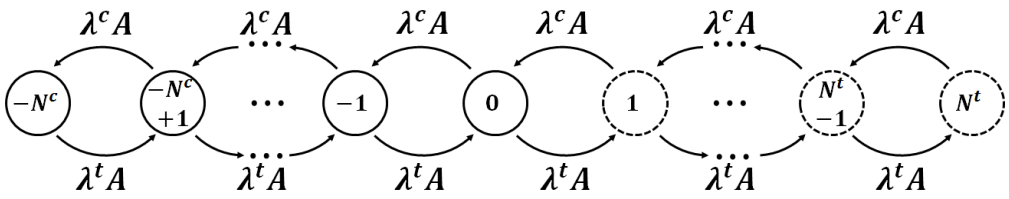


Figure 2.2: A double-ended queue characterized by each single matching block

where $N^c = A\lambda^t w_0^c$ and $N^t = A\lambda^c w_0^t$. Denote $\{\pi_i\}_{i\in\Omega}$ as the steady state probability. The state balances then yield the following equations:

$$\begin{aligned}
&A\lambda^t \cdot \pi_i = A\lambda^c \cdot \pi_{i+1} && i \in \{-N^c, N^t - 1\} \\
&(A\lambda^c + A\lambda^t) \cdot \pi_i = A\lambda^c \cdot \pi_{i+1} + A\lambda^t \cdot \pi_{i-1} && i \in \Omega/\{-N^c, N^t\} \\
&\sum_{i=-N^c}^{N^t} \pi_i = 1
\end{aligned}$$

which further yield the following:

$$\pi_i = \frac{\rho^{i+N^c} \cdot (1-\rho)}{1-\rho^{N^c+N^t+1}} \qquad \forall i \in \Omega \tag{2.8}$$

where $\rho = \lambda^t/\lambda^c < 1$.

2.4.2 Effects of matching block size (radius) on system performance

By fixing the fluxes λ^c and λ^t, this subsection investigates the consequence of varying the size of matching blocks. To facilitate the discussion, we first consider scenarios of severe supply shortage when ρ stays well below 1. Since drivers will enjoy relatively short time of idleness, we ignore their cruising behavior by assuming $w_0^t = +\infty$. Later in Section 2.4.4 when studying mild supply shortage scenarios, we bring back the finite w_0^t and discuss how different system performance measures are affected.

Below we derive four performance measures of primary concern to a platform operator. The monotonicity of each measure on the flux ratio ρ and the capacity of request queue (the maximum request queue length) N^c are presented.

1. *Percentage of abandonment* - p_a

$$p_a = \pi_{-N^c} = 1 - \rho$$

Clearly, p_a is decreasing on ρ and unrelated to N^c. The irrelevance of N^c is due to our assumption that idle drivers do not leave the block even with long time of waiting.

2. *Average queuing time for riders* - w^c

$$w^c = \frac{1}{\rho A\lambda^t} \cdot \sum_{i=1}^{N^c} i \cdot \pi_{1-i} = w_0^c \cdot \left(1 - \frac{\rho + \rho^2 + \cdots + \rho^{N^c}}{N^c}\right) \tag{2.9}$$

It is clear that w^c characterizes a function decreasing on ρ and increasing on N^c. Given ρ, w^c converges to w_0^c when N^c is sufficiently large.

3. *Average pickup time* - w^m

$$\begin{aligned} w^m &= \frac{\sqrt{A}}{v\rho} \cdot \left(d(1) \cdot \sum_{i=-N^c+1}^{0} \pi_i + \sum_{i=1}^{+\infty} d(i) \cdot \pi_i \right) \\ &= \frac{\sqrt{A}}{v} \cdot \left[d(1) - \rho^{N^c} \cdot \left(d(1) - \frac{1-\rho}{\rho} \cdot \sum_{i=1}^{+\infty} d(i) \cdot \rho^i \right) \right] \end{aligned} \tag{2.10}$$

where v is defined as vehicles' average travel speed, and $d(i)$ denotes the function of customers' average distance to the closest vehicle in a unit-size block with i idle drivers available for matching. Since $d(i)$ is strictly decreasing on i and $\sum_{i=1}^{+\infty} \rho^i = \rho/(1-\rho)$, it is easy to prove that the normalized pickup time $w^m/\sqrt{A}$ also characterizes a function decreasing on ρ and increasing on N^c. Again, when N^c becomes large and the customers' queue length stabilizes to be nonzero, the matching distance approximates the average distance between two randomly distributed points within a block.

4. *Queuing time variance for riders* - $\sigma_{w^c}^2$

$$\sigma_{w^c}^2 = \left(\frac{w_0^c}{N^c} \right)^2 \cdot \left[\sum_{i=1}^{N^c} i^2 \cdot \frac{\pi_{1-i}}{\rho} - \left(\sum_{j=1}^{N^c} i \cdot \frac{\pi_{1-i}}{\rho} \right)^2 \right]$$

It appears intractable to directly work on $\sigma_{w^c}^2$ due to N^c being discrete. We thus resort to its continuous approximation by treating Ω as a continuous state space over $[-N^c, +\infty)$:

$$\begin{aligned} \hat{\sigma}_{w^c}^2 &= \left(\frac{w_0^c}{N^c} \right)^2 \cdot \left[-\rho^{N^c} \ln \rho \cdot \int_0^{N^c} x^2 \rho^{-x} \mathtt{d}x - \left(\rho^{N^c} \ln \rho \cdot \int_0^{N^c} x \rho^{-x} \mathtt{d}x \right)^2 \right] \\ &= w_0^{c2} \cdot \left[1 - \frac{\left(\rho^{N^c} - \ln \rho^{N^c} \right)^2 - 1}{\ln^2 \rho^{N^c}} \right] \end{aligned}$$

Under the condition that $\rho^{N^c} \ll 1$, $\hat{\sigma}_{w^c}^2$ is increasing on ρ^{N^c}, which further yields $\hat{\sigma}_{w^c}^2$ being increasing on ρ and decreasing on N^c. The monotonicity of $\hat{\sigma}_{w^c}^2$ on N^c implies that when the platform shrinks the matching block, customers' waiting time in matching exhibits higher heterogeneity.

By fixing λ^c, λ^t and substituting N^c with $A\lambda^t w_0^c$, all the above performance measures become univariate functions on the block size A. Except for the percentage of abandonment p_a which is

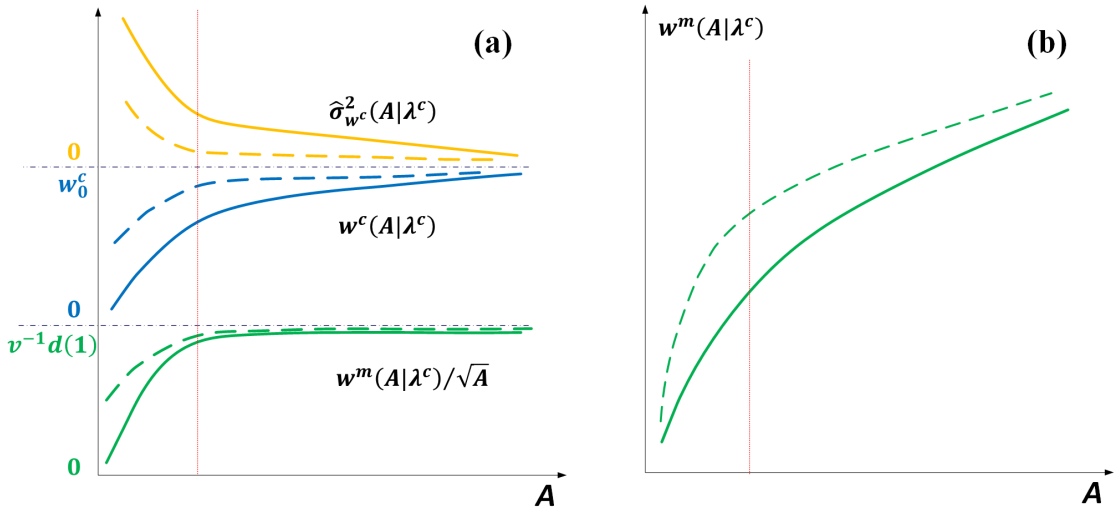


Figure 2.3: Effects of block size A on various system performance measures given λ^c and λ^t. (a) From top to bottom, the pairs of curves respectively show how $\hat{\sigma}^2_{w^c}$, w^c and $w^m/\sqrt{A}$ change against A; (b) the relationships between w^m and A. Notes: (1) For each pair of curves, the solid one is associated with a λ^c lower than the dashed counterpart; (2) The vertical red lines in both figures label a rough tipping point for A, below which drivers' idleness starts to intensify.

invariable, Fig. 2.3a depicts the relationships between the other three measures and A. As the space is isotropic, the ratio of arrivals ρ is invariable regardless of the size of A. Thus, starting from the state with the maximum request queue length N^c, each system state on average is realized with a decreasing probability that only depends on its relative distance to N^c. Suppose the system only stays over the n most-probable states, i.e. the queue length of requests being $N^c, N^c-1, \cdots, N^c-n+1$. Then, in the rightmost of Fig. 2.3a with A being very large, N^c is much greater than n. Since requests entering at the queue length of k will experience an average waiting time of $w_0^c \cdot k/N^c$, a small ratio of n over N^c implies that the corresponding queuing time falls closely below w_0^c. Therefore, the mean is close to w_0^c and its variance stays small. When A gradually shrinks, so does N^c, and n thus becomes more comparable to N^c, thereby driving the distribution of riders' queuing time to be scattered below w_0^c. As observed, before the tipping point (labeled as the red line in Fig. 2.3a), in response to a decreasing block size, the queuing time w^c gradually reduces while its variance $\hat{\sigma}^2_{w^c}$ mildly increases. The normalized pickup time $w^m/\sqrt{A}$ will consistently hold at $v^{-1}d(1)$, as it is highly probable that the system still holds a queue of requests and the rider at the head of the queue will be matched with the first coming driver in the block.

The mild variations start to intensify on the other side of the tipping point. As A continues to decrease, more and more riders enjoy zero queuing time while a constant portion, specifically $(1-\rho)\rho$, of riders will need to wait for a duration of w_0^c. In the case of extremely small A, riders will either experience no queue or a queuing time of w_0^c. Spatially, a portion of blocks will undergo

severe supply shortage, while the other accumulate excessive idle drivers waiting to be matched. We note that such extreme cases resemble the conditions of a street-hailing market, where riders' hailing experience becomes highly divergent during peak hours due to great search frictions.

Additionally, a pair of curves are provided in each subfigure of Fig. 2.3a to showcase the impacts of the request flux λ^c under different block sizes. The dashed curve is associated with higher λ^c. As λ^c influences the system solely via the flux ratio ρ, it is straightforward to conclude from our previous monotonicity analyses that by increasing λ^c, the per-rider cost of using the service w^c and w^m increase while the queuing experience tends to be more homogeneous for riders. Identically, due to the same mechanism discussed above, the effect from λ^c also intensifies on the left of the tipping point, i.e., the red line.

2.4.3 Supply curve of a system with matching blocks (radii)

The above subsection singles out the matching component and studies the effects of block size A on the system performances under severe supply shortage. In this subsection, by treating the idle vehicle flux λ^t as an endogenous variable, we now return to the system perspective as in Section 2.3.1 to investigate the supply curve of the ride-hailing system under the implementation of matching blocks.

One crucial "brand promise" of ride-sourcing services is to provide riders instantaneous matches to idle drivers nearby. As per Castillo et al. (2018), such a "promise" is so important that platforms are reluctant to undermine it to exchange for potential efficiency gains. Strategically, they attempt to avoid the situation where neighboring blocks exhibit too much difference in their level of service, e.g., one block has a long queue of requests while idle drivers are waiting inside its bordering blocks. Consequently, the size of matching blocks (radii) shall be determined to maintain the system to stay at states on the right side of the tipping point (i.e. the red line in Fig. 2.3), where drivers' idleness seldom arises[5].

Given the request flux λ^c and matching block size A, the system supply $S_e^r(C)$ thus solves the driver flux λ^t from the following system equation on C:

$$N = \lambda^t \cdot \left[w^m(\lambda^t|\lambda^c, A) + t\right] \tag{2.11a}$$

$$C = W_e\left(w^c(\lambda^t|\lambda^c, A),\ w^m(\lambda^t|\lambda^c, A)\right) \tag{2.11b}$$

Since the fare and delivery time are assumed fixed, the request flux λ^c directly reflects the potential customer demand. We note that the above equation set combines the demand and supply functions and it is difficult to separate them apart. Thus, we resort to varying λ^c and then deriving the

[5]Mathematically, the cumulative probability of having zero customers in the request queue ρ^{N^c} stays below a fairly small threshold.

response of λ^t and C to indirectly retrieve properties of the supply function. The process leads to the following theorems (mathematical proofs are provided in Appendix A.2):

Theorem 2.1. *When the block size (matching radius) is fixed, the supply curve still exhibits backward bending.*

Although the supply curves are still backward bending, the bend is weaker under a smaller block size.

Theorem 2.2. *A smaller block yields a milder bend in the supply curve. Specifically, corresponding to the same level of cost for riders, the backward-bending supply realized under a smaller block is greater than that under a larger block.*

Fig. 2.4 is a graphical illustration of Theorem 2.2. Typically, different sizes of matching blocks should capacitate the system at a similar level, because deadheading distances remain small around the transition point and the rationing of trips should be rare. However, after entering the domain of WGCs, the deadheading distances of drivers almost grow linearly proportional to the matching radius. Thus, the supply curves become more divergent as the supply shortage worsens, and systems with smaller matching radius can greatly reduce the distant pickups and sustain themselves at states of relatively high trip flux.

As illustrated in Fig. 2.4 or stated by Theorem 2.1, the supply curve is always backward bending for a ride-hailing system with constant matching radius, even if it is very small. However, we hypothesize that if the platform manages to adaptively adjust the matching radius, the backward

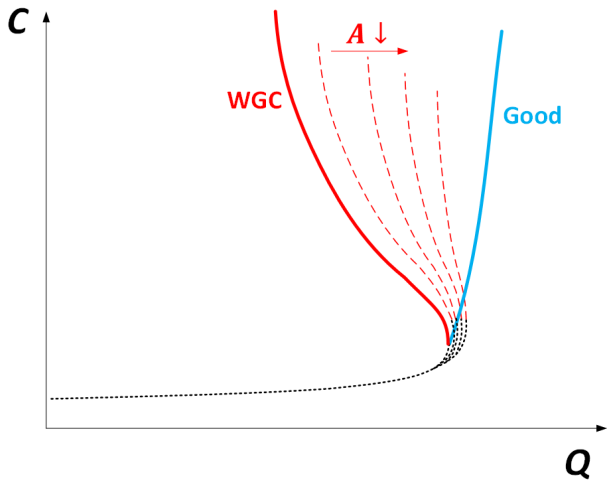


Figure 2.4: Supply curve of a system with matching blocks (radii). Note: Red curves depict the supply under constant block sizes, while the blue curve conceptualizes the realization of supply under adaptive block sizes.

bend can be completely avoided in an isotropic market. Below we provide evidence for the existence of such an operation.

Theorem 2.3. *For each supply state in the backward bending regime, there always exist smaller matching blocks that push the supply curve outward.*

Theorem 2.3 implies that there exists an operation strategy that adaptively adjusts the matching radius to completely erase the backward bend in the supply curve. The blue curve in Fig. 2.4 represents the outcome of such an operation. It can be seen that the curve from bottom to top subsequently intersects the backward bending supply curves with shrinking matching blocks, maintaining its monotonicity. Although such a supply would be more desirable, how to achieve it in practice remains an open question. Practically, given platforms' determination to deliver their "brand promise", it is advisable for them to adapt the block size (matching radius) around the tipping point shown as the red line in Fig. 2.3. As previously discussed, such operation reduces the deadheading time (see Fig. 2.3b) while maintaining spatial homogeneity in terms of request queuing within individual blocks.

2.4.4 Effects of drivers' random cruising

The previous discussions focus on severe supply shortage conditions, where drivers' idle time is kept nil and thus their cruising to neighboring blocks can be neglected. In this subsection, we turn to the mild shortage condition where the idleness of drivers becomes more significant. Thus, we consider drivers' waiting tolerance in matching w_0^t, as defined in Section 2.4.1, and investigate how the resultant cruising will impact customers' experience in matching.

In the context of cruising, the vehicle flux shall be separated into two groups, one representing the vehicles who arrive after delivering their riders in the block and the other being the vehicles cruising from other blocks. Assume drivers do not have any information on other blocks' real-time condition, and thus randomly pick a destination when they decide to leave the current block. Then, given the request flux λ^c, the delivery vehicle flux λ_e^t and the matching block size A, the total vehicle flux λ^t should satisfy the following condition:

$$\lambda_e^t = \lambda^t \cdot \left[1 - \pi_{N^t}\left(\lambda^t\right)\right] \tag{2.12}$$

where π_{N^t} given by Section 2.4.1 is the probability of one driver abandoning her current block to cruise to others. Eq. (2.12) essentially mirrors the conservation of idle drivers in the matching pool. Let $\rho_e = \lambda_e^t/\lambda^c$ and $\rho = \lambda^t/\lambda^c$. Substituting λ_e^t and λ^t in Eq. (2.12) yields

$$1 - \rho_e = \frac{1-\rho}{1-\rho^{N^t+N^c+1}}$$

which further gives rise to

$$\pi_i = (1 - \rho_e) \cdot \rho^{i+N^c} \qquad \forall i \in \Omega$$

The impacts on customers are then summarized as the following theorem.

Theorem 2.4. *Under drivers' random cruising, both riders' average queuing time w^c and pickup time w^m are decreasing functions of the maximum driver queue length N^t.*

The mathematical proof is provided in Appendix A.3. Theorem 2.4 implies that in an isotropic market during supply shortage, drivers' random cruising could potentially harm customers' level of service. Two mechanisms possibly render such negative effects. First, the cruising of drivers on average dilutes the accumulations of idle driver in blocks, causing more frequent supply shortage. As a consequence, ride requests will be more likely to go through the queuing process and receive distant pickups. Secondly, intensive cruising leads to higher arrivals of idle drivers, which further increase customers' tolerance on the queue length. Longer queuing capacity then drives the system towards the states with longer queues of requests, thereby sustaining the system over request queuing and raising the average queuing time. It should be noted that the above conclusion is derived under the assumption of random cruising. In practice, drivers may access real-time market information provided by the platform and thus their cruising behavior is more "targeted". How a targeted cruising will affect the system performance remains to be investigated.

2.5 Empirical Evidences

Two datasets have been used, each consisting of four-month records of trips in the mid-sized city in China. The first dataset was collected in periods when a capped surge pricing is implemented, while in the period of the second dataset the platform called off surge pricing and resorted to the queuing strategy. For each dataset, we aggregate the records by 15-min intervals to generate series of different system measures, and then drop those instances happened outside of 8:00-22:00.

We first apply the former dataset to verify that the pickup time w^m is indeed a monotonically decreasing function of the idle drivers' density N^v. Since the traffic condition is time-varying in reality, we verify the assumption by resorting to the relationship between the pickup distance d^m (i.e., getting rid of the speed effect) and the density of idle drivers N^v. The pickup distances we use are the estimated distances, which have been recorded for each match performed no matter whether it is completed eventually. The reason we use the estimated distances instead of those being realized is that the latter is likely to be biased (or filtered), as riders/drivers at times cancel orders if they are matched with long distance of pickups. Further, to show that $d^m \propto N^{v\alpha}$ with

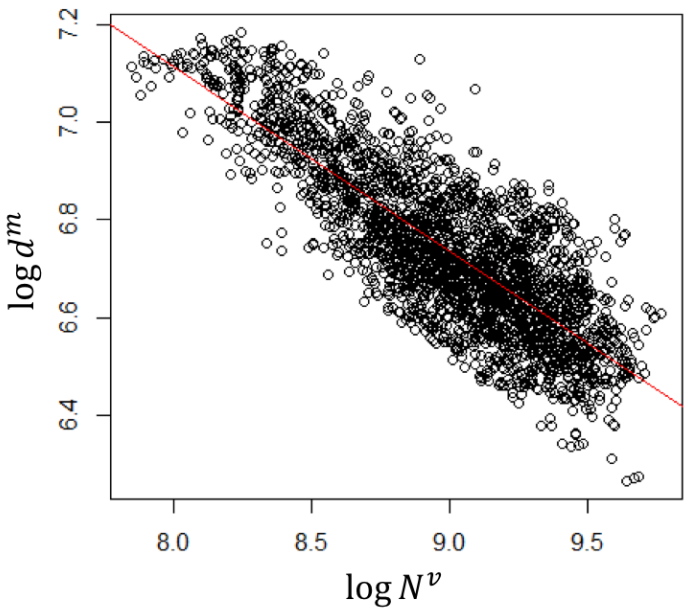


Figure 2.5: Relationship of the pickup distance d^m versus the number of drivers in idle N^v, as $\log d^m \sim \log N^v$. The red regression line is generated by OLS.

$\alpha < 0$, we plot the relationship of $\log d^m \sim \log N^v$ using data from 10:00-16:00 on weekdays. Such a range of time provides sufficient variations for observing the trends, while keeping the impact of matching radius marginal as the cap gets reached less frequently compared to the case in rush hours. Fig. 2.5 plots out the relationship of $\log d^m \sim \log N^v$. It is clear that the two measures are negatively correlated. By fitting a linear model (see the red line in Fig. 1), we find that the relationship $d^m \propto N^{v-0.38}$ explains over 67% of the variance in $\log d^m$. We thus think it is safe to assume in our model that without being capped the average pickup time will be shorter given denser idle drivers.

Both datasets are then used to show the effectiveness of the queuing strategy. Following Castillo et al. (2018), we group data points by the number of working drivers in each interval, and then graph the relationship between the estimated pickup time C and the supply of trips S in each group (with relatively consistent number of drivers) by adopting the local polynomial regression fitting. Fig. 2.6 outlines the supply curves for both cases, with the supply of trips normalized over [0,1] due to the confidentiality issues. Shaded regions represent 95% confidence intervals. Note that the case in 0%-20% quintile of drivers was hidden, because that group contains a number of outlier cases with very small size of drivers which violate the intention of consistent driver size among within-group samples. As can be observed, both figures show that the supply curves in

(a) (b)

Figure 2.6: Pickup time versus supply of trips in markets implemented (a) surge pricing but no queuing; (b) queuing but no surge pricing.

some circumstances tend to be backward-bended. Specifically, since larger sizes of working drivers usually imply high demand of rides during that time, the backward-bending supply in the former (Fig. 2.6a) and latter (Fig. 2.6b) periods respectively arises at the high-level and medium-level of demand. A possible reason is that in the former period there was no queuing system but a capped surge pricing. Consequently, the system was able to sustain the efficiency at the medium-level of demand but failed at high demand when the cap of the surge ratio became insufficient. Later, however, the system called off surge pricing and instead implemented the queuing system during peak hours. As mentioned, the queuing strategy was only active when the system experiences severe supply shortage. Consequently, the system supply gets restored at high-demand cases but worsens in the medium-demand regime.

2.6 Summary

This chapter investigates the efficiency of a ride-hailing system by studying its supply curves, which we define as the relationship between the throughput rate of a system and the cost of its riders. We model the supply functions of ride-hailing markets under different service conditions, and discuss the reasons and solutions for cases when inefficiencies arise. Our discussions start from a recent discovery of WGCs in isotropic ride-sourcing markets by Castillo et al. (2018). WGCs yield a waste of resource and cause the supply curve to be backward bending. To better address the issue, we resort to the idea of adjusting matching radius. A double-ended queuing model is established to explicitly capture the impact of finite matching radius. We prove that a smaller matching radius can yield a weaker bend, and further demonstrate the possibility of completely

avoiding the bend by adaptively adjusting the matching radius.

Future investigations may proceed along the following directions. First, even though this chapter reveals the feasibility of adaptive matching radius in avoiding backward-bending supply, an implementable scheme remains to be discovered. As platforms own much more market information beyond the arrival fluxes of riders and drivers, we expect to take additional variables into account for the design of adaptive schemes. Second, actual cruising behaviors of drivers can be far more complicated than random. As cruising considered by our model could harm the system performance, it is meaningful to study various ways to feed drivers with certain real-time market information to influence their cruising behaviors. Third, there are many variants in the ride-for-hire market whose supply curves are not covered by discussions in this chapter. For example, in taxi markets, drivers now widely adopt both street hailing and e-hailing and serve requests whichever come first from either mode; as for ride-sourcing markets, it is also crucial to examine how ride pooling/sharing affects the system supply and efficiency.

CHAPTER 3

"Countryside Delivery"

Price of Indiscriminate Matching

3.1 Introduction

By examining the supply curves of a homogeneous spatial market, Chapter 2 identifies a local matching failure in the ride-hailing system, named "Wild Goose Chases" (WGC). However, seldom in reality can we find a market that is perfectly isotropic, and the degree of anisotropy varies across the time of day and the day of week. Note that in an anisotropy ride-sourcing market, the supply condition depends on not only the local operations at subregions but also the interactions between them via the inter-subregion passenger trips and cruising idle drivers. In this chapter, we show another matching failure which arises when requests with different negative effects to the system efficiency are treated the same way in the matching process. Such a matching failure commonly exists in both e-hailing (Afèche et al., 2018) and street-hailing markets (Buchholz, 2018), and typically occurs in peak hours when travel demand highly surges along one specific direction. We further demonstrate that the failure can be solved by either using differentiated pricing or demand rationing.

The question then becomes the selection of an optimal policy for supply management. Ride-sourcing platforms nowadays have unprecedented controls over the system, but finding optimal control policies is very challenging due to the complexities underlying the system dynamics. The operations and management of ride-hailing services has attracted extensive research interests covering broad aspects of the service. See Wang and Yang (2019) for a recent, comprehensive review. This chapter contributes to this ongoing quest by proposing a macroscopic fluid model to help with strategic decision-makings of a ride-hailing platform. The model describes the system dynamics over a spatiotemporal network under control polices by a platform, and can thus evaluate the impacts of various policies. It can also be applied to seek for better policies to further achieve the managerial objectives of the platform.

The literature has relied excessively on canonical models that assume steady states in a ride-hailing system to derive various regulation, control, or policy to manage the system. A number of models are constructed based on highly simplified market/system settings, without or with stylized consideration of spatiotemporal variations (see e.g., Zha et al., 2016; Xu et al., 2017; Zha et al., 2018b; Yang et al., 2020b; Zhang et al., 2019). These models aim to offer profound theoretical or policy insights, but fall short of supporting strategic decision making for operating a real ride-hailing system, especially those with drastic spatiotemporal heterogeneity. As demonstrated by Xu et al. (2020), the heterogeneity can cause severe operational inefficiencies. Since Yang and Wong (1998), a series of studies have been carried out to better describe the spatial structure of a street-hailing taxi market (see Salanova et al., 2011 for a comprehensive survey). Recent studies have also extended the analysis to app-based ride-hailing systems (see e.g., Ban et al., 2019; Xu et al., 2019a as well as the citations therein). Another school of work apply a queueing-network-based approach to better capture market frictions and examine how various operations strategies impact system performance (e.g., Braverman et al., 2019). A common assumption made by all these studies is that the market condition is stationary with no variations or correlations across periods. However, a real ride-hailing system is subject to time-varying supply and demand that interact over transient system states. Therefore, to better cope with the operations of real ride-hailing systems, we need models or approaches that accommodate the system dynamics and spatial heterogeneity. Notably, among a few exceptions in the literature, Buchholz (2018) proposed a dynamic model of spatial search and matching between passengers and street-hailing taxis. Nourinejad and Ramezani (2020) introduced a dynamic model that tracks the time-varying number of riders, vacant and occupied vehicles in a ride-hailing system. However, their model is constructed for an aggregate market without spatial heterogeneity.

Another approach that becomes increasingly popular is reinforcement learning, which has been adopted to assist ride-hailing platforms with online matching and dispatching policies (see e.g., Ke et al., 2019; Shou et al., 2020). These studies disentangle the system dynamics as Markov decision processes, and then estimate policy values via designated learning approaches. However, the underlying learning process is a "black box" to the model builders, which makes the resultant policies ineffective for generalization. This chapter takes another route by proposing a generalized fluid modeling framework of ride-hailing systems constructed closely following its spatiotemporal dynamics. Managerial strategies of a platform are embedded mathematically into the framework and thus can be optimized systematically with observable market responses. The transparency of our framework's physical structure makes it highly transferable for customization and update. We demonstrate this by conducting an empirical study and customizing a model for a ride-hailing system in a large city in China. We calibrate the model with a high degree of granularity using a large-scale dataset from Didi Chuxing. It is demonstrated that the model is capable of describing

the market conditions and predicting the consequences of control policies; moreover, it can be incorporated into an optimization model to prescribe optimal operation strategies. To our best knowledge, the proposed modeling framework is one of the first to offer a tractable way to support the analysis and optimization of real, large-scale ride-hailing systems. The empirical study also provides one of the first testbeds for examining various operations strategies for ride-sourcing services.

The rest of the chapter flows as follows. Section 3.2 adopts a parsimonious two-node network to showcase the failure arising due to indiscriminate matching within an anisotropic market, and then discuss alternative solutions to resolve the issue. Section 3.3 presents a generalized fluid-based modeling framework for ride-hailing systems by assembling two groups of relationships, respectively resulting from the conservation law and statistical characteristics in fluid dynamics. Section 3.4 introduces a model customized and calibrated for a real ride-hailing system in a large city in China. In Section 3.5, we demonstrate how the model can be applied to prescribe optimal control policies and discuss a solution procedure to solve the resulting optimization model. Section 3.6 then presents numerical results that show the potentials of price leverages for the ride-hailing platform in service management. Lastly, Section 3.7 summarizes the chapter.

3.2 "Countryside Delivery" in Anisotropic Market

This section show that the indiscriminate treatment of riders during the matching process will give rise to a failure, named "countryside delivery", in an anisotropic market. To facilitate the understanding, we introduce the failure as well as its countermeasures based on a parsimonious two-node network.

3.2.1 A two-node network example

The two-node network is designed as Fig. 3.1a. Suppose a ride-hailing market can be abstracted into two nodes A and B, respectively representing two functional regions of a city. There are two types of travel demand originating from node A, intra-node demand Q_A that ends within A, and inter-node demand Q_B that ends at node B. For simplicity, we assume no demand generated from node B. Each request of either Q_A or Q_B takes an identical time of t to deliver, but drivers who drop off a passenger at node B shall spend another time t on driving back to node A, to continue her service. Further, assume there is no meeting frictions between idle drivers and riders on node A and pickup time is ignorable compared to the time of delivery. The demand profiles of Q_A and Q_B are assumed to follow a linear relationship to the customers' waiting time C shown in Fig. 3.1b. Denote Q_A^0 and Q_B^0 respectively as the potential requests for intra-node and inter-node

travel demand, and C_m as the maximum waiting time that a request can be held in the system. Specifically, it should be noted that the potential demand $\{Q_i^0\}_{i\in\{A,B\}}$ in this case only includes those who accept the price and trip time and decide to request for rides. And the maximum waiting time C_m can either be customers' tolerance in waiting or a time threshold imposed by the platform.

Suppose there are N vehicles in service and all requests are matched in an indiscriminate fashion, i.e. both types of trips will be imposed with identical waiting costs. We then fix Q_A^0 and vary Q_B^0 to see how the system supply will take its shape. Without loss of generality, let N be greater than tQ_A^0. Then, for Q_B^0 falling in the range of $[0, (N - tQ_A^0)/2t]$, customers' waiting costs remain zero and the supply of trips S increases on Q_B^0. But when the potential demand Q_B^0 goes beyond $(N - tQ_A^0)/2t$ and supply comes to a shortage, the realized demand (Q_A, Q_B) shall subject to the following equations:

$$\begin{aligned} \frac{Q_A}{Q_A^0} &= \frac{Q_B}{Q_B^0} \\ tQ_A + 2tQ_B &= N \end{aligned}$$

which further yields the service cost C and supply S below,

$$\begin{aligned} C\left(Q_B^0\right) &= C_m \cdot \left(1 - \frac{N}{t \cdot (Q_A^0 + 2Q_B^0)}\right) \\ S\left(Q_B^0\right) &= \frac{N}{t} \cdot \frac{Q_A^0 + Q_B^0}{Q_A^0 + 2Q_B^0} \end{aligned}$$

It is easy to prove that the service cost C will be increasing on Q_B^0, while the supply S monotonically decreases. Therefore, the supply curve again appears backward bending, as if the system had become congested on the process from B to A. Since the "congestion" arises owing to the indiscriminate matching of requests whose destinations exhibit long time of idleness for

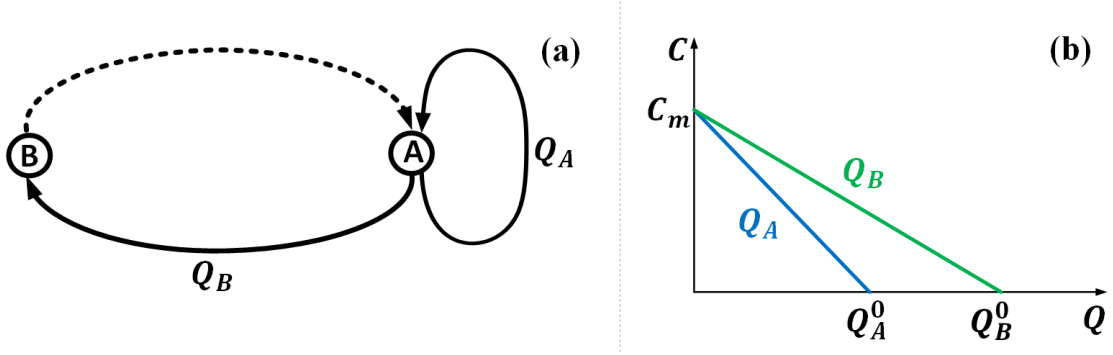


Figure 3.1: A two-node case. (a) A concept map of the two nodes. (b) Demand profiles for intra-node and inter-node trips on node A

drivers, we name it "countryside delivery (CSD)". Similar to WGC, CSD wastes the precious supply resource. Worse yet, it may greatly worsen the situation if drivers terminate their shifts after finishing an order at cold spots, due to the high perceived cost for serving another order. We note that CSD widely exists in practice, e.g., during peak hours, before a concert/game or even in off peaks with spatially unbalanced demand, when drivers have to passively serve trips from hot to cold spots, due to the no-rejection pledge (taxi market) or the penalty for declining ride requests (ride-sourcing market).

Further, it is worth mentioning that a paradox may arise when the ride-sourcing platform implements an indiscriminate surging pricing[1] in a supply shortage area, for the purpose of balancing the supply and demand therein. However, demands to cold spots in many instances have fewer alternative modes available, and thus are less sensitive to the price surge compared to those demands to hot spots. As a result, the orders that may lead the system to "congestion" will occupy higher proportion, further deteriorating the supply shortage.

3.2.2 Alternative solutions to "countryside delivery"

Two strategies widely adopted to relieve congestion in a congestible system are rationing and pricing (MacKie-Mason and Varian, 1995). In this subsection, we discuss how to use price or rationing discrimination (e.g., Zangui et al. 2013; Afèche et al., 2018) to address CSD. The basic idea is to adjust the portions of different demands in the matching pool and ensure the matching rates to reflect the saturation level of idle drivers at the trip destinations.

A recent paper from Didi Chuxing reports a matching algorithm that the platform recently

[1]To our best knowledge, most ride-sourcing platforms implement zone-based surge pricing, where all the rides generated from the same area during the surge period will be subject to the same surge ratio.

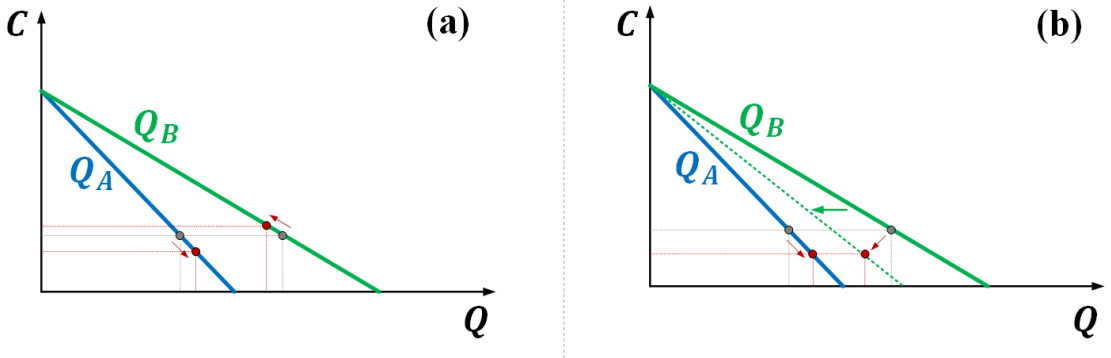


Figure 3.2: Changes of system states under (a) rationing; (b) pricing. Note: The gray dots represent the system states without policy intervention while the red dots denote the intervened equilibrium states. Arrows show the direction of system evolution under the corresponding intervention.

implemented (see Xu et al., 2018). Instead of performing "anonymous" bipartite matching, the platform associates a value with each destination zone that represents drivers' potential future gains after a driver delivers the rider and finishes the order in the zone. The value is typically higher for a zone that experiences supply shortage. Requests who head to this type of zones will be prioritized for matching. Such a differentiated matching algorithm essentially performs an intelligent rationing over requests under supply shortage. The demands traveling to hot spots will enjoy shorter matching time and become less likely to drop out, while their cold spot counterparts will suffer longer waiting time and are thus more likely to abandon the request. The potential system change can be conceived by projecting the mechanism onto the two-node network case shown as Fig. 3.2a. The rationing approach elevates the portion of hot-spot-heading demand by differentiating the matching time. According to Xu et al. (2018), this approach remarkably improves the system efficiency, raising the platform's revenue by 0.5% to 5% in major cities of China.

Another way to adjust the demand portions in the matching pool is via price discrimination with respect to their destinations. Bimpikis et al. (2019) conducted a pioneering work along this direction by investigating revenue maximization on a simplified ride-hailing network. According to their derivation (see Appendix C.2 in Bimpikis et al., 2019 for details), the revenue-maximizing OD-specific price should reflect the relative marginal value between having one driver available at the destination and the origin. Indeed, we note that such a price discrimination enhances system efficiency by optimizing the matching portions over different OD demands. The mechanism can be simply demonstrated by the two-node case in Fig. 3.2b, where the potential demand heading to hot/cold spots regions will be inflated/suppressed. If all the remaining requests receive the same level of service (indiscriminate in the matching process), the demand heading to hot spots will be served in higher portions.

3.3 Fluid-based System Model

To facilitate the implementation of the above strategies in real-world operations, we formulate a generalized fluid-based model to describe the system dynamics of a ride-hailing system under various control measures. Consider a city with $\mathcal{N}$ regions, where a ride-hailing platform manages a group of affiliated drivers to serve customers, each characterizing a ride from one region $i(\in \mathcal{N})$ to another $j(\in \mathcal{N})$. The dynamics of such a ride-hailing system can be largely captured by 10-tuple processes. Below details each process over a period of time $\mathcal{T}$, with a complete notational glossary provided in Appendix B.1 for reference.

3.3.1 Conservation in system dynamics

Denote $\mathcal{W}$ as the set of origin-destination (OD) pairs for ride requests. For brevity, we refer type-(i,j) riders to those whose requests are from i to j with $(i,j) \in \mathcal{W}$. Then, the processes $\{a_{ij}^c(t), t \in \mathcal{T}\}$ and $\{m_{ij}(t), t \in \mathcal{T}\}$ are both $\mathcal{W}$-dimensional, where $a_{ij}^c(t)$ and $m_{ij}(t)$ respectively denote the cumulative number of type-(i,j) riders having arrived and been matched at region i by time t. The process in $\{q_{ij}^c(t), t \in \mathcal{T}\}$ represents the number of type-(i,j) riders at region i waiting to be matched at time t. Naturally, we have the following relationship among these three processes,

$$a_{ij}^c(t) = m_{ij}(t) + q_{ij}^c(t) \qquad (i,j) \in \mathcal{W} \tag{3.1}$$

In contrast to $\{q_{ij}^c(t)\}$, the processes $\{w_{ij}^q(t), t \in \mathcal{T}\}$ represent the frictions in serving ride requests. We thus have:

$$a_{ij}^c(t) = m_{ij}(t + w_{ij}^q(t)) \qquad (i,j) \in \mathcal{W} \tag{3.2}$$

We note that $w_{ij}^q(t)$ reflects the average time that type-(i,j) riders need to wait for matching around time t under a general matching algorithm.

Let each process in $\{d_{ij}(t), t \in \mathcal{T}\}$ be the cumulative number of type-(i,j) riders delivered to their destination j by time t. Define $w_{ij}^s(t)$ as the service time (pickup plus in-vehicle time) for type-(i,j) riders matched at time t. Accordingly, the cumulative numbers of matches and deliveries are connected as follows,

$$d_{ij}(t) = \sum_{\tau \in \mathcal{T}} \mathbb{1}(\tau + w_{ij}^s(\tau) \leqslant t) \cdot \delta m_{ij}(\tau) \qquad (i,j) \in \mathcal{W} \tag{3.3}$$

where $\delta m_{ij}(\tau)$ represents the number of type-(i,j) riders newly matched at time τ.

At each region, idle drivers are generated from three sources: external arrivals, dropping off riders at the region, or cruising from neighboring areas. Define the processes $\{e_i(t), t \in \mathcal{T}\}$ and $\{a_i^d(t), t \in \mathcal{T}\}$ respectively as the external and total arrivals of idle drivers at region $i \in \mathcal{N}$ by time t. Let $\mathcal{W}^c$ be the full set of pairs for neighbouring or adjacent regions, and define the processes $\{c_{kl}(t), t \in \mathcal{T}\}$ as the cumulative number of idle drivers that successfully transit from region k to l with $(k,l) \in \mathcal{W}^c$ and contribute to the supply at l by time t. Then, we obtain the following equation regarding the entry of idle drivers:

$$a_i^d(t) = \sum_{j:(j,i)\in\mathcal{W}} d_{ki}(t) + \sum_{k:(k,i)\in\mathcal{W}^c} c_{ki}(t) + e_i(t) \qquad i \in N \tag{3.4}$$

On the other hand, for drivers already in a region, they could either stay idle, be matched to riders,

get offline, or transit to another region. Define the process in $\{q_i^d(t), t \in \mathcal{T}\}$ as the number of idle drivers at region i at time t, and each in $\{b_i(t), t \in \mathcal{T}\}$ to be the cumulative number of drivers getting offline at region i by time t. We thus have another conservation for drivers as follows:

$$a_i^d(t) = q_i^d(t) + \sum_{j:(i,j)\in\mathcal{W}} m_{ij}(t) + b_i(t) + \sum_{k:(i,k)\in\mathcal{W}^c} c_{ik}(t) \qquad i \in N \tag{3.5}$$

The above five conservation equations hold in general for non-pooling ride-hailing systems. They reveal the intrinsic interactions and dynamics of rider and driver flows that are isolated from market turbulence and system realizations. We will leverage these relationships to trace causal transmissions throughout the spatiotemporal network.

3.3.2 Statistical characteristics of system dynamics

Governed by the aforementioned conservation conditions, a ride-hailing system evolves with stochasticity and randomness, under the control policies by the platform. Thanks to the law of large numbers, below we capture statistical characteristics of the system dynamics.

At the demand side, the number of rides realized in the system is largely influenced by the service costs that riders experience at the stages of matching, pickup and delivery. The impacts can be summarized as the following latent function D_{ij} for each pair of $(i,j) \in \mathcal{W}$ during small interval δt around time t,

$$\mathbb{E}[\delta a_{ij}^c(t)]/\delta t = D_{ij}\left(\mathbb{D}^t[w_{ij}^m(t)],\ \mathbb{D}^t[w_{ij}^p(t + w_{ij}^m(t))],\ \mathcal{P}_{ij}(t),\ t\right) \tag{3.6}$$

where δa_{ij}^c represents the increment of type-(i,j) rider arrivals in δt, while $w_{ij}^m(\cdot)$ and $w_{ij}^p(\cdot)$ respectively denote arriving riders' matching and pickup time. The symbol $\mathbb{D}^t[\cdot]$ denotes the probabilistic distribution over time realized by the variable inside. The last term $\mathcal{P}_{ij}(t)$ denotes service fare of type-(i,j) riders at time t, which can be adjusted dynamically in accordance with the platform's pricing policy (Zha et al., 2018a; Bimpikis et al., 2019).

The online matching process performed by the ride-hailing platform can be generalized as the following many-to-one function $M_{ij}(\cdot)$ for each pair of $(i,j) \in \mathcal{W}$ and $t \in \mathcal{T}$, i.e.,

$$\mathbb{E}[\delta m_{ij}(t)]/\delta t = M_{ij}(\mathbb{D}^s[\boldsymbol{m_i}(\leqslant t)],\ \mathbb{D}^s[\boldsymbol{q_i^c}(\leqslant t)],\ \mathbb{D}^s[q_i^d(\leqslant t)],\ \boldsymbol{\mathcal{A}_i}(t)) \tag{3.7}$$

where $\boldsymbol{m_i}(\leqslant t)$, $\boldsymbol{q_i^c}(\leqslant t)$, and $q_i^d(\leqslant t)$ respectively indicate the historical realizations of processes $\{m_{ij}(t)\}$ and $\{q_{ij}^c(t)\}$ with $(i,j) \in \mathcal{W}$ and $\{q_i^d(t)\}$ at each region $i \in \mathcal{N}$ up to the time t. The action term $\boldsymbol{\mathcal{A}_i}(t)$ denotes the matching policy implemented by the platform at region i at time t. The uncertainty implied by the expression $\mathbb{D}^s[\cdot]$ results from the spatial distributions of riders and

drivers, as the platform considers not only their arriving sequence/time waited but also the costs of meeting physically (Yang et al., 2020a).

As aforementioned, the service time of riders $w_{ij}^s(t)$ sums up the time in pickup and delivery, i.e.,

$$w_{ij}^s(t) = w_{ij}^p(t) + w_{ij}^l(t + w_{ij}^p(t)) \qquad \forall (i,j) \in \mathcal{W} \tag{3.8}$$

where $w_{ij}^l(\tau)$ denotes the delivery time of the type-(i,j) rider who gets picked up at time τ. Since each region covers an area of space, pickup time $w_{ij}^p(t)$ and delivery time $w_{ij}^l(t)$ are both random. As an immediate result of matching, the statistics on pickup time $w_{ij}^p(t)$ are likely to rely on the same parametric setting as Eq. (3.7), i.e.,

$$\mathbb{E}[w_{ij}^p(t)] = T_{ij}(\mathbb{D}^s[\boldsymbol{m_i}(\leqslant t)],\ \mathbb{D}^s[\boldsymbol{q_i^c}(\leqslant t)],\ \mathbb{D}^s[q_i^d(\leqslant t)],\ \boldsymbol{\mathcal{A}_i}(t),\ t) \qquad \forall (i,j) \in \mathcal{W} \tag{3.9}$$

where $T_{ij}(t)$ characterizes the time-dependent function for type-(i,j) riders' expected pickup time that is variable to the real-time traffic condition in reality.

The states of a ride-hailing system also depend on drivers' decision-makings. For idle drivers, they can either wait in the current region or cruise to another one in search of the next rider. The aggregate movements of drivers can be measured by the following statistics,

$$\mathbb{E}[\delta c_{ij}(t)]/\delta t = C_{ij}(q_i^d(t),\ \boldsymbol{\mathcal{R}_i}(t),\ t) \qquad \forall (i,j) \in \mathcal{W}^c \tag{3.10}$$

$$\mathbb{E}[\delta b_i(t)]/\delta t = B_i(q_i^d(t),\ \boldsymbol{\mathcal{R}_i}(t),\ t) \qquad \forall i \in \mathcal{N} \tag{3.11}$$

where $C_{ij}(t)$ and $B_i(t)$ are two functions respectively related to drivers' expected rates of zonal transitions and service abandonment. The additional term $\boldsymbol{\mathcal{R}_i}(t)$ in the above functions stands for the intervention actions taken by the platform on drivers at region $i \in \mathcal{N}$ and time $t \in \mathcal{T}$.

3.3.3 Generalized fluid-based system model

The statistical relationships identified in the previous section reveal key physics or mechanisms of a ride-haling system. However, the involvement of random variables and statistical distributions makes it hard to connect with practical implementation. As a remedy, we apply fluid approximation. Fluid approximation first treats discrete riders and drivers as continuum and then takes the deterministic analog of system dynamics by omitting the uncertainty (Newell, 1971). Specifically for the processes previously discussed, the fluid approximation requires the removal of system uncertainty both in time and space (mathematically, by substituting the distributions of temporal and spatial processes with the corresponding mean values). Although the fluid

approximation undermines to a certain level the accuracy of system state estimates, it yields a continuous and differentiable mathematical framework that provides mathematical tractability for system analysis and optimization. In this chapter, we will demonstrate the descriptive capability of the resulting fluid model through an empirical study, but leave its qualitative properties for future investigation.

Applying the two approximation procedures to Eqs. (3.1-3.11) yields the following generalized fluid-based model for ride-hailing systems:

$$a^c_{ij}(t) = m_{ij}(t) + q^c_{ij}(t) \qquad (i,j) \in \mathcal{W} \tag{3.12a}$$

$$a^c_{ij}(t) = m_{ij}(t + w^q_{ij}(t)) \qquad (i,j) \in \mathcal{W} \tag{3.12b}$$

$$a^d_i(t) = \sum_{j:(j,i)\in\mathcal{W}} d_{ki}(t) + \sum_{k:(k,i)\in\mathcal{W}^c} c_{ki}(t) + e_i(t) \qquad i \in \mathcal{N} \tag{3.12c}$$

$$a^d_i(t) = q^d_i(t) + \sum_{j:(i,j)\in\mathcal{W}} m_{ij}(t) + b_i(t) + \sum_{k:(i,k)\in\mathcal{W}^c} c_{ik}(t) \qquad i \in \mathcal{N} \tag{3.12d}$$

$$d_{ij}(t) = \int \mathbb{1}(\tau + w^s_{ij}(\tau) \leqslant t) \cdot \mathrm{d}m_{ij}(\tau) \qquad (i,j) \in \mathcal{W} \tag{3.12e}$$

$$w^s_{ij}(t) = w^p_{ij}(t) + w^l_{ij}(t + w^p_{ij}(t)) \qquad \forall (i,j) \in \mathcal{W} \tag{3.12f}$$

$$w^p_{ij}(t) = T_{ij}(\boldsymbol{m_i}(\leqslant t),\ \boldsymbol{q^c_i}(\leqslant t),\ q^d_i(\leqslant t),\ \boldsymbol{\mathcal{A}_i}(t),\ t) \qquad \forall (i,j) \in \mathcal{W} \tag{3.12g}$$

$$\frac{\mathrm{d}a^c_{ij}(t)}{\mathrm{d}t} = D_{ij}\left(w^q_{ij}(t),\ w^p_{ij}(t + w^q_{ij}(t)),\ \mathcal{P}_{ij}(t),\ t\right) \qquad \forall (i,j) \in \mathcal{W} \tag{3.12h}$$

$$\frac{\mathrm{d}m_{ij}(t)}{\mathrm{d}t} = M_{ij}(\boldsymbol{m_i}(\leqslant t),\ \boldsymbol{q^c_i}(\leqslant t),\ q^d_i(\leqslant t),\ \boldsymbol{\mathcal{A}_i}(t)) \qquad \forall (i,j) \in \mathcal{W} \tag{3.12i}$$

$$\frac{\mathrm{d}c_{ij}(t)}{\mathrm{d}t} = C_{ij}(q^d_i(t),\ \boldsymbol{\mathcal{R}_i}(t),\ t) \qquad \forall (i,j) \in \mathcal{W}^c \tag{3.12j}$$

$$\frac{\mathrm{d}b_i(t)}{\mathrm{d}t} = B_i(q^d_i(t),\ \boldsymbol{\mathcal{R}_i}(t),\ t) \qquad \forall i \in \mathcal{N} \tag{3.12k}$$

By specifying the functional forms for Eqs. (3.12g-k) and the platform's policies in pricing $\boldsymbol{\mathcal{P}}$, matching $\boldsymbol{\mathcal{A}}$, and dispatching $\boldsymbol{\mathcal{R}}$, the dynamics Ω of a ride-hailing system, represented by the following 10-tuple processes, solves the system (3.12):

$$\Omega_{\boldsymbol{\mathcal{P}},\boldsymbol{\mathcal{A}},\boldsymbol{\mathcal{R}}} \triangleq \left\{ \begin{matrix} a^c_{ij}(t),\ m_{ij}(t),\ q^c_{ij}(t),\ w^q_{ij}(t),\ d_{ij}(t) \\ e_i(t),\ a^d_i(t),\ q^d_i(t),\ c_{ik}(t),\ b_i(t) \end{matrix} \middle| \begin{matrix} \text{Subject to Eqs. (3.12) under } \boldsymbol{\mathcal{P}}, \boldsymbol{\mathcal{A}}, \boldsymbol{\mathcal{R}} \\ \forall i \in \mathcal{N},\ (i,j) \in \mathcal{W}, t \in \mathcal{T},\ (i,k) \in \mathcal{W}^c \end{matrix} \right\}$$

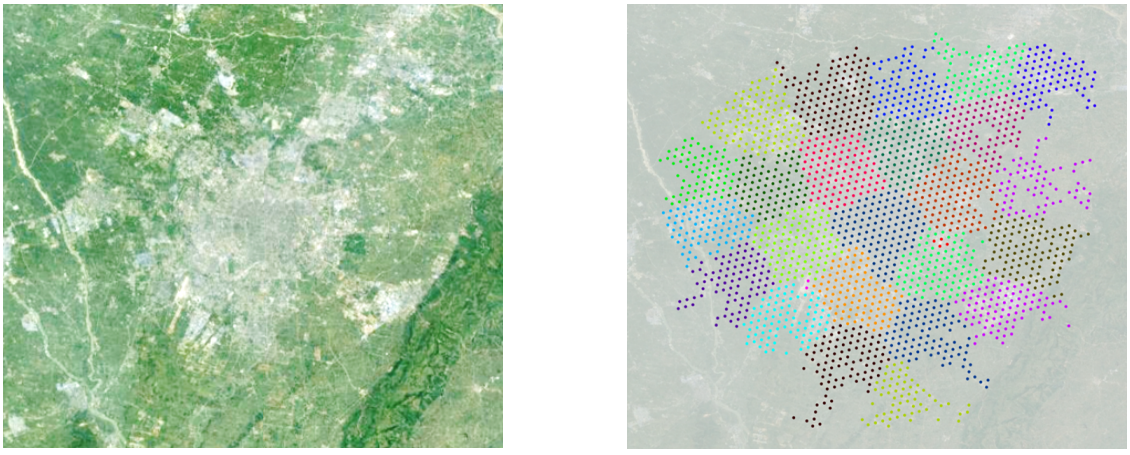
The above fluid modeling framework allows us to quickly identify the system states and capture the interaction of time-varying demand and supply across the spatiotemporal network. It is very general and can be customized to represent various types of strategies of pricing, matching, or dispatching, by virtue of the controllers embedded in the fluid model, and then evaluate their

impacts on system performance. Moreover, to further improve the performance, it is trivial to add another layer of optimization to prescribe optimal operation strategies, although it often requires non-trivial efforts to develop efficient solution algorithms. In the following sections, we exemplify how the proposed modeling framework can be customized and calibrated for a realistic system, and how the resulting model can be incorporated into an optimization model to prescribe optimal pricing policies for the system.

3.4 Parameterization, Calibration, and Validation of an Empirical Model

This section customizes the proposed generalized framework to model and analyze a real-world ride-hailing system. We consider the system operated by DiDi Chuxing in a large city in China. The city is partitioned into 27 regions, consisting of 24 spatial areas (neighborhoods) and three service spots (two railway stations and one airport). See Figure 3.3 for the spatial partitioning. Meanwhile, to facilitate the model calibration, we transform the above framework into a discrete-time counterpart and then parameterize the characteristic functions based on empirical observations and physical insights. The empirical data we have spans over an eight-week period in Summer 2019 (from June 24 to August 11). A granularity of five-minute interval is chosen for model calibration given the highly variable market condition, especially during the peak periods.

Assume all the state transitions of drivers and riders happen only at the beginning of each time interval and the system conditions remain stationary within the interval. Catering to such a discrete-



(a) (b)

Figure 3.3: Maps of (a) the city and (b) its 27 clustered regions, including 24 spatial service areas and 3 service spots (2 railway stations and 1 airport).

time context, we re-specify the label t to refer the time interval in $\mathcal{T}$ rather than a timestamp. The state variables (e.g. $q_{ij}^c(t)$ and $q_i^d(t)$) now denote the corresponding system condition at the beginning of interval t, while the prefix Δ in combination with the cumulative state variables (e.g., $a_{ij}^c(t)$ and $a_i^d(t)$) represent the respective increments during interval t. Below we then specify, calibrate, and validate different components of an empirical model with the contextual transitions.

3.4.1 Flow propagation and conservation

First of all, the flow conservation relationship can naturally be adapted as follows,

$$q_{ij}^c(t+1) = q_{ij}^c(t) + \Delta a_{ij}^c(t) - \Delta m_{ij}^c(t) \qquad (i,j) \in \mathcal{W} \quad (3.13a)$$

$$q_i^d(t+1) = q_i^d(t) + \Delta a_i^d(t) - \sum_{j:(i,j)\in\mathcal{W}} \Delta m_{ij}^d(t) - \Delta b_i(t) - \sum_{j:(i,j)\in\mathcal{W}^c} \Delta c_{ij}(t) \qquad i \in \mathcal{N} \quad (3.13b)$$

$$\Delta a_i^d(t) = \sum_{j:(j,i)\in\mathcal{W}} \Delta d_{ji}(t) + \sum_{j:(j,i)\in\mathcal{W}^c} \Delta c_{ji}(t) + \Delta e_i(t) \qquad i \in \mathcal{N} \quad (3.13c)$$

$$\Delta d_{ij}(t) = \sum_{\tau\in\mathcal{T}} \mathbb{1}\left(\tau + w_{ij}^p(\tau) + w_{ij}^l(\tau + w_{ij}^p(\tau)) = t\right) \cdot \Delta m_{ij}^d(\tau) \qquad (i,j) \in \mathcal{W} \quad (3.13d)$$

$$\Delta m_{ij}^d(t) = \eta_{ij}^s(t) \cdot \Delta m_{ij}^c(t) \qquad (i,j) \in \mathcal{W} \quad (3.13e)$$

where Eqs. (3.13a,b) respectively describe the updating of riders and drivers' queues; Eqs. (3.13c,d) respectively inherit Eqs. (3.12c,e) for each time slice, where the external arrivals of idle drivers $\Delta e_i(t)$ and the delivery time $w_{ij}^l(t)$ are both exogenous parameters directly retrieved/estimated from the empirical data; the last relationship (3.13e) is newly added, with the matching tuple $\{m_{ij}\}$ now split by drivers $\{m_{ij}^d\}$ and riders $\{m_{ij}^c\}$ to accommodate the unbalanced matching quantities resulting from the ride-pooling service, which shares the same pool of drivers together with the solo service during operations. The parameter $\eta_{ij}^s(t)(\leq 1)$ characterizes an OD-specific discounting ratio for the service demand given the availability of ride-pooling, where multiple customers can be served by a single driver simultaneously.

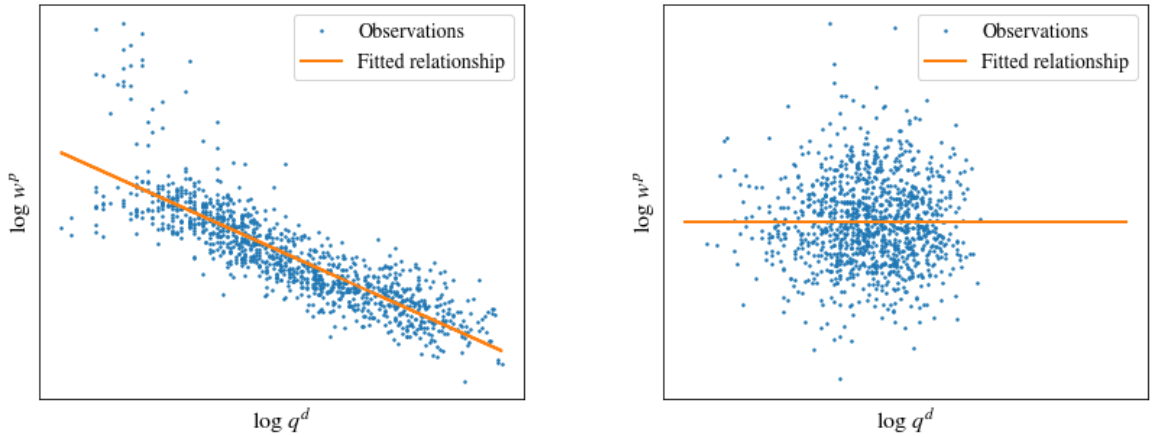
3.4.2 Parameterization and calibration of input functions

To fully capture the system dynamics, we then parametrize and calibrate the five input functions (3.12g-k) in the fluid model.

Pickup time function

The pickup time of riders who get matched in interval t is specified as a function of the instantaneous number of idle drivers within the region, i.e.,

$$w_i^p(t) = \alpha_i(t) \cdot [q_i^d(t)]^{\gamma_i(t)} \qquad \forall i \in \mathcal{N}, t \in \mathcal{T}$$



(a) Neighborhood (b) Airport

Figure 3.4: Relationships of the average riders' pickup time versus the number of idle drivers at (a) the urban district and (b) the airport between 13:00 to 16:00 of weekdays. Logarithms have been taken for both indices for the purpose of visualization.

where parameters α_i and γ_i are time-dependent constants to be calibrated. Such a power-function form has been rationalized in theory by many previous studies (see e.g., Daganzo, 1978; Arnott, 1996 for pioneering derivations). Figure 3.4 visualizes the relationships between the average riders' pickup time and the number of idle drivers in two example regions. Specifically, our calibration differentiates two types of service areas, neighborhood or spot. In neighborhoods, drivers are scattered spatially and matched by the platform to faraway riders who then require longer-distance pickups. Thus, the power index γ_i in this case is negative (see Figure 3.4a), implying the physics that the average pickup distances of riders stay shorter under higher density of idle drivers. In contrast, for service spots like airport and train station, which usually accumulate a large number of drivers waiting at designated areas, the pickup time of riders can be fairly constant (see Figure 3.4b). In these areas, the power γ_i is set to be zero. We note that such a specification aligns reasonably well with empirical data.

Demand function

The demand of rides $\Delta a^c_{ij}(t)$ in each interval t is calibrated via the following specification,

$$\Delta a^c_{ij}(t) = B_{ij}(t) \cdot \exp\left(\beta^q_{ij}(t)\bar{w}^q_{ij}(t) + \beta^p_{ij}(t)\bar{w}^p_{ij}(t) + \beta^f_{ij}(t)\mathcal{P}_{ij}(t)\right) \quad \forall (i,j) \in \mathcal{W}, t \in \mathcal{T} \quad (3.14)$$

where coefficient $B_{ij}(t)$ denotes the potential demand for the ride-hailing service during interval t and the vector $\boldsymbol{\beta}$ represents users' sensitivity to costs in different service stages, i.e., online matching ($\boldsymbol{\beta^q}$), pickup ($\boldsymbol{\beta^p}$), and delivery ($\boldsymbol{\beta^f}$). Conceptually, each of the latter exponents

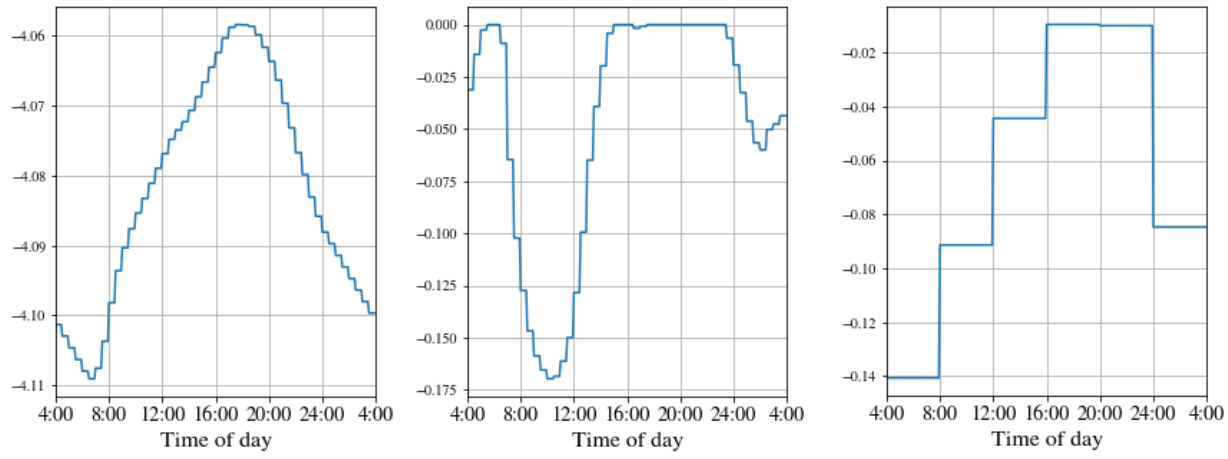
represents the successful conversion rate of users after they sense the costs in waiting and price. In our calibration, the average trip distance and trip time are also incorporated to control for their possible impacts on the realized demand. The time and price sensitivities are calibrated altogether across different period of a day but with different granularity. An optimization is constructed to perform the least square estimation with side-constraints restricting the estimates to stay in the negative regime. Figure 3.5 presents a set of calibrated sensitivities for the demand at the central district in a typical weekday. Interestingly, the calibration suggests a huge contrast between the morning and evening peaks. The demand is much more sensitive to additional waiting or price surge during the morning rush hours, but becomes neutral to the pickup time and price during the evening peak.

Note that $\bar{w}_{ij}^q(t)$ and $\bar{w}_{ij}^p(t)$ in the original demand function (3.14) are the average matching time and pickup time experienced by the riders who send out request at time t. In the discrete-time context, the two items are correspondingly calculated as

$$\bar{w}_{ij}^q(t) = \frac{\sum_{\{s\in\mathcal{T}|s\geqslant t\}} \pi_{ij}^c(t,s)\cdot(s-t)}{\sum_{s\in\mathcal{T}} \pi_{ij}^c(t,s)} \qquad \forall (i,j)\in\mathcal{W}, t\in\mathcal{T} \tag{3.15a}$$

$$\bar{w}_{ij}^p(t) = \frac{\sum_{\{s\in\mathcal{T}|s\geqslant t\}} \pi_{ij}^c(t,s)\cdot w_i^p(s)}{\sum_{s\in\mathcal{T}} \pi_{ij}^c(t,s)} \qquad \forall (i,j)\in\mathcal{W}, t\in\mathcal{T} \tag{3.15b}$$

where $\pi_{ij}^c(t,s)$ denotes the number of type-(i,j) riders who enter the system at interval t while



(a) β^q (b) β^p (c) β^f

Figure 3.5: Demand sensitivities in the central district on the matching time, pickup time, and fare for a typical weekday. The time sensitivities are calibrated with a 30-min granular, while the price sensitivity updates every four hours. The matching and pickup time is processed in the unit of hour, while the price is in Chinese Yuan (¥).

getting matched at interval $s(\geq t)$. The determination of $\pi_{ij}^c(t,s)$ will depend on the specific matching disciplines adopted by the platform, but is always guarded by the following two relationships,

$$\Delta m_{ij}^c(s) = \sum_{t \in \mathcal{T}} \pi_{ij}^c(t,s) \qquad \forall (i,j) \in \mathcal{W}, s \in \mathcal{T}$$
$$\Delta a_{ij}^c(t) = \sum_{s \in \mathcal{T}} \pi_{ij}^c(t,s) \qquad \forall (i,j) \in \mathcal{W}, t \in \mathcal{T}$$

In the later empirical analysis, we assume that riders of each OD type are matched with the first-come-first-serve (FCFS) discipline to pin down $\{\pi_{ij}^c(t,s)\}$.

Further, as implied by Eq. (3.15), both cost terms of riders are indeed realized endogenously. In other words, the market is dynamically equilibrated over time, where the demand rate of riders is realized based on the future supply condition.

Aggregate matching function

As evidenced by Figure 3.6, the numbers of riders and drivers waiting for matching in a neighborhood largely adhere to the relationship below,

$$\left(q_i^d(t) - q_i^0\right) \cdot \sum_{j:(i,j) \in \mathcal{W}} q_{ij}^c(t) = 0 \qquad \forall i \in \mathcal{N}, t \in \mathcal{T}$$

where q_i^0 denotes a threshold on the number of idle drivers. The equation essentially summarizes

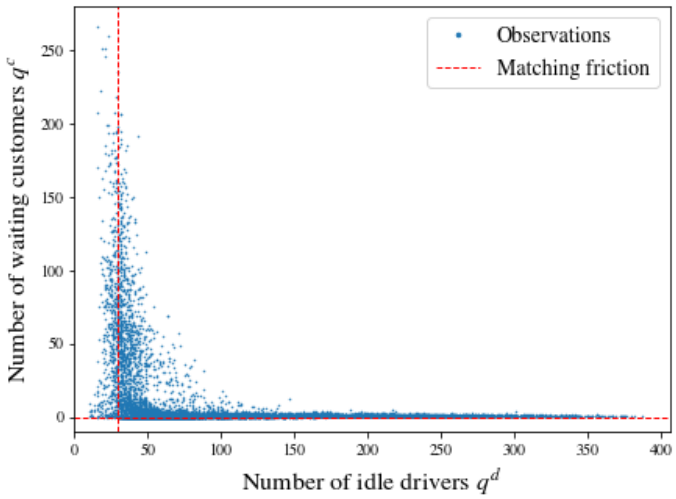


Figure 3.6: Relationship between the numbers of riders and drivers waiting to be matched at an urban district. The auxiliary red lines rationalize the use of a threshold-based matching function to model matching frictions.

the following matching regime: with a sufficient number of drivers (above the threshold that corresponds to the intersection of two red dash lines in Figure 3.6), incoming riders receive immediate matches. Otherwise, to avoid devastating long pickup, the platform intends to maintain the number of drivers to be above the threshold. In this case, riders have to wait. The waiting riders with different destinations are then matched proportionally.

Enlighten by the aforementioned regime, the aggregate matching function under the discrete-time context is specified as the equation below,

$$\begin{aligned}
\Delta m_{ij}^d(t) = \min\left(1,\ \frac{n_i^d(t) - q_i^0}{\sum_{k:(i,k)\in\mathcal{W}} \eta_{ik}^s(t) n_{ik}^c(t)}\right) \cdot \eta_{ij}^s(t) n_{ij}^c(t) \cdot (1 + \mathcal{A}_{ij}(t)) \quad \forall (i,j) \in \mathcal{W}, t \in \mathcal{T} \\
\text{subject to} \sum_{j:(i,j)\in\mathcal{W}} \eta_{ij}^s(t) n_{ij}^c(t) \cdot \mathcal{A}_{ij}(t) = 0 \\
\text{with} \quad \Delta m_{ij}^d(t) \in [0,\ \eta_{ij}^s(t) n_{ij}^c(t)].
\end{aligned}$$

where $n_i^d(t)$ and $n_{ij}^c(t)$ respectively denote the numbers of idle drivers and type-(i,j) riders participating in matching in region i at interval t. As per the equation, the platform in each interval either matches up all riders given sufficient supply or pairs them proportionately to maintain the supply level at the threshold. Specifically, the platform's action $\mathcal{A}_{ij}(t)$ is materialized as the weights/priorities in matching riders based on their destinations. In particular, if the platform does not prioritize riders according to their destinations (i.e., $\mathcal{A}_{ij}(t) = 0$), then all the riders waiting in the same area experience the same matching probability.

For our empirical analysis, the matching threshold $q_i^0(t)$ is determined to be the third percentile of the idle driver accumulation distribution for each region in each four-hour time interval of a day. This yields six thresholds for each region, each corresponding to one period of the day.

State transition rate

The last two functions are idle drivers' rate of regional transition and service abandonment, specified as the multiplicative form below for calibrations,

$$\begin{aligned}
\Delta c_{ij}(t) &= q_i^d(t) \cdot \Lambda_{ij}^c(t) \cdot (1 + \mathcal{R}_{ij}(t)) && \forall (i,j) \in \mathcal{W}^c \\
\Delta b_i(t) &= q_i^d(t) \cdot \Lambda_i^b(t) && \forall i \in \mathcal{N}
\end{aligned}$$

where $\Lambda_{ij}^c(t)$ and $\Lambda_i^b(t)$ respectively represent the probability of an individual driver at region i repositioning to k and abandoning the service during period t. The platform's action of dispatching idle drivers $\mathcal{R}_{ik}(t)$ is represented as the influences posed on the probability of repositioning towards adjacent regions.

In this chapter, we assume that idle drivers' probabilities of zonal repositioning $\{\Lambda_{ij}^c(t)\}$ and

service abandoning $\{\Lambda_i^b(t)\}$ are constants, unaffected by the changes of market condition. Thence, the two components are calibrated by taking the sample average probabilities for each region in each period. However, we caution that in reality drivers do notice the real-time market condition and respond accordingly. Thus, the state-transition probabilities may need to be specified as the function of variable market conditions. Our framework is open for accommodating more sophisticated considerations of drivers' behaviors once the behaviors are modeled and calibrated in the future. See, e.g., Urata et al. (2020) for some initial effort.

3.4.3 Empirical validation of the parametric model

Summing up all the above specifications yields an empirical model customized for the ride-hailing system of interest. This section validates the model using the processed eight-week service dataset. We intend to examine how well the macroscopic model performs on describing a real system, given all the spatiotemporal discretization and the variety of approximations made.

Figure 3.7 summarizes the weekly demand and supply pattern of the whole city across the eight weeks. The dotted blue lines show the average number of drivers in service (top) and requests being matched (bottom) respectively for five-min intervals in each hour of the week, while the surrounding shaded areas illustrate the 5th-95th percentile range of the corresponding distributions. As presented by the figure, both demand and supply exhibit well-shaped daily patterns with low scattering. The limited day-to-day variations suggest that after years of operations, the ride-sourcing service in this city has already reached a certain level of equilibrium. On the other hand, the substantial within-day variations in both state metrics call for a dynamic system model. To ensure the credibility of our calibrations and analysis, we choose to focus on the weekdays by narrowing the sample in each week to the period from Monday 4:00 AM to Friday 4:00 AM (as presented by the orange shared area in Figure 3.7), resulting in a dataset of 32 repetitive service cycles. All the empirical analyses in the following are performed in such a daily setting that considers a typical service cycle starting at 4:00 AM of a weekday.

Given each component calibrated following the above specifications, we now validate the proposed empirical model. We feed the currently realized trip demand into our calibrated spatio-temporal network and then compare the resultant system performance with the observed ones throughout the time of a day. We note that the fixed-demand scenario characterizes a special case that is easier to solve via a forward propagation scheme (see Section 3.5 for the details of the solution procedure). Figure 3.8 presents the comparisons of the predictions (the blue solid line) from the empirical model and the 32 observations from the daily samples (the grey dotted lines). Two performance measures are compared, respectively being the number of active drivers in service (fleet supply) and the number of riders delivered to their destinations (system throughput)

for each five-min interval. As can be clearly seen, the aggregate supply and demand predictions adhere reasonably well with the observed counterparts, validating the capability of our empirical model to capture the macroscopic dynamics of a realistic system.

The counterfactual policy analysis in Section 3.6 will further demonstrate the utilization of this calibrated empirical model for prescribing optimal policies. Before that, we first introduce the solution procedure to solve the policy optimization problems formulated for various managerial actions.

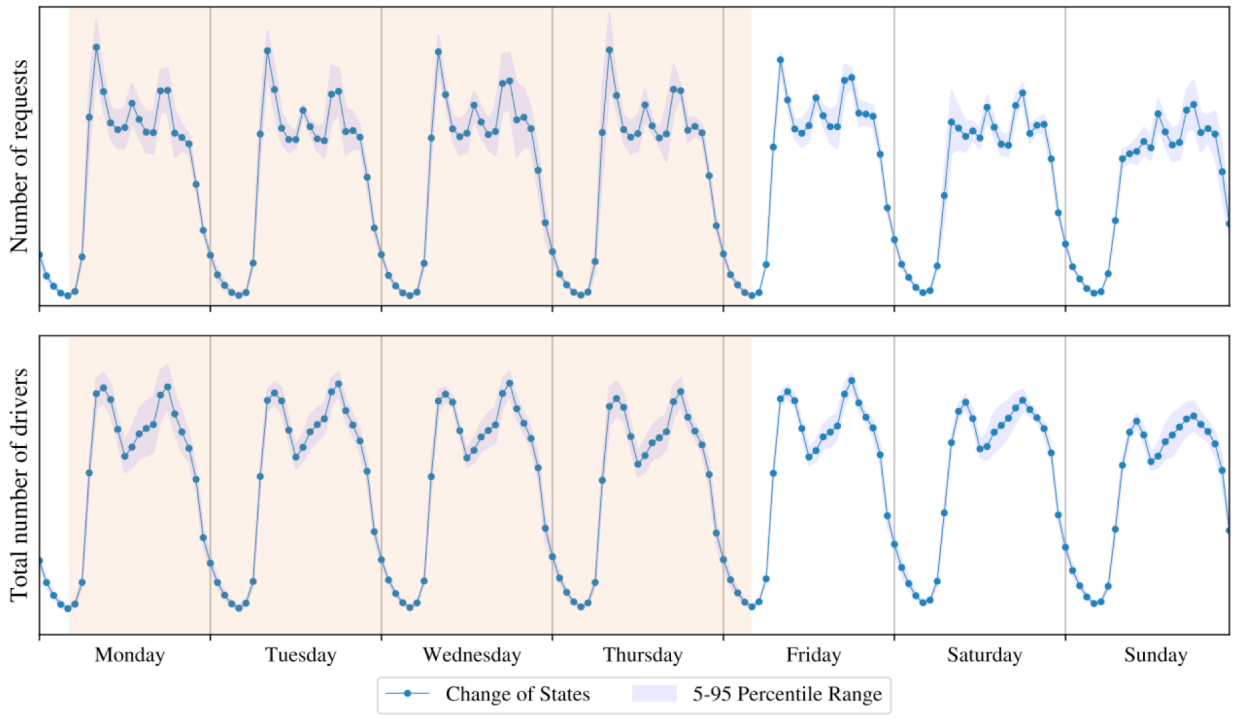


Figure 3.7: Variation of supply and demand by hour of week, using the eight-week dataset from June 24 to August 11, 2019. The blue dotted lines record the mean values averaged over the corresponding metrics across all the five-min intervals within each hourly period, while the blue shaded area around denotes the 5th to 95th percentile range of respective distributions. The orange shaded area presents the target time period that comes into our empirical study, taking the weekdays from Monday 4:00 AM to Friday 4:00 AM. Note that the scalings of the two metrics are hidden for confidentiality.

3.5 Solution Procedure for Policy Optimization

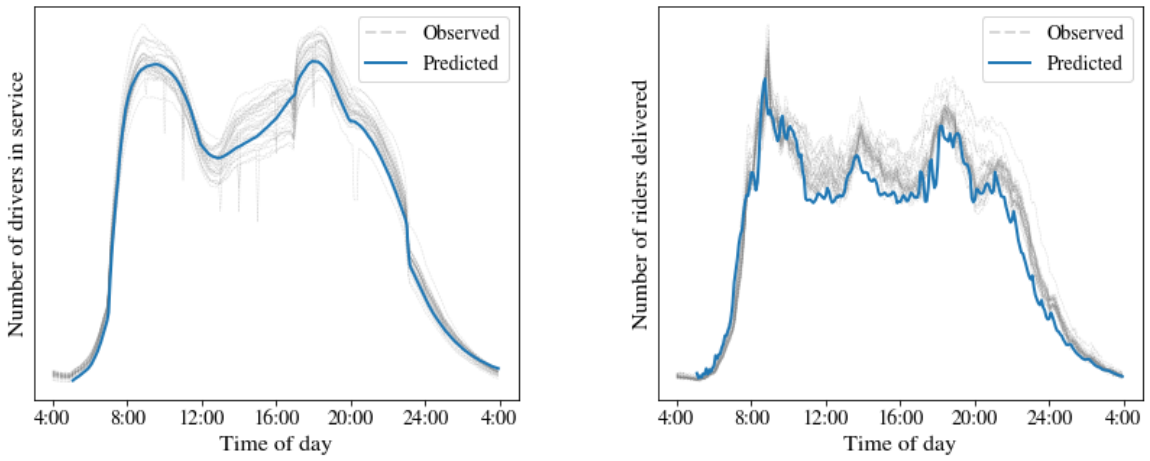
By virtue of the controllers embedded in the fluid model, it is conceptually straightforward to build optimization models over the system dynamics to seek policy improvements, i.e.,

$$\max_{\boldsymbol{\mathcal{P}},\boldsymbol{\mathcal{A}},\boldsymbol{\mathcal{R}}} Z = \sum_{(i,j)\in\mathcal{W},\ t\in\mathcal{T}} \Phi_{ij}(t)\cdot m_{ij}(t) \tag{3.16}$$

subject to the system equations (3.12).

where a platform adjusts the policies in pricing $\boldsymbol{\mathcal{P}}$, matching $\boldsymbol{\mathcal{A}}$, and/or dispatching $\boldsymbol{\mathcal{R}}$, to improve the system performance under certain managerial objective; and $\Phi_{ij}(t)$ denotes a measure of value taken for each request fulfilled. Note that such a multiplicative objective Z offers great flexibility and can be adapted to reflect a variety of managerial goals. For example, let $\Phi_{ij}(t) = 1$ reflects the goal of maximizing the system throughput, while prescribing $\Phi_{ij}(t)$ as $\mathcal{P}_{ij}(t)$ actually yields the case of a revenue-maximizing platform. The matching flows $\{m_{ij}(t)\}$ is derived subject to the relationships specified by the system dynamics. Simple as it appears though, the difficulty here is really about how to solve the resulting large-scale problem coupled with highly nonlinear constraints and possibly integer variables.

We thus devote this section to a heuristic algorithm that decomposes the whole problem into four subproblems or modules with each corresponding to a step in the proposed iterative procedure (see Figure 3.9 for a graphical description of the procedure). Below we introduce these modules.



(a) Number of drivers in service

(b) Number of riders delivered

Figure 3.8: Comparisons of the predicted and observed system states on supply and demand. The blue solid line denotes the predicted values suggested by the empirical model, while the grey dotted lines show the real observations from the 32 daily samples.

3.5.1 Mathematical modules in policy optimization

Four modules are presented, which are solved sequentially and iteratively to prescribe an optimal policy.

flow rebalancing module

Given the policy and trip demand, this module solves the bilateral matching and flow propagation problems in a forward propagation fashion. In other words, fixing the demand breaks the cyclic chains among the market conditions across periods, thereby enabling us to derive the flow prorogation by rolling one-way through the spatiotemporal network. The following demonstrates the flow rebalancing (FB) process at interval t for any region $i \in \mathcal{N}$:

(FB) Initialize with given $q_{ij}^c(t),\ q_i^d(t)$, and $\Delta d_{ij}(t)$.

S1. Calculate the numbers of riders $n_{ij}^c(t)$ and drivers $n_i^d(t)$ participating in matching by

$$n_{ij}^c(t) = q_{ij}^c(t) + \Delta a_{ij}^c(t)$$
$$n_i^d(t) = q_i^d(t) + \sum_{j:(j,i)\in\mathcal{W}} \Delta d_{ji}(t) + \Delta e_i(t)$$

where the newly arrivals of riders $\Delta a_{ij}^c(t)$ and drivers $\Delta e_i(t)$ are supplied exogenously.

S2. Calculate the matching flows $\Delta m_{ij}^c(t)$ and $\Delta m_{ij}^d(t)$ as

$$\Delta m_{ij}^c(t) = \min\left(1,\ \frac{n_i^d(t) - q_i^0}{\sum_{k:(i,k)\in\mathcal{W}} \eta_{ik}^s(t) n_{ik}^c(t)}\right) \cdot n_{ij}^c(t) \cdot (1 + \mathcal{A}_{ij}(t))$$

$$\Delta m_{ij}^t(t) = \eta_{ij}^s(t) \cdot \Delta m_{ij}^c(t)$$

and update the queue length $q_{ij}^c(t+1)$ and $\hat{q}_i^d(t)$ through

$$q_{ij}^c(t+1) = n_{ij}^c(t) - \Delta m_{ij}^c(t)$$
$$\hat{q}_i^d(t) = n_i^d(t) - \sum_{j:(i,j)\in\mathcal{W}} \Delta m_{ij}^d(t)$$

where $\hat{q}_i^d(t)$ is an intermediate state created for the driver's queue.

S3. Calculate the zonal transition flows $\Delta c_{ij}(t)$ and then update the queue $q_i^d(t+1)$ by

$$\Delta c_{ij}(t) = \hat{q}_i^d(t) \cdot \Lambda_{ij}^c(t)$$

$$q_i^d(t+1) = \left(1 - \Lambda_t^b(t)\right) \cdot \hat{q}_i^d(t) - \sum_{j:(i,j)\in\mathcal{W}^c} \Delta c_{ij}(t) + \sum_{j:(j,i)\in\mathcal{W}^c} \Delta c_{ji}(t)$$

S4. Update the pickup time $w_{ij}^p(t)$ and the future deliveries $\Delta d_{ij}(s)$ via

$$w_i^p(t) = \alpha_i(t) \cdot [q_i^d(t)]^{\gamma_i(t)}$$

$$\Delta d_{ij}(s) = \Delta d_{ij}(s) + \mathbb{1}\left(t + w_i^p(t) + w_{ij}^l(t + w_i^p(t)) = s\right) \cdot \Delta m_{ij}^d(t), \quad \forall s \geqslant t$$

S5. Transfer $q_{ij}^c(t+1),\ q_i^d(t+1)$, and $\Delta d_{ij}(t+1)$ to the next period.

The above procedure can be implemented repetitively across space and time to recover the balanced flows of riders and drivers under a given trip demand and policy.

Additionally, based on the balanced flows, we define the service ratio $\bar{S}_{ij}^c(t)$ of each type of riders at each interval t as,

$$\bar{S}_{ij}^c(t) = \min\left(1,\ \frac{n_i^d(t) - q_i^0}{\sum_{k:(i,k)\in\mathcal{W}} \eta_{ik}^s(t) n_{ik}^c(t)}\right) \cdot (1 + \mathcal{A}_{ij}(t))$$

Also, the number of drivers available for matching $\bar{U}_i^d(t)$ is defined as

$$\bar{U}_i^d(t) = n_i^d(t) - q_i^0$$

Both notions as well as the balanced flows will be fed into the later module to produce the marginal values for various flows.

Matching assignment module

The first module determines the number of riders matched in each interval, i.e. $\{m_{ij}^c(t)\}$. But as mentioned, to fully recover the trip demand, we need the average waiting and pickup time of incoming riders, which further require the more detailed flow assignment $\{\pi_{ij}^c(t,s)\}$, i.e., the number of type-(i,j) riders who request for trips at period t and get matched at period s.

Assume that riders of each OD type are matched with the FCFS discipline. Then, the assignment $\{\pi_{ij}^c(t,s)\}$ for given $\{\Delta a_{ij}^c(t)\}$ and $\{\Delta m_{ij}^c(t)\}$ solves the following linear program (MA) on each $(i,j) \in \mathcal{W}$:

$$\begin{aligned}
\text{(MA)}\ \max_{\boldsymbol{\pi}^c} & \sum_{t,s\in\mathcal{T}} \phi(s-t) \cdot \pi_{ij}^c(t,s) \\
\text{s.t.} & \sum_{t\in\mathcal{T}} \pi_{ij}^c(t,s) \leqslant \Delta m_{ij}^c(s) && \forall s \in \mathcal{T} \\
& \sum_{s\in\mathcal{T}} \pi_{ij}^c(t,s) \leqslant \Delta a_{ij}^c(t) && \forall t \in \mathcal{T}
\end{aligned}$$

$$\pi_{ij}^c(t,s) \geqslant 0, \text{ with } \pi_{ij}^c(t,s) = 0 \text{ if } t > s \qquad\qquad \forall t,\ s \in \mathcal{T}$$

where $\phi(s-t)$ represents an arbitrary value function that is strictly decreasing on the time duration $s-t$.

With $\{\pi_{ij}^c(t,s)\}$ ready, we then calculate the waiting time and pickup time for riders who request trips at t as per Eqs. (3.15), and apply these costs to Eq. (3.14) to update the trip demand.

Marginal effect module

Implementing the forward propagation scheme gives rise to the balanced flow pattern. However, with the flow information alone, it is difficult to identify where the operation bottlenecks are in the ride-hailing system. More importantly, we are blind to the global effects of the platform's interventions through local actions.

This module thus leverages a network flow model (NF) to make up the missing components by quantifying the marginal value of each movement/trip of drivers/riders, i.e., the change in the objective value caused by a unit change in a flow quantity. The model characterizes a linear program degenerated from the original policy optimization problem (3.16), i.e.,

$$\text{(NF)} \quad \max_{\boldsymbol{q^c},\boldsymbol{q^d},\boldsymbol{\Delta d},\boldsymbol{\Delta m^c},\boldsymbol{\Delta m^d},\boldsymbol{\Delta c}} Z = \sum_{(i,j)\in\mathcal{W},\ t\in\mathcal{T}} \Phi_{ij}(t)\cdot m_{ij}^d(t) \tag{3.17a}$$

$$\text{s.t.} \quad q_{ij}^c(t+1) = q_{ij}^c(t) + \Delta a_{ij}^c(t) - \Delta m_{ij}^c(t) \qquad (i,j)\in\mathcal{W} \tag{3.17b}$$

$$\hat{q}_i^d(t) = q_i^d(t) + \sum_{j:(j,i)\in\mathcal{W}} \Delta d_{ji}(t) + \Delta e_i(t) - \sum_{j:(i,j)\in\mathcal{W}} \Delta m_{ij}^d(t) \qquad i\in\mathcal{N} \tag{3.17c}$$

$$q_i^d(t+1) = \hat{q}_i^d(t) - \Delta b_i(t) - \sum_{j:(i,j)\in\mathcal{W}^c} \Delta c_{ij}(t) + \sum_{j:(j,i)\in\mathcal{W}^c} \Delta c_{ji}(t) \qquad i\in\mathcal{N} \tag{3.17d}$$

$$\Delta d_{ij}(t) = \sum_{\tau\in\mathcal{T}} \mathbb{1}\left(\tau + w_{ij}^s(\tau) = t\right)\cdot \Delta m_{ij}^d(\tau) \qquad (i,j)\in\mathcal{W} \tag{3.17e}$$

$$\Delta m_{ij}^d(t) = \eta_{ij}^s(t)\cdot \Delta m_{ij}^c(t) \qquad (i,j)\in\mathcal{W} \tag{3.17f}$$

$$\Delta b_i(t) = \hat{q}_i^d(t)\cdot \Lambda_t^b(t) \qquad i\in\mathcal{N} \tag{3.17g}$$

$$\Delta c_{ij}(t) = \hat{q}_i^d(t)\cdot \Lambda_{ij}^c(t) \qquad (i,j)\in\mathcal{W}^c \tag{3.17h}$$

$$\Delta m_{ij}^c(t) \geqslant \bar{S}_{ij}^c(t)\cdot\left(q_{ij}^c(t) + \Delta a_{ij}^c(t)\right) \qquad (i,j)\in\mathcal{W} \tag{3.17i}$$

$$\sum_{j:(i,j)\in\mathcal{W}} \Delta m_{ij}^d(t) \leqslant \bar{U}_i^d(t) \qquad (i,j)\in\mathcal{W} \tag{3.17j}$$

All the constraints are essentially the flow propagation procedures specified in the flow rebalancing module. We only replace the matching component with the last two equations (i.e., 3.17i,j), where $\bar{S}_{ij}^c(t)$ and $\bar{U}_i^d(t)$ are notions previously created based on the balanced flows. As one may notice, under such a construction, the balanced flow pattern solved by the forward propagation scheme is

indeed the only solution that satisfies all the constraints of the system (3.17). Thus, the problem is readily solved.

Our purpose of solving this problem is to quickly retrieve the dual for different variables and constraints, which indicates the marginal value of various flows in the spatiotemporal network. For example, the dual to constraint (3.17b) suggests the marginal value of increasing the trip demand from type-(i, j) riders at time t. Meanwhile, the variables $\delta_{ij}^c(t)$ and $\Delta c_{ij}(t)$'s dual can imply the marginal effect of speeding up the matching for type-(i, j) riders and the repositioning of idle drivers from region i to j at period t, respectively. All this information can be useful in analyzing pricing, matching, and repositioning strategies, and is thus collected and fed into the last module for policy improvement.

Policy improving module

The retrieval of marginal flow values enables us to formulate the policy improving problem as a quadratically constrained program that can be solved efficiently. To better present the idea, we define $\boldsymbol{f}$ as an augmented vector that spells all types of flows, including those of customer arrivals $\boldsymbol{\Delta a}^c$, matching $\boldsymbol{\Delta m}$, and repositioning of drivers $\boldsymbol{\Delta c}$, etc. Also, define $\boldsymbol{P}$ as a summary of all the existing policies considered. Let $\boldsymbol{\Delta f}$ and $\boldsymbol{\Delta P}$ denote the corresponding incremental changes in flow volumes and policies, respectively.

The policy improving module then solves the program (PI) below for $\boldsymbol{\Delta P}$:

$$\text{(PI)} \quad \max_{\boldsymbol{\Delta P}, \boldsymbol{\Delta f}} \ \boldsymbol{\psi_f} \cdot \boldsymbol{\Delta f} + \boldsymbol{f} \cdot \boldsymbol{\Psi}(\boldsymbol{\Delta P}) \tag{3.18a}$$

$$\text{s.t.} \ \ \boldsymbol{\Delta f} = \frac{\boldsymbol{\partial f}}{\boldsymbol{\partial P}} \cdot \boldsymbol{\Delta P} \tag{3.18b}$$

$$\|\boldsymbol{K} \cdot \boldsymbol{\Delta f}\| \leqslant \boldsymbol{\nu} \tag{3.18c}$$

where $\boldsymbol{\psi_f}$ denotes the marginal values for different flows obtained from the previous module; $\Psi(\cdot)$ specifies a linear function that captures the changes of values for each unit of driver and rider flows due to the updates of the policy ΔP. The two terms $\boldsymbol{\psi_f} \cdot \boldsymbol{\Delta f}$ and $\boldsymbol{f} \cdot \boldsymbol{\Psi}(\boldsymbol{\Delta P})$ in (3.18a) capture the consequences on the original optimization objective (3.16) from the incremental and existing flows, respectively. The constraint (3.18b) characterizes a possibly huge system of linear equations that connect the changes of flows to the policy adjustments. The Jacobian matrix $\frac{\boldsymbol{\partial f}}{\boldsymbol{\partial P}}$ here is taken as the constant that valued at the current system state $(\boldsymbol{f}, \boldsymbol{P})$. Note that the derivations of the Jacobian matrix may involve some efforts of system approximations, as not all the relationships are explicitly functional. Since all the impacts are evaluated with the first-order approximations, quadratic regularization constraints (3.18c) are imposed to ensure the local search of a new policy, where $\boldsymbol{K}$ is a coefficient matrix and $\boldsymbol{\nu}$ is a set of bounds. Some fine-tuning of both parameters could significantly accelerate the search of improving policies in implementation.

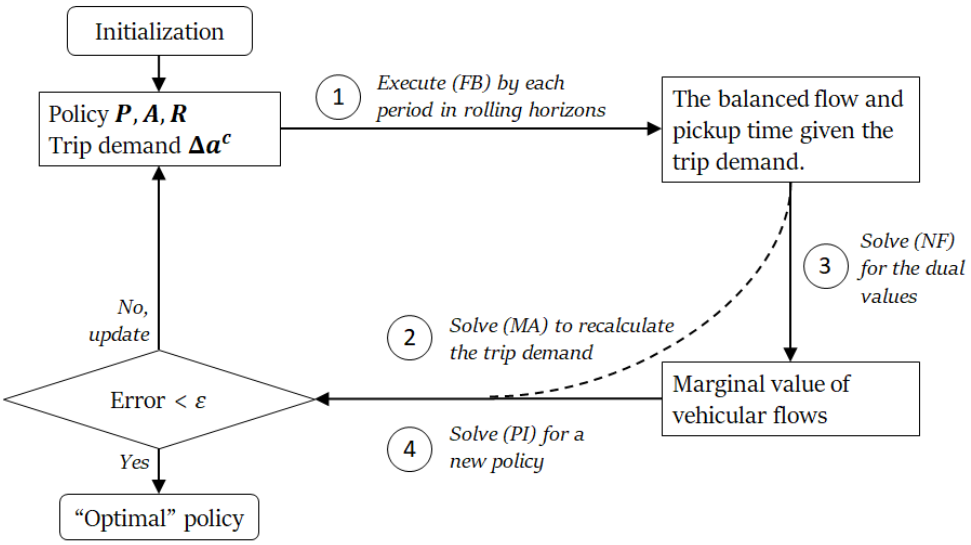


Figure 3.9: Solution algorithm for policy optimization

3.5.2 Solution procedure

Given the above deductions, we now present the following solution procedure to solve the policy optimization (see Figure 3.9 for the corresponding flow chart):

Initialize the policy $\boldsymbol{\mathcal{P}}, \boldsymbol{\mathcal{A}}, \boldsymbol{\mathcal{R}}$ and trip demand $\boldsymbol{\Delta a}^c$.

S1. Derive the balanced flow pattern and pickup time by executing (FB) in forward propagation, and calculate the objective value $\hat{Z}$ given by Eq. (3.16).

S2. Obtain the matching assignment $\{\pi_{ij}^c(t,s)\}$ by solving (MA) and then update trip demand $\Delta\hat{a}^c$ for Step 5.

S3. Solve the network flow problem (NF) to retrieve the marginal value of different actions/policies.

S4. Obtain the new policy $\{\hat{\boldsymbol{\mathcal{P}}}, \hat{\boldsymbol{\mathcal{A}}}, \hat{\boldsymbol{\mathcal{R}}}\}$ by solving (PI).

S5. If $\|\boldsymbol{\Delta a^c} - \boldsymbol{\Delta \hat{a}^c}\| < \varepsilon_a$, then

- Update the "optimal" policy with $\{\boldsymbol{\mathcal{P}}, \boldsymbol{\mathcal{A}}, \boldsymbol{\mathcal{R}}\}$, if $\hat{Z}$ is higher than the best value achieved so far,
- Stop the search and return the "optimal" policy, if $\kappa_p\|\boldsymbol{\mathcal{P}} - \hat{\boldsymbol{\mathcal{P}}}\| + \kappa_a\|\boldsymbol{\mathcal{A}} - \hat{\boldsymbol{\mathcal{A}}}\| + \kappa_r\|\boldsymbol{\mathcal{R}} - \hat{\boldsymbol{\mathcal{R}}}\| < \varepsilon_P$;

Otherwise, update the current policy and demand $\{\boldsymbol{\mathcal{P}}, \boldsymbol{\mathcal{A}}, \boldsymbol{\mathcal{R}}, \boldsymbol{\Delta a}^c\}$ using its linear combination with $\{\hat{\boldsymbol{\mathcal{P}}}, \hat{\boldsymbol{\mathcal{A}}}, \hat{\boldsymbol{\mathcal{R}}}, \boldsymbol{\Delta \hat{a}}^c\}$ and repeat the process from S1.

where κ in the last step characterizes a set of weights and ε is a given tolerance. Given the complexity of the procedure, its convergence properties are not yet theoretically investigated. However, for our all numerical experiments conducted for counterfactual analyses in the subsequent section, the procedure has always converged efficiently.

3.6 Counterfactual Analysis on Pricing Policy

Adopting the above solution procedure, this section conducts counterfactual analyses to demonstrate the applicability of our framework as well as to examine the potentials of different pricing strategies for optimizing system performance. The impacts of four different pricing schemes are respectively quantified and compared. The baseline is the status-quo pricing scheme with a fixed price structure, regardless of riders' location and time. The second scenario considers a time-of-day (dynamic) pricing that at each period applies a surge multiplier to adjust the price proportionately for riders. On top of dynamic pricing, the third and fourth scenario further varies the price by riders' origins (spatial pricing) and OD pairs (OD-specific pricing), respectively. For the latter three scenarios, the optimal pricing strategies are investigated to achieve different goals of service management.

Our counterfactual experiments are conducted in the morning (6:00-10:00 AM) of the averaged weekday (see Section 3.4.3 for the production of the "typical" weekday). The constrained trip fares are allowed to vary in the range between a 15% discount to a 50% surge based on the status-quo prices. The proposed procedure has converged successfully in all these numerical cases within four to five hundreds of iterations, each taking about six seconds to execute with our GAMS code on a desktop with Intel i5-6600 CPU@3.30Ghz and 8GB memory.

3.6.1 Throughput-maximizing pricing policies

We first examine the policies that maximize the system throughput, which is defined as the number of customers delivered to their destinations per unit of time, an important metric for the production of a ride-hailing system. It is also directly related to the utilization of ride-hailing vehicles. For a system with relatively homogeneous trip length among passengers, the system throughput is proportional to the percent time that in-service vehicles are occupied and paid.

Figure 3.10 first compares the system throughput (measured by the total time being occupied or paid) yielded by four pricing scenarios. The blue columns represent the total occupied time of drivers, while the orange polyline presents their percent time being paid during the investigated

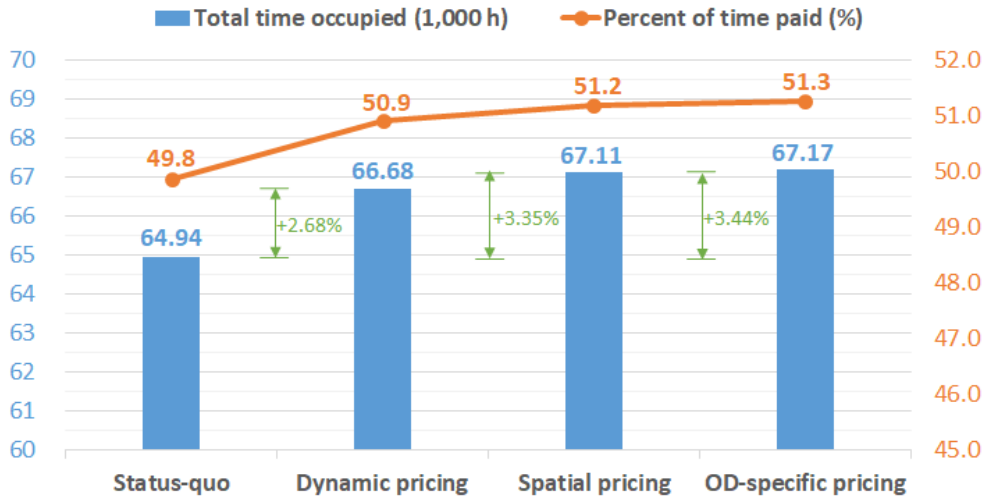
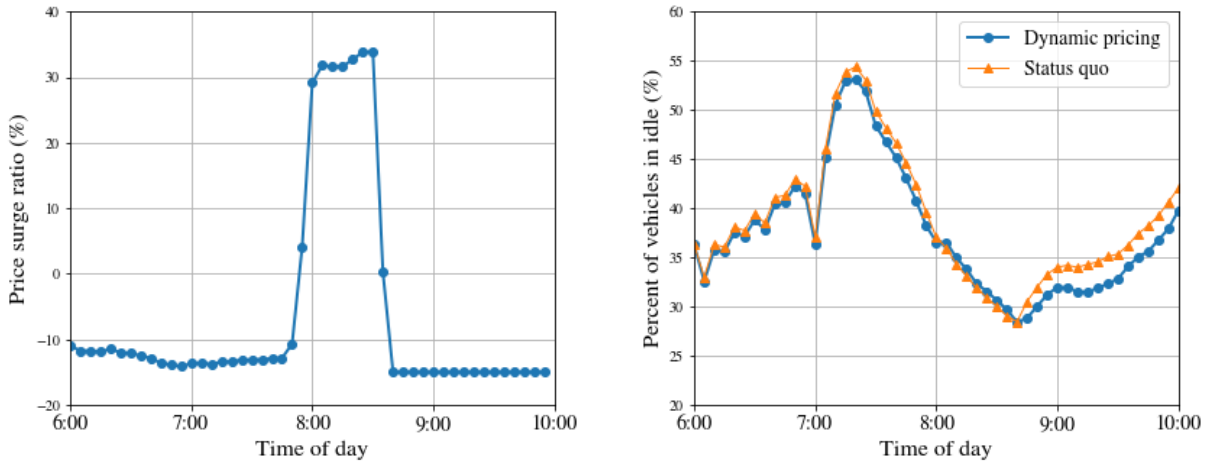


Figure 3.10: Comparison of the optimal pricing schemes under throughput maximization

period. As we can see in Figure 3.10, both measures improve constantly as the pricing policy becomes more differentiated. Specifically, the ride-hailing system will be able to accommodate 2.68% more passenger trip time by allowing the price to vary in time, and the number further rises to more than 3% as additional flexibility is provided for handling the spatial heterogeneity. Meanwhile, drivers in service also achieve higher utilization as the percent of time paid increases from 49.8% to 51.3%. These improvements imply the potentials of the system to accommodate more passenger trips and utilize the supply resources more efficiently through price adjustments.

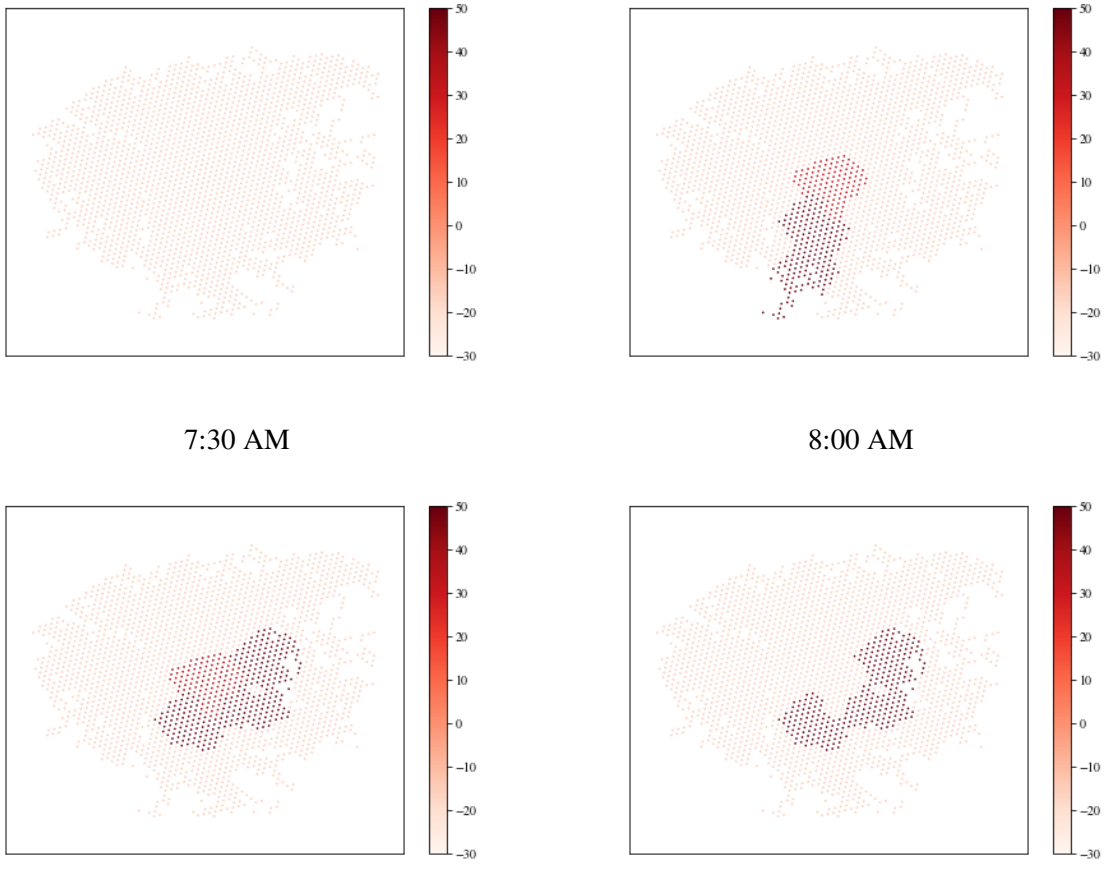
To help us better understand the platform's strategies, Figure 3.11 and 3.12 respectively visualize the surge ratios (or adjustment ratios) of prices for maximizing the system throughput under dynamic and spatial pricing. The surge ratios suggested by both figures indicate the necessity of context-dependent strategies in service management, as the market condition varies drastically in space and time. Figure 3.11a shows that a throughput-maximizing platform would lower the prices in non-peak periods, while lifting them during the rush hours. For non-peak periods, the system with sufficient supply retains extra capacity to accommodate more trip of rides. By lowering down the prices, the platform aims to attract and serve more customers. In contrast, the platform opts to increase the price to suppress the demand for achieving higher throughput during the peak period, i.e., 8:00-8:30 AM, for this city. Such a strategy is actually consistent with the suggestion discussed in the literature to address the "wild goose chases" phenomenon, where a platform may need to buffer the idle vehicle accumulations by suppressing the demand to avoid the enormous supply wastes in devastatingly long pickups during the peak periods (see Castillo et al., 2018; Xu et al., 2020 for specific discussions). Figure 3.11b evidences such reasoning by comparing the percent of idle vehicles for dynamic pricing and status quo, respectively. It is observed that the dynamic



(a) Price surge ratio (%)

(b) Percent of vehicles in idle (%)

Figure 3.11: Dynamic pricing for throughput maximization in the morning.



7:30 AM

8:00 AM

8:30 AM

9:00 AM

Figure 3.12: Surge ratios of the throughput-maximizing spatial pricing during the morning peak.

pricing scheme lowers the percent of vehicles in idle during most of the time except for the peak period, where the value stays constantly higher than the corresponding level at the status quo. The same contrast is again captured in Figure 3.12, which presents four representative instances of price-surge maps with darker colors marking higher surge ratios. As can be seen, the platform may still lower the price in most of the time and areas to increase the system throughput. However, for the peak period, a few regions in the central downtown need price surges.

Since lower prices are suggested in most cases by the above implementation, the service revenue of the platform will not increase despite the efficiency gains. As drivers expect higher revenue to offset the wear and tear caused by additional service provision, the enhancements come at the cost of a platform aggressively subsiding the system, which may not be sustainable in the long run. A mechanism is thus needed to fairly allocate the costs and benefits of efficiency gains among stakeholders, making the schemes more competitive for implementation.

3.6.2 Revenue-maximizing pricing policies

We proceed to investigate another case where the platform optimizes the pricing policy aiming to maximize the total service revenue. Two performance metrics are compared in Figure 3.13 for the four schemes. The blue columns now represent the total service revenue of the ride-hailing market, while the orange polyline presents the total number of requests fulfilled. As clearly shown by the figure, the platform can benefit substantially from higher levels of pricing differentiation. Only allowing price discrimination in time could give rise to a 22.4% increment in service revenue, while the optimal OD-specific pricing yields 27.6% more revenue than the current fixed-price condition. On the other hand, the number of customers served does not necessarily increase amid

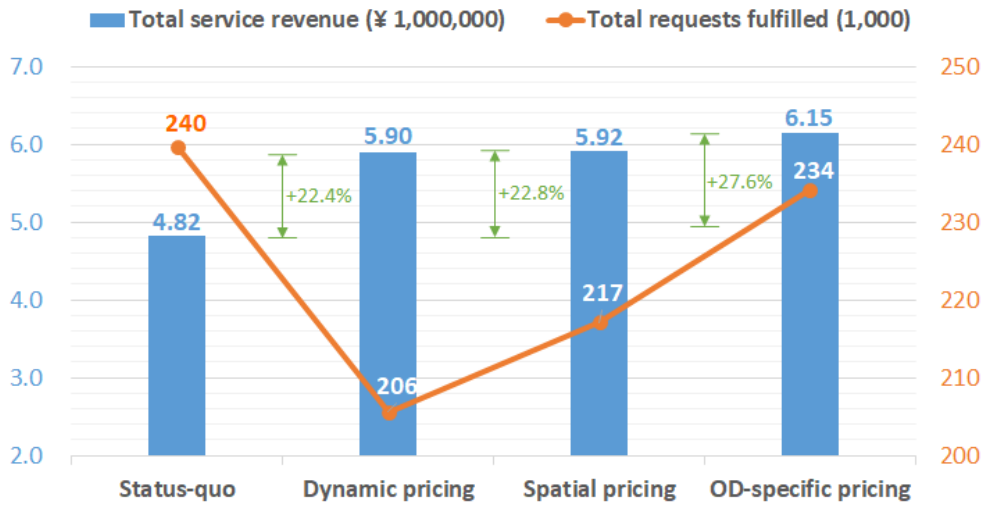


Figure 3.13: Comparison of the optimal pricing schemes under revenue maximization

the revenue boosts. As we can see from the figure, by allowing a revenue-maximizing platform to implement time-dependent pricing could incur a 14% reduction in the number of requests fulfilled. The platform can strategically raise the price to serve fewer customers but receive higher revenue. In fact, in our analyses the platform would opt to price at the maximum surge ratio in most of the time and areas with only a few exceptions. our counterfactual analysis suggests that the current pricing structure adopted by the platform does not maximize its revenue for this particular market and doing so via price differentiation in time and space will compromise social welfare and consumers' surplus of the market.

3.7 Summary

This chapter presents a type of matching failure, named as "countryside delivery", that can arise in anisotropic ride-sourcing markets due to the indiscriminate matching between riders and drivers. A two-node network is designed to showcase the existence of such a matching failure as well as its consequences of causing inefficient supply states. We point out that such a matching failure can possibly arise in a wide variety of circumstances in real-world operations of ride-hailing services, but can be solved through price or rationing discrimination. To cope with the need for real-world operations, we further propose a generalized fluid framework for modeling ride-hailing systems and demonstrated its applicability by customizing and calibrating the framework for a real system operated by DiDi Chuxing. The customized model was further applied to prescribe optimal pricing policies for the system. The results reveal the potentials of pricing policies for improving the performance of the system. The model can be further applied to explore other operational strategies such as matching and dispatching.

We believe that the proposed fluid framework has made a good balance between mathematical tractability and behavioral realism to allow service providers to answer what-if questions quickly and optimize their strategic operation decisions. The framework can also be further extended to capture the interaction between ride-hailing services with other modes of transportation such as private cars and transit and facilities such as parking and dedicated pick-up/drop-off locations for ride-hailing services. An important consideration in this extension is to capture the congestion effect of in-service ride-hailing vehicles, which can be achieved by leveraging recent developments in network macroscopic fundamental diagrams, as demonstrated by, e.g., Xu et al. (2017) and Ramezani and Nourinejad (2018). The resulting framework will be able to support the planning practice and policy making of governmental agencies in applications such as regulation of ride-hailing services, congestion pricing, and leveraging ride hailing to solve the first mile/last mile problem of transit systems.

Driver Behaviors

Unlike traditional firm-employee relationship, the ride-sourcing market features more flexible labor supply (Chen and Sheldon, 2016). Drivers now are free to determine where, when and how long they want to work. Some may work full time like professional taxi drivers while others only provide service for limited hours (Zha et al., 2018a). However, the flexibilities enjoyed by ride-sourcing drivers pose challenges for ride-sourcing companies to manage the system, either because of the limited information drivers possess or their individual motivation for revenue maximizing. Accordingly, ride-sourcing companies may not be able to attract proper labor supply despite the accurate prediction on labor-constrained days or areas with demand surges. Therefore, there is an increasing need for developing a framework to learn and predict drivers' service behavior in ride-sourcing markets. The overarching goal of Chapter 4 and 5 is to better comprehend drivers' decision making of their service areas using data from Didi Chuxing. These empirical results could help ride-sourcing companies design strategies and mechanisms to influence drivers' labor supply, reduce their search costs and improve the system operations.

CHAPTER 4

"Search under the nose"

Downside of Freedom

4.1 Introduction

While the great success of ride-sourcing services has triggered heated discussions, less attention has drawn on the behavior of ride-sourcing drivers, partially due to the confidentiality of service data. By virtue of comprehensive empirical data from Didi Chuxing, this chapter aims to comprehend the behavior of ride-sourcing drivers in customer search, which constitutes a significant portion of drivers' service time and strongly associates with their profitability.

Undoubtedly, the customer search behavior of ride-sourcing vehicles shares a great similarity with that of traditional taxis, which has already been studied extensively in empirical literatures. Szeto et al. (2013) and Wong et al. (2014c) first applied the Multinomial Logit (MNL) model to capture strategic zonal choices of vacant drivers' customer search. Wong et al. (2015b) then enhanced the MNL framework to a sequential Logit model to consider the opportunities along the search paths. Meanwhile, Wong et al. (2014b) proposed a cell-based Logit-opportunity model to tackle the local customer search behavior of taxis. A two-stage approach was then developed by Wong et al. (2015a) by combining both the far-distance zonal choices and local search components. Qin et al. (2017) applied a generalized multi-level ordered Logit model on the search-delivery records of nearly 8,000 taxis in Shanghai to identify the significant factors that impact drivers' income. Recently, Tang et al. (2019) argued that between different destination choices of vacant taxis, there are substantial overlaps in paths, which invalidate the use of MNL models. Instead, they proposed a mixed path size Logit-based customer search model and tested its effectiveness in predicting routing choices over the trajectory of 36,000 taxis in Beijing. Although these static search models substantially facilitate empirical calibration, they fall short in dealing with highly time-varying market conditions and the time effects in drivers' choices. Besides, one of the major difficulties in calibrating the search behavior is that drivers' trajectories do not fully reflect their real preferences. It is common for drivers to get matched to passengers before reaching the actual

cruising destinations, especially in app-based ride-hailing markets. Sometimes, drivers do not even have specific search destinations in mind. Therefore, we are in need of a behavioral model that can cope with the intense market variations and identify drivers' latent search patterns with modeling differentiation.

An alternative way of modeling drivers' customer-searching movement is to formulate it as Markov decision process (MDP). Oftentimes, an MDP framework is coupled with learning approaches to seek the optimal searching policy for idle drivers. Recently, Liu et al. (2013), Verma et al. (2017), Gao et al. (2018), and Lin et al. (2018) investigate the optimal dynamic routing strategy of drivers through Q-learning, where the intervening opportunities are incorporated implicitly into the action rewards. Besides, Qu et al. (2014), Rong et al. (2016), Yu et al. (2019a), and Shou et al. (2020) specified structured reward functions based on various zonal features and then solved the problem using dynamic programming. However, the weights of different features were exogenously given to concretize certain search objectives. In general, these learning-based approaches aim at the optimal policy of repositioning drivers in view of centralized platforms, and lack the behavioral implications of individual drivers.

To fill these research gaps, a dynamic discrete choice model is developed in this chapter to investigate the customer-searching behavior of ride-sourcing drivers. The model translates market conditions, including both supply and demand information, into a spatiotemporal continuum of opportunity values to idle drivers. Then, by adopting an absorbing MDP framework, we formulate and evaluate drivers' manifold decisions underlying customer search. To foster a comprehensive understanding, heterogeneous behaviors among different driver types and time of day are further explored and compared. It is worth noting that our model differentiates three universal modes of search movements, respectively being staying motionless, cruising around without target areas, or repositioning towards specific destinations. These movements, although corresponding to completely different mentality of drivers, are challenging to separate through trajectory data. Indeed, to the best of our knowledge, none of the previous studies have investigated or calibrated the driver behavior in this regards. Leveraging large-scale empirical datasets (with uninterrupted trajectories of 32,000 drivers and transaction information of 5 million trip requests), this chapter is dedicated to deepening our understanding of ride-sourcing drivers' customer search behaviors and provide policy insights for the platform's supply management.

The remainder of this chapter is organized as follows. Section 4.2 details the dynamic discrete choice model designed for learning ride-sourcing drivers' customer-searching behavior, while Section 4.3 illustrates the data preparation for model calibration. Section 4.4 presents the results of parametric estimation and then interprets the behavioral implications of drivers in zonal search. At the end, Section 4.5 concludes the chapter and point out the future research avenue.

4.2 Dynamic Discrete Choice Model

This section first introduces the dynamic discrete choice model that we formulate in line with ride-sourcing drivers' customer-searching movements. We first dissect the process of how idle drivers search for customers, and then formulate a mathematical model to delineate drivers' sequential choice-making. The method for parametric estimation is also discussed.

4.2.1 Drivers' customer-searching movements

When ride-sourcing drivers are idle, they enjoy full freedom to decide where and how to search for the next customer based on their perception of the market condition. They may either remain motionless with the satisfaction of local matching opportunities, cruise around the neighborhood to actively search for customers, or reposition themselves towards target hotspots. To cope with these different ways of customer search, we materialize idle drivers' movements as cyclic decision-making by segmenting them into series of steps. Within each step, a driver either stays in the current zone, or chooses his/her next destination from the finite set of zones adjacent to the current location, and then moves to that target zone. A series of decisions are made sequentially by a searching driver along a chain of connecting zones, until he/she is successfully matched with a customer (see Figure 4.1 for a graphical explanation).

For the convenience of behavioral analysis, we further treat the sequential movements as Markov decision processes, assuming that drivers' decision-making at each step is independent of her previous choices. Notwithstanding, drivers plan for several steps ahead when making each decision. Note that such a setting adheres well with that of the dynamic discrete choice theory

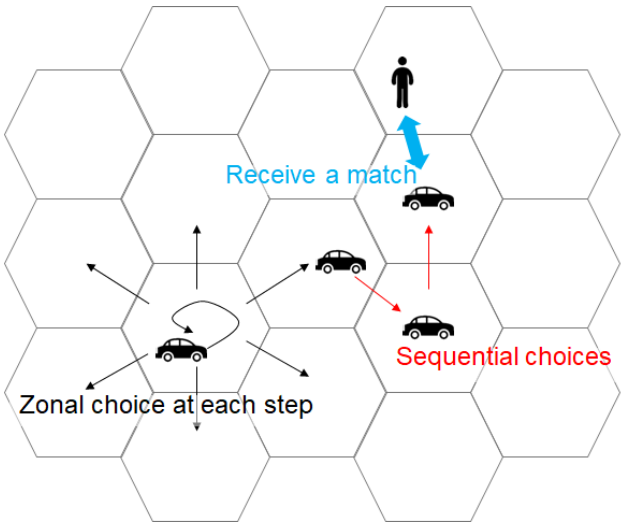


Figure 4.1: Sequential movements of idle drivers in customer search

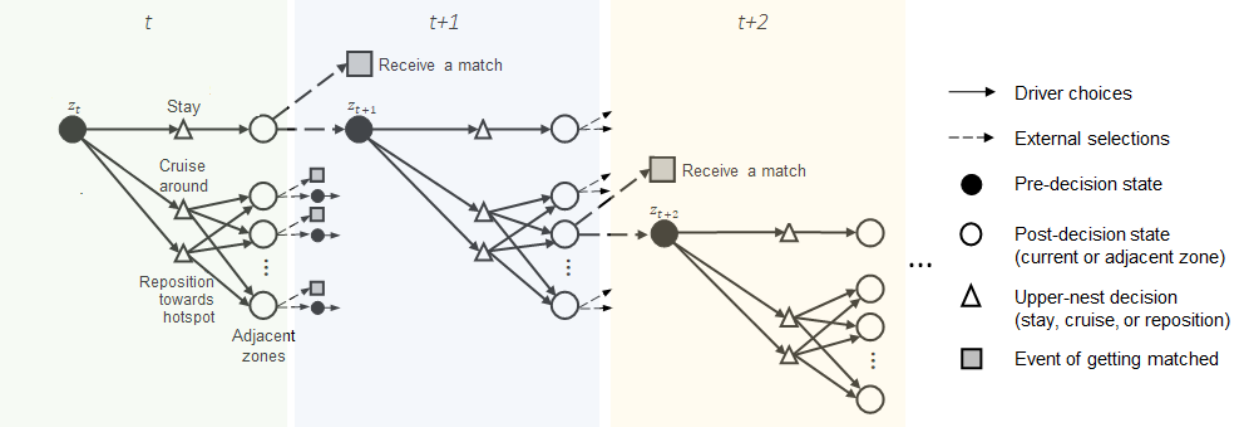


Figure 4.2: Decision-making processes of ride-sourcing drivers

(Rust, 1987). Thus, we introduce an absorbing Markov chains in this chapter to describe the dynamic discrete choice of drivers under finite evaluation horizons.

Figure 4.2 illustrates the cross-nested structure of drivers' sequential searching choices. During the process, a driver at each stage can select to either rest for a while in the current zone or move to one of the adjacent zones, based on the expected utility of each choice. The process continues as the cumulative probability of being matched increases along the trajectory. Within each stage, the upper nest indicates three scenarios for drivers' zonal transition: to either remain motionless, cruise nearby without specific destinations, or reposition toward hotspot areas, while the lower nest specifies the potential destinations under the corresponding scenario. Cater to such a cross-nested structure, our model differentiates the latent searching scenarios and describes the search movements of drivers more precisely.

4.2.2 Model formulation

We formulate the model based on the multinomial-type dynamic discrete choice model proposed by Rust (1987). Let z_t denote the location of a driver at time t, while $v(z_t)$ and $\varepsilon(z_t)$ respectively be the observed value and unobserved error associated with the corresponding choice. To facilitate understanding, we first present a base model for the non-nested scenarios and then extend it to fit with the cross-nested contexts. The driver-specific indicators are omitted in the expressions below for clarity.

Base model

With a non-nested choice structure, the anonymous value function $V^T(z_t)$ under a time horizon

of T can be formulated as

$$V^T(z_t|\theta) = \frac{1}{\mu} \cdot \mathbb{E}\left[\max_{z_{t+1} \in C(z_t)} \left(\mu\left(v\left(z_{t+1}|z_t;\theta\right) + \rho_{z_t}\beta V^T\left(z_{t+1};\theta\right)\right) + \varepsilon\left(z_{t+1}\right)\right)\right]$$

where θ represents a parametric vector; μ is a scaling factor ($\mu \geq 1$), indicating the degree of independence among the alternatives' unobserved confounders; $C(z_t)$ denotes the choice set of drivers at z_t; ρ_{z_t} is the probability of drivers remaining unmatched at z_t; and β denotes a time-discounting factor ($0 \leq \beta \leq 1$). Note that ρ_{z_t} in the equation acts similarly as the survival probability in a general dynamic programming (Rust, 1987). The value iteration terminates at the end of each horizon with $V^T(z_T|\theta) = 0$.

Assuming that the error term ε follows a standard Gumbel distribution yields the following choice probability (with parameter θ omitted for clarity),

$$P^T\left(z_{t+1}|z_t\right) = \frac{\exp\left(\mu\left(v\left(z_{t+1}|z_t\right) + \rho_{z_t}\beta V^T\left(z_{t+1}\right)\right)\right)}{\sum_{z \in C(z_t)} \exp\left(\mu\left(v\left(z|z_t\right) + \rho_{z_t}\beta V^T\left(z\right)\right)\right)}, \qquad \forall\, z_{t+1} \in C\left(z_t\right)$$

where the expected value $V^T(\cdot)$ of each choice is given as

$$V^T\left(z_t\right) = \frac{1}{\mu} \cdot \ln \sum_{z_{t+1} \in C(z_t)} \exp\left(\mu\left(v\left(z_{t+1}|z_t\right) + \rho_{z_t}\beta V^T\left(z_{t+1}\right)\right)\right)$$

The choice probability $P^T(z_{t+1}|z_t)$ is then utilized as a surrogate of the state transition probability from z_t to z_{t+1}, following the same treatment adopted by the recursive logit model (Fosgerau et al., 2013). It is worth noting that the time horizon T is incorporated to avoid the state explosion of a time-expanded network amid the transient market conditions. Behaviorally, it could be seen as the cognitive limitation of drivers in customer search. Then, with each specific time horizon T given, the value of V^T can be calculated via backward induction.

Nested model

We then extend the above base model to accommodate the context of a nested structure (with no cross alternatives), where drivers first make an upper-nest decision among stay, cruise, or reposition, and then select a target out of the zonal nest to move one-step forward. The value function under the nested choices is reformulated as follows,

$$V^T\left(z_t\right) = \mathbb{E}\left[\max_{l \in C_u(z_t)} \left(\epsilon\left(l\right) + \max_{z_{t+1} \in C(z_t,l)} \left(v\left(z_{t+1}|z_t\right) + \rho_{z_t}\beta V^T\left(z_{t+1}\right) + \epsilon\left(z_{t+1},l\right)\right)\right)\right]$$

where l is the choice in the upper nest $C_u(\cdot)$. Let $\epsilon\left(z_{t+1},l\right) = \frac{1}{\mu_l}\varepsilon\left(z_{t+1}\right)$ and the error term $\varepsilon\left(z_{t+1}\right)$ follow a standard Gumbel distribution. The value function and inclusive (log-sum) utility v_u are

given by

$$V^T(z_t) = \mathbb{E}\left[\max_{l \in C_u(z_t)} \left(\epsilon(l) + v_u(l|z_t) + \epsilon_u(l)\right)\right]$$
$$v_u(l|z_t) = \frac{1}{\mu_l} \ln \sum_{z_{t+1} \in C(z_t,l)} \exp\left(\mu_l \left(v(z_{t+1}|z_t) + \rho_{z_t} \beta V^T(z_{t+1})\right)\right), \quad \forall\, l \in C_u(z_t)$$

Again, assuming $\epsilon(l) + \epsilon_u(l) = \frac{1}{\mu_n}\varepsilon(l)$ with $\varepsilon(l)$ following a standard Gumbel distribution yields

$$V^T(z_t) = \frac{1}{\mu_n} \ln \sum_{l \in C_u(z_t)} \exp\left(\mu_n v_u(l|z_t)\right), \quad (\mu_l \geq \mu_n \geq 1) \tag{4.1}$$

Additionally, the choice probability exhibits the following multiplicative form,

$$\begin{aligned} P^T(z_{t+1}|z_t) &= P^T(z_{t+1}|l) \cdot P^T(l|z_t) \\ &= \frac{\exp\left(\mu_l \left(v(z_{t+1}|z_t) + \rho_{z_t} \beta V^T(z_{t+1})\right)\right)}{\sum_{z \in C(z_t,l)} \exp\left(\mu_l \left(v(z|z_t) + \rho_{z_t} \beta V^T(z)\right)\right)} \cdot \frac{\exp\left(\mu_n v_u(l|z_t)\right)}{\sum_{k \in C_u(z_t)} \exp\left(\mu_n v_u(k|z_t)\right)} \end{aligned}$$

Cross-nested model

Further extensions for a cross-nested model is obtained through the generating function. In the cases of nested logit (NL) and cross-nested logit (CNL), the generating functions are respectively given as follows (Train, 2009),

$$\text{NL}: G(\boldsymbol{Y}) = \sum_{l=1}^{K} \left(\sum_{z \in B_l} Y_z^{\mu_l}\right)^{\frac{1}{\mu_l}}$$
$$\text{CNL}: G(\boldsymbol{Y}) = \sum_{l=1}^{K} \left(\sum_{z \in B_l} (\alpha_{z,l} Y_z)^{\mu_l}\right)^{\frac{1}{\mu_l}}$$

where $\alpha_{z,l}$ is an allocation parameter that reflects the likelihood of alternative z being a member of nest B_l with $\alpha_{zl} \geq 0$ and $\sum_l \alpha_{zl} = 1$; the function Y_z characterizes $\exp(V(z))$ in both models, where $V(z)$ stands for the value of choice z. By comparison, we have $V^T(z_t)$ continue to hold as Eq. (4.1) and the log-sum (inclusive) utility v_u under the cross-nested logit be respecified as

$$v_u(l|z_t) = \frac{1}{\mu_l} \ln \sum_{z_{t+1} \in C(z_t,l)} \left(\alpha_{z_{t+1},l} \exp\left(v(z_{t+1}|z_t) + \rho_{z_t} \beta V^T(z_{t+1})\right)\right)^{\mu_l}, \quad \forall\, l \in C_u(z_t) \tag{4.2}$$

and the choice probability now sums over all the probability multiplications,

$$P^T\left(z_{t+1}|z_t\right) = \sum_{l\in C_u(z_t)} P^T\left(z_{t+1}|l\right)\cdot P^T\left(l|z_t\right) \tag{4.3}$$
$$= \sum_{l\in C_u(z_t)} \frac{\left(\alpha_{z_{t+1},l}\exp\left(v\left(z_{t+1}|z_t\right)+\rho_{z_t}\beta V^T\left(z_{t+1}\right)\right)\right)^{\mu_l}}{\sum_{z\in C(z_t,l)}\left(\alpha_{z,l}\exp\left(v\left(k|z_t\right)+\rho_{z_t}\beta V^T\left(k\right)\right)\right)^{\mu_l}}\cdot\frac{\exp\left(\mu_n v_u\left(l|z_t\right)\right)}{\sum_{k\in C_u(z_t)}\exp\left(\mu_n v_u\left(k|z_t\right)\right)}$$

with $\mu_l \geq \mu_n \geq 1$.

4.2.3 Model estimation

All parameters in the model Θ can be estimated by maximizing the following log-likelihood (LL), i.e.,

$$\mathrm{LL}(\Theta) = \sum_{i,t}\delta_{i,t}\left(z_{t+1}|z_t\right)\cdot\log P^T\left(z_{t+1}|z_t;\Theta\right)$$

where $\delta_{i,t}$ indicates the choice made by individual driver i at time t. Student's t-tests are conducted to verify the significance of parametric estimates in the model, while Watson and Westin pooling tests are applied for model comparisons (Watson and Westin, 1975).

4.3 Data Description

The two datasets that we use for model calibration respectively contain the transaction information of trip requests and the trajectory information of drivers. They were both collected by Didi Chuxing from one medium-sized city in China, spanning over the 10 weekdays in August 7-18, 2017.

The trajectory data comprises the spatiotemporal records of 32,000 DiDi drivers. Each record characterizes a timestamp, the longitude/latitude, as well as the service state of a driver at that moment. The label of service state differentiates whether a driver is idle (waiting to be matched), deadheading (picking up a customer), or occupied (delivering a customer to the destinations). The entire city is partitioned into regular hexagonal lattices, each gets side-connected with six adjacent ones. The major advantage for hexagonal partition is that each unit has six symmetrically equivalent and unambiguous side-adjacent neighbors, while square partition results in two different types of neighboring, respectively being side-connected or corner-connected. We map the latitude/longitude sequences to the 660-meter-side-length hexagonal lattices (Sahr et al., 2003) to produce the base sample for movement identification. The case when a driver remains in one specific zone for more than a time threshold δ is regarded as a "stay", while the case when a driver reposition to a neighboring zone within δ is taken as a "move". Those search trajectories

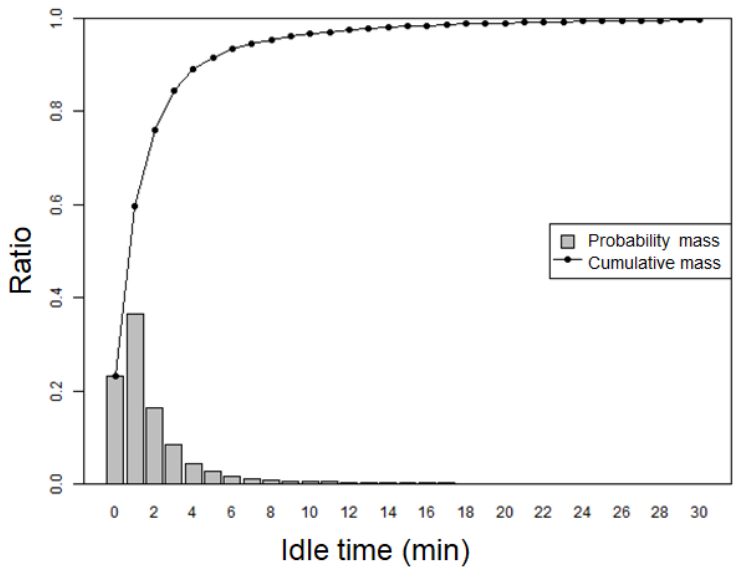
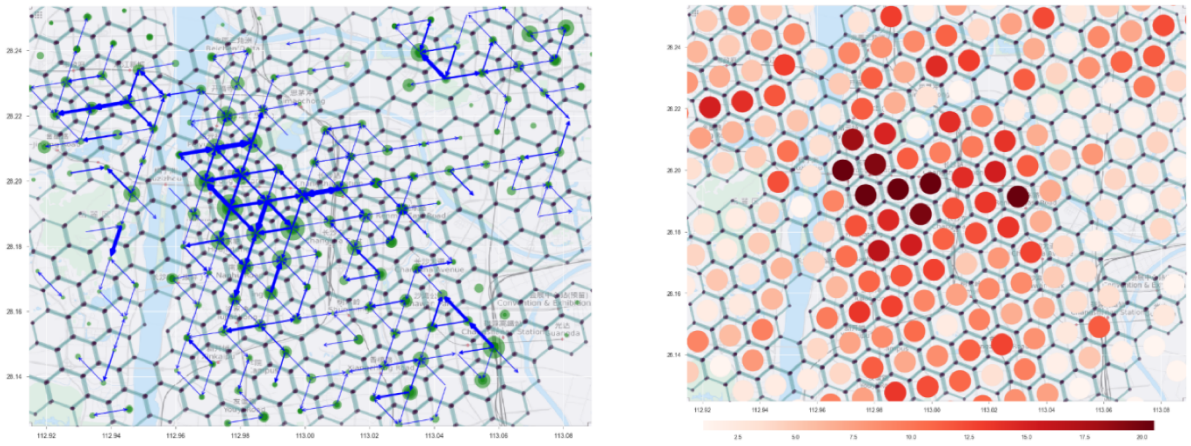


Figure 4.3: Histogram of drivers' idle time staying in one zone

without any zonal transitions or with a total duration less than δ minutes are removed from our sample. Figure 4.3 plots the distribution of drivers' consecutive idle time of staying at one single zone. We finalize the selection of δ to be 4 min to filter out the time needed by passing-through drivers waiting for traffic signals.

The transaction dataset contains the information of approximately 5 million request records. Each record comprises the timestamp when the trip gets requested, the matching/pickup/delivery time of the passenger, the origin/destination of the trip, the driver in service, and the trip fare. The transaction information is then aggregated by hexagon zones and time intervals. It is worth noting that in this study we do not differentiate between the days in a week. To overcome the sparsity of data under the high-granularity partition, all the orders fall in the same time interval across different days are merged together to create the within-day explanatory variables. Consequently, the coefficient estimates in this study likely reflect the behavior of drivers in response to the regular market variations from day to day (or say, drivers' long-term learning experiences), rather than the transient fluctuations.

As an overview of our datasets, Figure 4.4 visualizes the spatial distribution of the drivers' movements and orders requested. In Figure 4.4a, the blue arrow represents the number of samples moving between two adjacent zones ("move" action), while the green circle refers to the number of samples staying in the current zone ("stay" action). Thicker arrows imply a greater number of "move" actions, while larger circles stand for more "stay" actions. Figure 4.4b shows the heat map of the number of orders averaged over all the 5-min intervals for each hexagonal zone. The deeper color represents more orders generated. It can be observed that the zones with the higher



(a) Drivers' movements (b) Order requests

Figure 4.4: Spatial distribution of drivers' search trajectories and orders requested

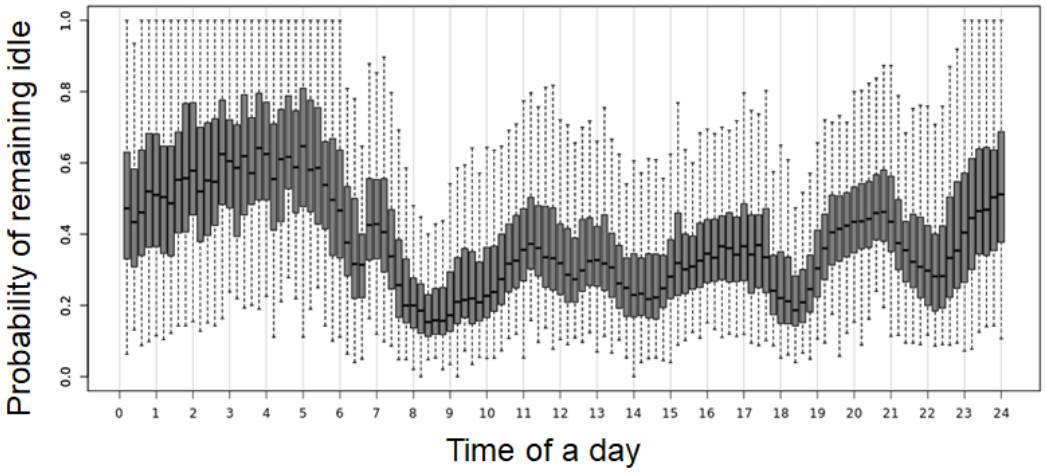


Figure 4.5: Retention ratio of idle drivers in each 12-min interval throughout a day

density of requests generally stay more attractive to idle drivers. Figure 4.5 presents a boxplot of the retention ratio of idle drivers $\rho_z^t (= 1 - MP_z^t)$ in each 12-min interval of a day, where the bold line shows the median and the dark gray boxes denote the 25th-75th quantile ranges. The statistics on the plot are taken over the retention ratios of all zones within each interval. As shown by Figure 4.5, the matching probability of drivers changes drastically throughout the time of a day, thereby indicating the applicability of the dynamical discrete choice model we propose.

In addition, based on the cumulative working duration, drivers in the study are divided into two classes, respectively working in full-time and part-time. Full-time drivers are those with the most extended service hours but also the highest earners and most profitable ones. They are mostly daytime workers and display very little within-group heterogeneity in terms of profitability. In contrast, part-time drivers are active for less amount of time and differ greatly in their driving and

working strategies, which affects their profitability and earnings.

4.4 Model Estimation and Behavioral Interpretation

A set of explanatory variables are generated from the above datasets and then fed into the dynamic discrete choice model for parametric estimation. In this section, we base on the coefficient estimates to interpret drivers' factorial focuses underlying the customer search movements.

4.4.1 Parametric setup

The following explanatory variables are generated to represent the variable market condition:

- Trip fare TF_z^t: The average trip fare of orders originated in zone z during time interval t;
- Number of requests NR_z^t: The total number of requests sent out in zone z during time interval t;
- Pickup time PT_z^t: The average pickup time of passengers requesting for trips in zone z during time interval t;
- Matching probability MP_z^t: The ratio of idle drivers receiving matches in zone z during time interval t;

To be consistent with the movement threshold δ, the time granularity for the four market variables are set as 4 minutes. The first three variables are further normalized to distributions with zero mean and unit standard deviation. It is also verified that the partial correlation among the explanatory variables is fairly low, and collinearity should not be a concern here. All these variables are then invited to construct the observed utility $v\left(z_{t+1}|z_t\right)$ in Eq. (4.2) as follows,

$$v\left(z_{t+1}|z_t,\theta\right) = \theta_{TF}TF_{z_{t+1}}^t + \theta_{NR}NR_{z_{t+1}}^t + \theta_{PT}PT_{z_{t+1}}^t + \theta_{PM}PM_{z_{t+1}}^t + \theta_{SZ}\mathbb{1}(z_{t+1}=z_t)$$

where coefficient $\boldsymbol{\theta}$ represents drivers' sensitivity to different factors; Specifically, θ_{SZ} is a movement constant that indicates drivers' preference for remaining motionless when idle.

To better capture the strategic movements of idle drivers, we introduce another variable - Distance to the hotspot DH_z^h - that calculates the Euclidean distance from the centroid of zone z, where the driver stays, to that of the hotspot h. The hotspot areas are carefully selected in each of the five periods of a day, i.e. morning (6AM-10AM), daytime (10AM-4PM), evening (4PM-8PM), night (8PM-11PM), and midnight (11PM-6AM). For each zone, we sum up the number of requests by 20-min intervals, and then calculate the 80^{th}-percentile count for each period. A zone

can be identified as a hotspot only if the gap between the 80^{th}-percentile and the maximum counts is less than 10%. The selected hotspots are mostly the downtown areas, the commercial areas in the suburb, and the railway stations, which are further classified into the downtown-hotspot area (DTH) and the outer-hotspot area (OUH). Our cross-nested logit model treats these two hotspot categories separately in the upper nest, and the variable DH_z^h calculates the distance to the closest hotspot in each category. Note that the outer-hotspot category only exists from 10AM to 11PM. The allocation parameter $\alpha_{z_t,l}$ in Eq. (4.3) are specified as follows:

$$\alpha_{z_t,h} = \frac{\exp\left(\gamma_{DH} DH_{z_t}^h\right)}{S}, \quad \forall\, h \in \{\text{DTH, OUH}\}$$
$$\alpha_{z_t,c} = \frac{\exp\left(\gamma_{NR} NR_{z_t}^t\right)}{S}$$
$$S = \exp\left(\gamma_{DH} DH_{z_t}^{\text{DTH}}\right) + \exp\left(\gamma_{DH} DH_{z_t}^{\text{OUH}}\right) + \exp\left(\gamma_{NR} NR_{z_t}^t\right)$$

where $\alpha_{z_t,h}$ and $\alpha_{z_t,h}$ denote the weights for the target-specific repositioning and aimless-cruising movements, respectively.

The choice set of drivers C at each zone is extracted from the observations. Meanwhile, the sample is split evenly into two for the purpose of learning and cross validation, respectively. We track the cross-validation log-likelihood (CVLL) constantly in case of possible overfitting.

4.4.2 Model refinement

With the value functions specified, we then refine the selection of critical parameters and structures in the model, i.e. the horizon of evaluation T, the probability of drivers remaining idle ρ_{z_t}, and the nest structure.

Table 4.1 compares the models under a set of different evaluation horizons T, based on the sample of full-time drivers. As shown in the table, the CVLL follows the same trend as the LL of model estimations, relieving the concerns for overfitting. The LL for the model when $T = 0$ (a static model) appears the lowest, implying the relative superiority of our proposed dynamic choice model. Besides, the LL increases constantly as T grows from 0 to 2, and stays constant afterwards. Considering the computational efficiency, we adopt $T = 2$ in the final implementation to account for the ahead-planning behavior of drivers in customer search.

Table 4.1: Model comparisons for different evaluation horizons T

Horizon T (step)	0 (static)	1	2	3	4	5
LL	-112311	-112235	-112221	-112220	-112220	-112220
CVLL	-109803	-109743	-109726	-109724	-109724	-109724

Table 4.2: Model comparisons for juxtaposed nest structures and granularities of ρ_{z_t}

Nest structure	Cross-nested			Nested	Plain
ρ_{z_t}'s updating frequency	4 min	2 h	Never	4 min	4 min
LL	-112221	-112266	-112261	-112253	-112875
CVLL	-109726	-109758	-109755	-109761	-110440

Table 4.2 compares the models with different sampling granularity of ρ_{z_t} and juxtaposed nest structures. First, for the cross-nested case, we present three models where ρ_{z_t} gets updated every 4 minutes, 2 hours, and never (by fixing ρ_{z_t} to 1), respectively. According to the LL, the model with the most frequently updated ρ_{z_t} works the best. Then, we zoom out to compare the modeling effectiveness of different nest structures. The nested logit model keeps only the stay/leave options in the upper nest (therefore, with no crossings in the zonal alternatives), while the plain structure practices the most basic logit model. The utility specifications of both models include the hotspot variables DH_z^h by linear combinations. Not surprisingly, the cross-nested model outperforms the other two and is thus selected for our later analyses.

4.4.3 Full-time versus part-time drivers

Based on the cross-nested model, we apply the Watson and Westin (1975) pooling test to examine whether full-time and part-time drivers in general behave differently in customer search. The restricted model applies a unified set of coefficients for both types of drivers, while the alternative model specifies two separate sets of coefficients. The likelihood-ratio (LR) test statistic is significant at the 1% level, which confirms the heterogeneous search behavior among different types of drivers. The specific parametric estimation as well as the significance of each estimate are presented in Table 4.3. The left two columns summarize the estimation results of coefficient Θ for the full-time driver group, while the right-side columns present the comparative differences $\Delta\Theta$ of part-time drivers in responding to the various factors.

As shown by Table 4.3, all the coefficient estimates $\hat{\Theta}$ corresponding to both full-time and part-time drivers are significant at the 1% level, with intuition-consistent signs. This implies an encouraging fact that in general ride-sourcing drivers do respond actively and positively to the repetitive market variations. The only counter-intuitive result is that part-time drivers tend to drive to the area with longer pickup time for passengers. In fact, different from the full-time drivers, part-time drivers may not be sensitive to the potential pickup time and thus should not purposely quest for long pickups. A possible reason causing such a correlation could be that part-time drivers prefer the areas where they could match to customers in a wider space (higher matching opportunity but longer pickup time), or the areas with more condensed customer demand but also more congestion in usual. Besides, as revealed specifically by our model, drivers' search

Table 4.3: Coefficient estimates of a cross-nested model with two types of drivers

Driver type	Full-time Θ		Part-time $\Delta\Theta$	
	Coeff.	*t*-statistic	Coeff.	*t*-statistic
Upper nest				
DH	-3.68	8.00∗	0.51	0.74
NR	0.52	8.46∗	-0.07	0.78
Lower nest				
TF	0.02	2.40†	0.03	1.77
NR	0.08	37.02∗	-0.01	3.77∗
PT	-0.02	3.11∗	0.10	8.27∗
PM	1.00	18.29∗	-0.40	4.30∗
Constant (Stay)	1.70	199.62∗	0.09	5.45∗
Discount β	0.31	23.08∗	-0.14	5.30∗
Scaling μ_l	1.74	68.90∗	0.17	3.16∗
Observations				249188
LL(0)				-277810
Final LL				-165264
CVLL				-161942
Adjusted ρ^2				0.41

Notes: The column of part-time drivers presents the difference in drivers' response compared to that of full-time drivers on the left.

† - Significance to the 5% level;

∗ - Significance to the 1% level.

movements are not confined only to the local considerations. Instead, they show a clear tendency of repositioning towards the hotspots, which strengthens significantly as they move closer to the hotspot. It is highly probable that drivers take the distant hotspots as back-up options given the certainty of receiving quick matches therein. But as a result, all the supply of idle drivers at the neighborhood of a hotspot might be drained up, causing deceptive supply shortage at local regions.

While drivers of different groups hold consistent preferences for the various factors in customer search, there is a significant disparity between full-time and part-time drivers in response sensitivity. In contrast to full-time drivers, drivers work in part-time behave much less sensitive to the number of requests and the matching probability in zonal search. This is consistent with our intuition that full-time drivers are more experienced in service provision and can thus respond scrupulously under different circumstances. Interestingly, part-time drivers apply to a significantly lower discount factor β than full-time drivers. In accordance to the function of β, this implies that full-time drivers are more far-sighted by planning ahead, while part-time drivers focus more on the near-future opportunities. Meanwhile, the scale parameter μ_l for full-time drivers is significantly lower, meaning that their customer search movements are dictated more strongly by the unobserved

confounders.

We then proceed to examine whether drivers' search behavior varies in time by allowing the coefficients to change by periods of time. To ensure the validity of calibration over the segmented datasets, we fix the discounting factor β and allocation γ in each sub-model to the value estimated previously using the full sample. Figure 4.6 displays the parametric estimates of full-time (red line) and part-time (blue line) drivers across the time of a day, respectively. On each parameter, the estimated $\hat{\Theta}$ corresponding to the daytime period (10:00 AM - 4:00 PM) is set as the base, while the estimated relative deviations $\Delta\hat{\Theta}$ of the other time periods are tested. Each dot represents the estimated value of $\hat{\Theta}$ for different periods, while the error bar around shows the standard deviation of $\Delta\hat{\Theta}$. According to Figure 4.6, the two classes of drivers again exhibit very similar temporal patterns in response to various factors. They both show higher preference on the trip

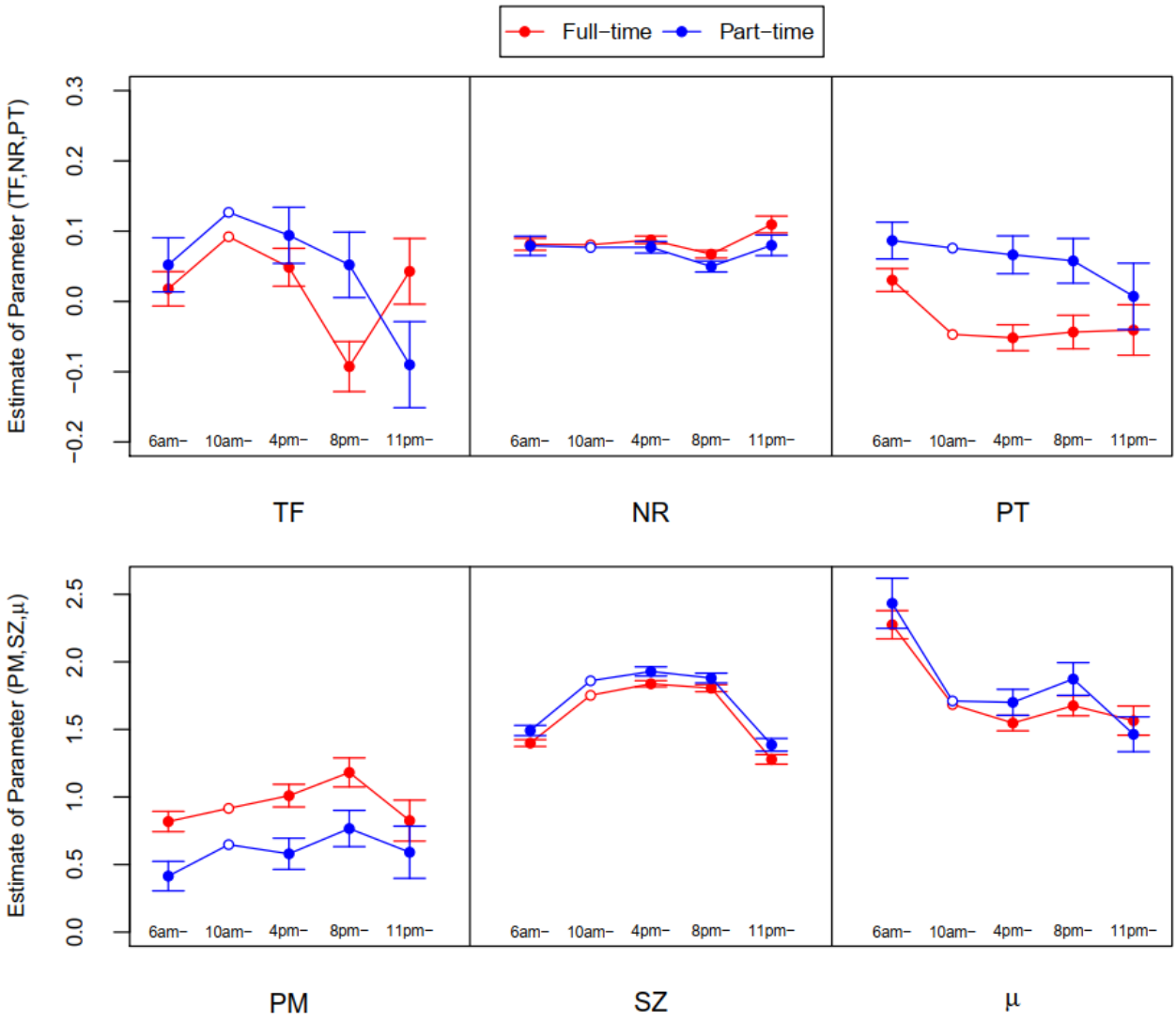


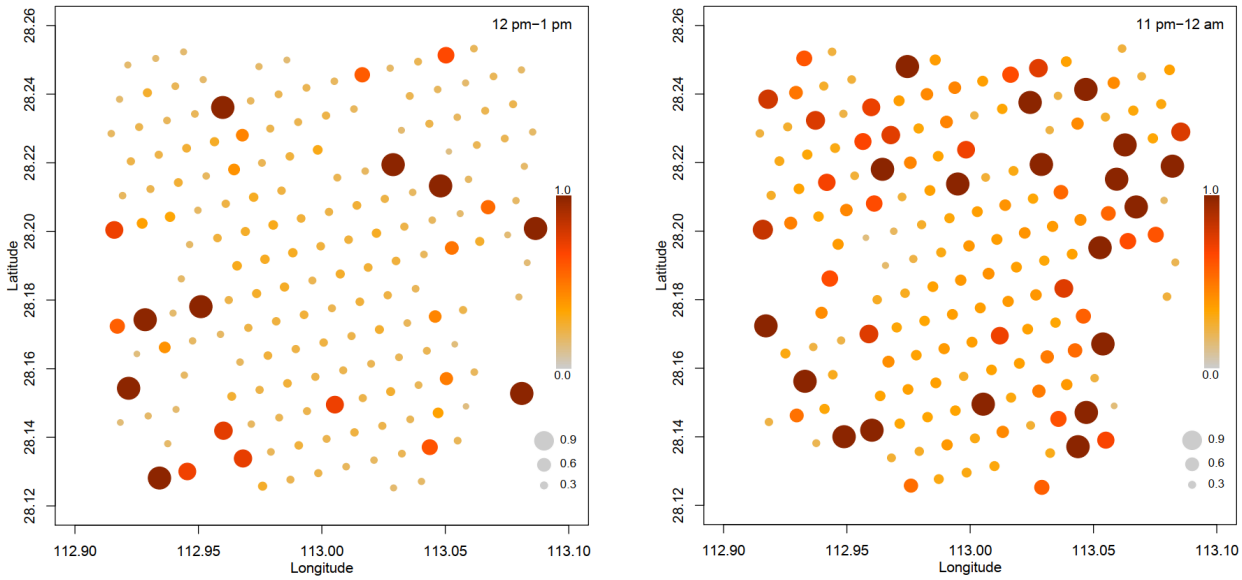
Figure 4.6: Comparisons of drivers' behavioral responses across the time of a day

fare during the daytime, while focusing more on the matching probability in the evening. Such a behavioral transition adheres with the nature of ride-sourcing markets, as the travel demand becomes much more sparse and heterogeneous spatially after the evening peak, when drivers need to switch searching goals to first secure the chances of getting matched. Meanwhile, the estimates of the movement coefficient θ_{SZ} indicate that drivers prefer to stay motionless in the afternoon and evening (specifically, from 10:00 AM to 11:00 PM), while moving more actively at late night and early morning. This contrast partially results from the fact that many drivers start their shifts before the morning peak and end at late night. At those periods, the searching behavior of drivers can be drastically influenced by their inclination to either reposition towards ideal service areas or move back home.

4.4.4 Space-time-dependent preference of search movements

Applying the calibrated model above, we then calculate the choice probabilities of search movements at different locations and time periods to see how the latent behavior adapts the aggregate zonal choices to the variable market conditions.

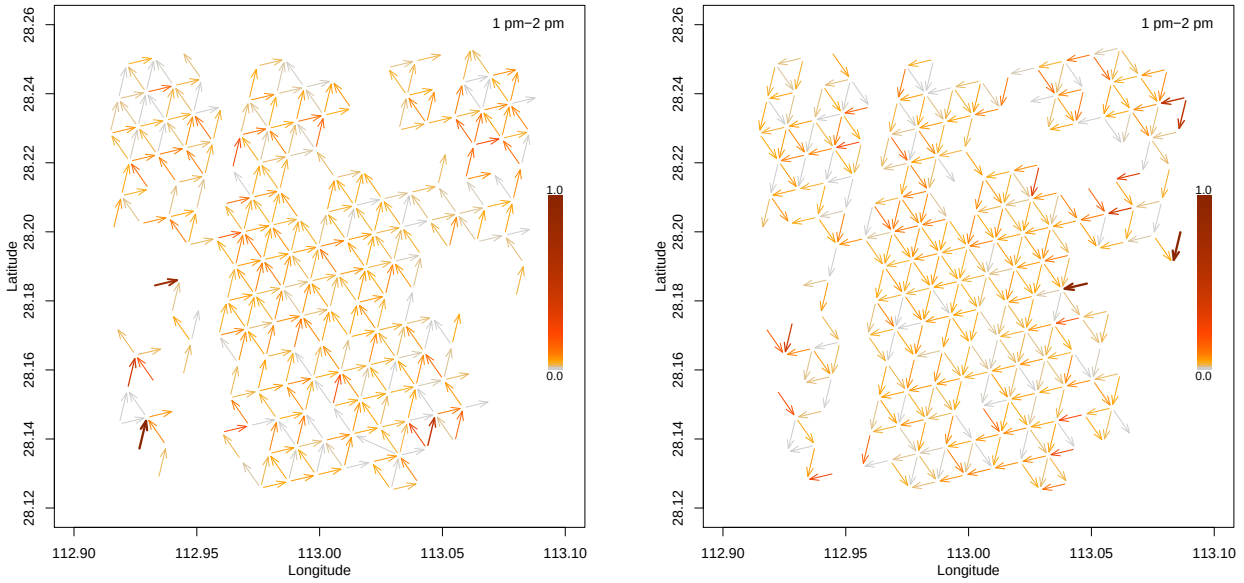
Figure 4.7 and 4.8 visualize the zonal choice probabilities of drivers at different areas in two typical periods to highlight the difference of drivers' latent searching movements between daytime and nighttime. Figure 4.7 first presents the probabilities for drivers in each zone to reposition to the downtown hotspot when being idle, with larger and darker pies marking the higher probabilities



(a) Daytime

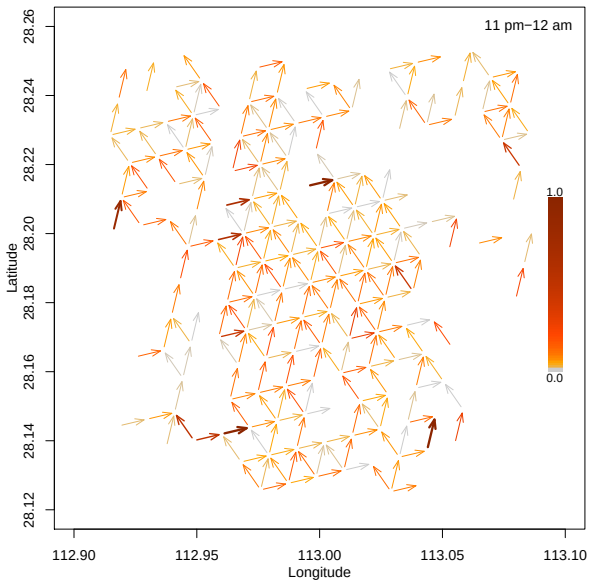
(b) Nighttime

Figure 4.7: Choice probability of idle drivers repositioning towards the downtown in each zone

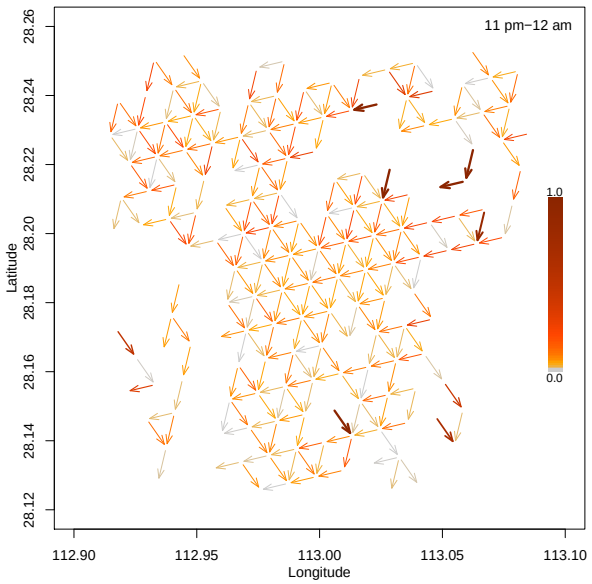


(a) Daytime, Upwards

(b) Daytime, Downwards



(c) Nighttime, Upwards

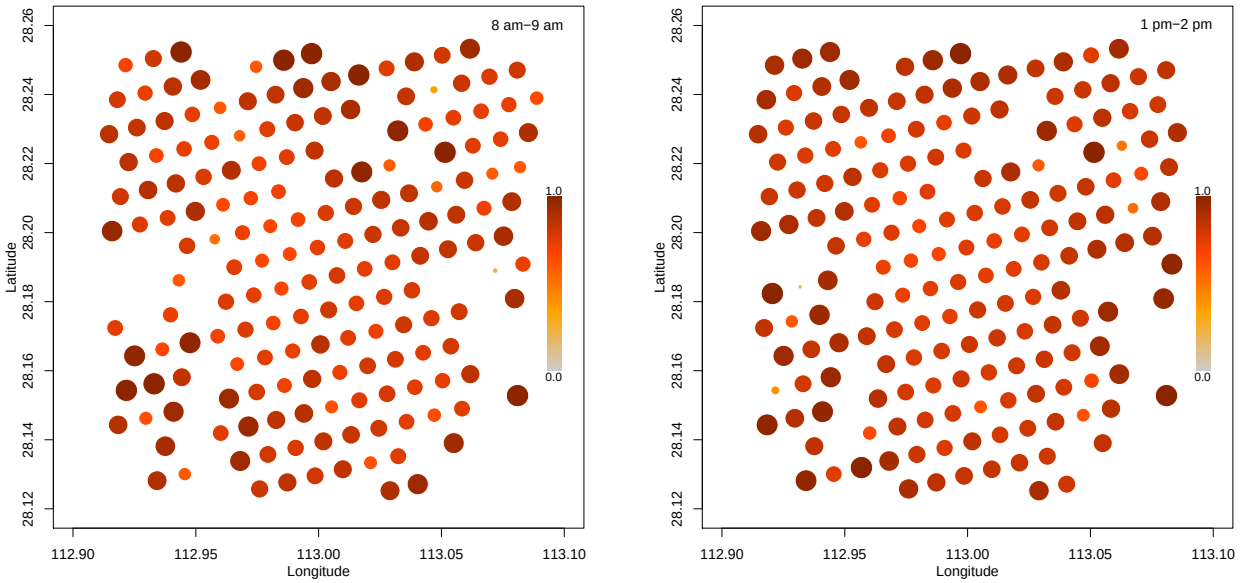


(d) Nighttime, Downwards

Figure 4.8: Choice probability of idle drivers on step-wise movements in zonal search. Each arrow denotes the movement of moving from the origin zone towards a neighboring zone; and darker color indicates that the labeled movement among all the choices available at the origin zone is chosen with higher probability.

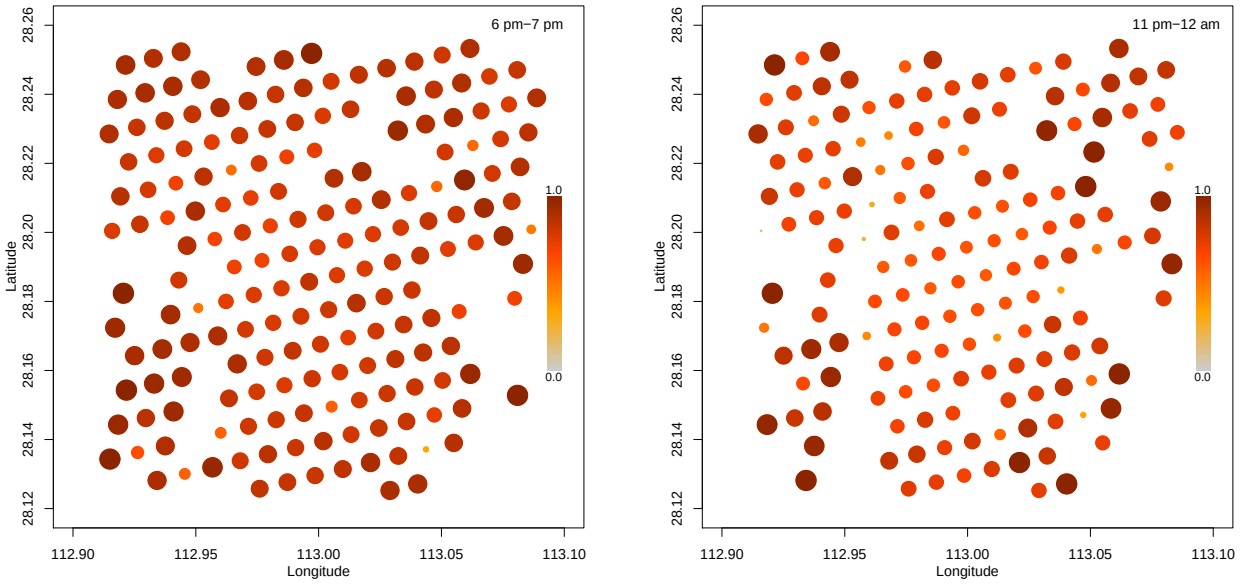
(same for the figures presented later in this section). As can be seen clearly, idle drivers, except for those at the few zones on the outskirt of the city, prefer less the choice of repositioning to the downtown hotspot during the daytime. In contrast, as the suburban market cools down significantly during the evening, idle drivers show much stronger willingness to reposition and escape the potentially long time of wait therein. Such insights are also suggested by Figure 4.8, which details the contrasts by visualizing the step-wise choice probabilities over the space. Each arrow denotes the choice of moving from the origin zone towards a neighboring zone, and again darker color indicates that the corresponding movement is chosen with a higher probability among all the choices available at the origin. Figure 4.8a,c present the probability of drivers searching towards the north (upwards), while Figure 4.8b,d show the counterparts down the south (downwards). Connecting all these preferences graphs the flow tendency of moving towards the downtown area (which is located at the center of the map), where drivers can be matched more easily. From the figures, we can observe that the upward(downward) arrows in the south(north) side generally carry darker colors, which essentially implies the inclination of idle drivers moving to the central area. Meanwhile, compared to the daytime (Figure 4.8a,b), idle drivers during the nighttime show much higher preferences for moving rather than staying motionless (Figure 4.8c,d). The pattern of drivers gathering from the suburban areas to the downtown is greatly strengthened at night.

Figure 4.9 and 4.10 display the choice probabilities of drivers for staying motionless and cruising around the neighborhood, respectively, at four representative periods of a day. Interestingly, we find that drivers are consistently more mobile in the central areas compared to the suburbs. At those regions with rare travel demand, drivers rather than moving prefer to stay motionless. Such an observation is somehow counter-intuitive as taxi drivers are usually considered to be more active in moving and searching at the low-demand areas, to help them find customers more quickly. We suspect that such a attitudinal difference might be due to the introduction of online matching, under which zonal search does not necessarily increase the chances of being matched but pushes up the costs in operations, especially in places with sparse trip demand. Besides, the temporal variations of choice probabilities are also intriguing. For example, the cruising behavior of drivers appears in a wide range of areas during the morning peak (Figure 4.9a), and then almost disappears afterward (Figure 4.9b). Later, starting from the evening peak, while the cruising effect remains weak in the suburbs, it rebounds in the downtown as well as around the railway station (at the southeast corner) and intensifies in the evening up until midnight. Again, such a time-varying contrast could also be a result of the strategic behaviors under online matching. For the idle drivers in the hotspot areas, zigzagging can substantially increase their matching probability, given the relatively denser fleet supply and thus more fierce competition there. But the same strategy may yield marginal improvements in suburbs where both demand and supply are sparsely distributed in space. It would be worthwhile to empirically verify these hypotheses in



(a) Morning peak

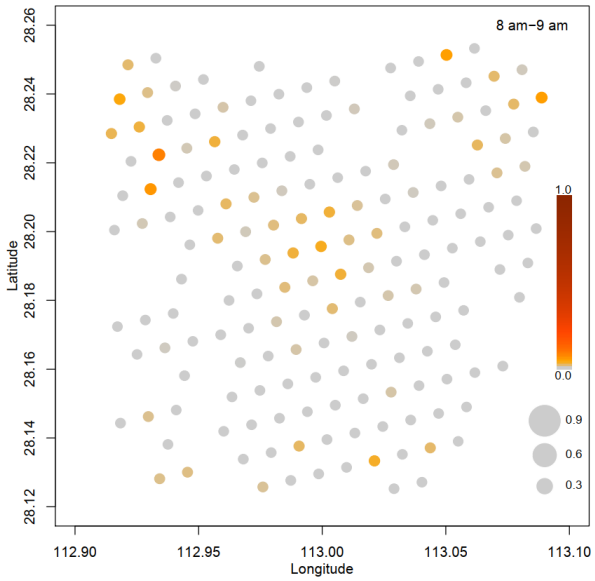
(b) Midday



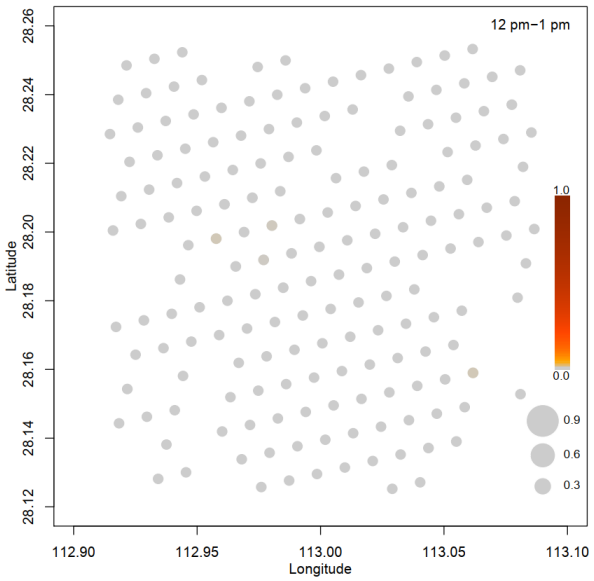
(c) Evening peak

(d) Midnight

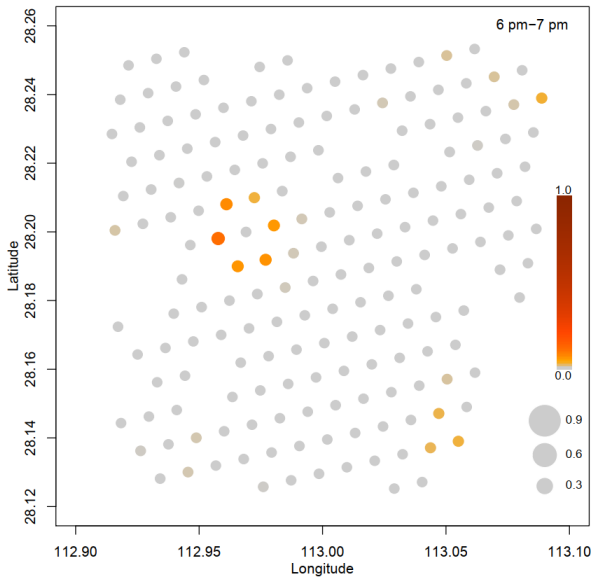
Figure 4.9: Choice probability of idle drivers staying motionless in each zone



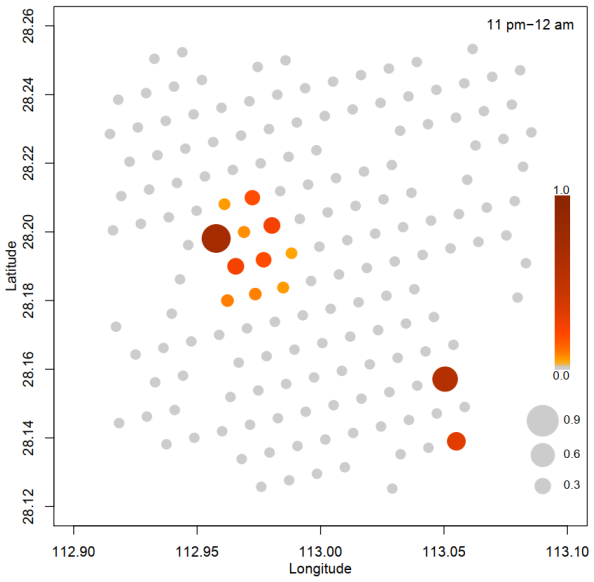
(a) Morning peak



(b) Midday



(c) Evening peak



(d) Midnight

Figure 4.10: Choice probability of idle drivers cruising nearby in each zone

the future, to help reinforce the supply management of app-based ride-hailing services with more sustainable allocations of idle drivers.

4.5 Summary

To the best of our knowledge, the analysis in this chapter is one of the first attempts to investigate ride-sourcing drivers' customer search behavior. A dynamic discrete choice model is proposed to rationalize the time-dependent search movements of idle drivers within the spatial market. The proposed model enables us to evaluate the impacts of spatiotemporal market condition and understand the search behavior of different classes of drivers. In particular, our model considers the unobservable intentions behind drivers' search movements: remaining motionless, cruising around without target areas, or repositioning toward specific destinations. Based on large-scale datasets from real-world operations, we calibrate drivers' time-variant sensitivity to various factors in their decision-making when being idle. Statistical testing results confirm that there exists significant disparity between full-time and part-time drivers, and drivers' preferences in customer search vary across the time of a day. The platform's supply management could be further enhanced by accounting for these preferences and providing more differentiated search guidance for drivers.

Another important message from this study is that drivers have lots of concerns, other than the regular market factors we understand, that may much dictate their search behavior. For example, as revealed by our testing results, drivers generally hold a strong preference for staying motionless rather than moving, partially due to their aversion of the cost in idle movements. For those idle drivers who decided to move, some may prefer to cruise around without a specific target in mind, while others may reposition towards the faraway hotspots with matching "assurance". All these complexities and uncertainties of drivers in idle/search movements pose challenges to the system operations. Therefore, to improve the supply management in the market, some ride-sourcing platforms have started recruiting contracted drivers, who are paid in fixed salary but with the requirement of following the platform's matching and repositioning orders. One of the meaningful tasks is to research how to effectively utilize this group of contractors and turn them into system actuators/controllers. Differentiated matching and repositioning of contractors could be further investigated to address the spatial imbalance of supply and demand, and improve the efficiency of a ride-hailing system.

CHAPTER 5

"Gambling Search"

Pains of Miscommunication

5.1 Introduction

This chapter considers drivers making zonal choices between different service areas, and focuses on their customer search behaviors at the spot markets. Besides the urban areas with spatially scattered customer demand, ride-sourcing services also serve another type of markets with spatially condensed demand, namely spot markets, such as train stations, airports, and amusement parks, etc. (we use the name "spot market" for clarity, with no implication for the spot market used in finance.) Usually, these spot markets are isolated from the business districts, and the ride requests originated there characterize relatively long-distance trips. Those long-distance trips could generate high profitability for drivers, which incentivizes them to cruise far away to or wait a long time at the spot markets, causing a massive accumulation of idle drivers therein. Taking a megacity airport in China as an example, the average queue length of idle drivers stays around 30 during the daytime. The average time for drivers to receive an order that originates from the airport is 36 minutes, while it would only take them 14 minutes to get a ride if they decide to leave the airport. The thick accumulation and the substantial difference in waiting time imply an enormous waste to the effective labor supply. To better target the issue, we should first understand how drivers behave in these spot markets.

Different from most of the previous studies on the spatial choice behavior of idle drivers, this chapter focuses on the temporal decisions in zonal transitions and investigates the process that drivers spend dwelling before they switch out of the spot markets. At those places, drivers' decision sets are more of a structure with binary choices, i.e., leaving or staying, and both choices exhibit high perceived costs for drivers. Instead of investigating which choice a driver takes, we care more about when and how that choice will take place. Thus, a localized zonal choice model with no temporal considerations is no longer applicable. In this chapter, we take drivers' decision of leaving as a hazard, model their stays at spot markets using survival models, and then apply

empirical service data to calibrate and interpret drivers' dwelling behaviors.

The rest of this chapter flows as follows. Relevant literature is first summarized to help position our discussions. Then, we present the survival models prescribed to identify drivers' dwelling behavior at spot markets. The established models are applied to the data from a Chinese megacity airport to study drivers' survival behaviors under random and sequential matching strategies, respectively. Results, insights as well as suggestions for better managing the system at the spot markets are discussed.

5.2 Literature Reviews

Ride-sourcing drivers share a great similarity with the traditional taxi drivers. Ever since Yang and Wong (1998) formulated a network equilibrium model on vacant and occupied taxis' movements, a series of studies have been carried out to better describe the spatial structure of taxi markets (see Salanova et al., 2011 for a recent review). However, due to the lack of real-world data, the results of those early modeling practices were not calibrated and validated. Empirical studies started to appear as the trajectory data of Hong Kong taxis became available. Szeto et al. (2013) and Wong et al. (2014c) first applied the Multinomial Logit (MNL) model to capture strategic zonal choices of vacant drivers' customer search. Since then, a series of research tackled drivers' behavior using various customized Logit models, which allow for drivers' heterogeneous willingness regarding zonal transfers. Zheng et al. (2018) recently modeled vacant taxi drivers' choices between spot and urban areas using a binary Logit model, with additional treatments designed to capture the learning and updating of drivers perceptions from past experience. Nevertheless, the choices of drivers in those contexts are modeled and calibrated as one-shot decisions with no temporal considerations. As a result, no matter how strong a driver is willing to switch, her transitions appear indistinguishable once she decides to switch to another service area. We note that the temporal choices as of how fast drivers make decisions characterize a crucial dimension of their choice behaviors. Missing such components may yield significant loss of references when interpreting drivers' choice behaviors.

Another bulk of research appeared in recent years as the advancements of mobile communication technologies allow navigation for taxis/ride-sourcing vehicles. Routing strategies have been designed and calibrated to reposition vacant vehicles to meet various managerial objectives in operations, such as minimizing the cruising distance (Zhang et al., 2014), or maximizing the potential profitability (see, e.g., Qu et al., 2014; Dong et al., 2014). Some recent studies formulated the vacant vehicle routing problem as a Markov Decision Process (MDP) to account for the long-term effects on the immediate decision-making (see, e.g., Lin et al., 2018; Yu et al., 2019b; Shou et al., 2019). These models treat the routing processes either as optimization

problems where drivers' behaviors are mostly ignored, or as MDP problems where the preceding and subsequent movements of drivers are disjoint. We note that these approaches may be more suitable for modeling a centralized platform's strategies. However, from a practical point of view, these assumptions are mostly invalid in the current decentralized ride-sourcing system, where drivers still dictate the search processes as per their behaviors and individual perceptions. In this case, strategy designs or implementations should take into account how different drivers respond differently to the varying situations, and how they alter their decisions based on their past choices. These considerations, however, are far less investigated among the vacant vehicle routing literature in the past.

This chapter attempts to narrow the gaps mentioned above. Specifically, we apply survival models to delineate the temporal correlations among ride-sourcing drivers' search decisions at spot markets, and identify how drivers of different characteristics differ in the dwelling behavior under various market conditions.

5.3 A Piecewise Survival Model with Proportional Risks

This section introduces the survival model that we develop to learn drivers' dwelling behavior at spot markets. Denote T as drivers' dwelling time at the spot market. We then model T as a survival process, with the "death" referring to the event when a driver decides to leave. Let $\lambda(t)$ be a driver's hazard for leaving at time t. Then, the probability of a driver staying to duration t, i.e., $S(t)$, can thus be written as

$$S(t) = \exp\left\{-\int_0^t \lambda(x)\mathrm{d}x\right\}$$

where the integral term $\int_0^t \lambda(x)\mathrm{d}x$ denotes the cumulative hazard for leaving up to time t.

It should be noted that a portion of the records are not ended with leaving in our case, as drivers may get matched after spending some time dwelling at the spot. For those who get matched directly at the spot market, the records are assumed to be right-censored (or non-informatively censored). All we know from the censored record is that the driver is willing to wait beyond the censoring

Figure 5.1: A typical state trajectory of ride-sourcing drivers at spot market

time. For instance in Fig. 5.1 where the driver gets matched at the time t^m, we know the driver is willing to wait longer than t^m, but are unsure about how long exactly she will wait beyond t^m. It is possible that the driver will leave the spot right after t^m or will stay forever until being matched.

As events occur in different time with different drivers, event-specific covariates $\boldsymbol{x_j}$ needs to be considered for differentiation. We incorporate them into the hazard function by modeling them as proportional risks (Cox, 1972), i.e.,

$$\lambda_j(t|x_j) = \lambda_0(t) \cdot \exp\left\{\boldsymbol{x_j'\beta}\right\}$$

where $\lambda_0(t)$ is the baseline hazard function, also a reference when $\boldsymbol{x_j} = \mathbf{0}$; and $\exp\{\boldsymbol{x_j'\beta}\}$ is the relative risk related to event j's characteristics $\boldsymbol{x_j}$. An increase/decrease in $\boldsymbol{x_j}$ will result in a proportionate increment/reduction in hazards (i.e. risks of leaving) at different time t.

Further, drivers at spot markets usually experience long time of idleness, during which market condition greatly varies. To cope with the time-dependency of covariates $\boldsymbol{x_j}$, we resort to the semi-parametric piecewise hazard functions. As required by the method, the lifetime of a dwelling event $[0,\ \tau]$ is partitioned into reasonably small intervals with cutpoints respectively being $\tau_0, \tau_1, \cdots, \tau_N$ (and $0 = \tau_0 < \tau_1 < \cdots < \tau_N = \tau < \infty$). We assume that $\boldsymbol{x_j}$ change values only at the boundaries between intervals, and then write the hazard for event j at interval i as,

$$\lambda_{ij}(t) = \lambda_i^0 \cdot \exp\left\{\boldsymbol{x_{ij}'\beta_{ij}}\right\},\ \forall t \in [\tau_{i-1}, \tau_i) \tag{5.1}$$

where $\boldsymbol{x_{ij}}$ denotes event j's characteristics/factors in interval i, and $\boldsymbol{\beta_{ij}}$ represents the respective coefficients of the impacts. The piecewise hazard function (5.1) is employed later as a tool to help us uncover drivers' dwelling behavior at spot markets.

5.4 Data Overview

The data we use consist of all the driver trajectories around a megacity airport in China as well as drivers' second-level status (either being idle, picking up customers or delivering them) spanning over three natural weeks from 2018.03.12 to 2018.04.01. The airport in our data is defined as an area including all the entrances and terminals, while the neighborhood surrounding the airport covers an urban area consisting of 97% of the destinations reached by idle drivers after they leave the airport. All the trajectories as well as market conditions were segmented and processed into a granularity of 1 min. As for the market conditions, we calculated the following five covariates for each minute within the investigated periods,

- *ap_queue* - Accumulations of idle drivers at the airport

- *ap_bubble* - Frequencies of App access from riders at the airport
- *twait_stay* - Drivers average idle time of staying at the airport
- *nb_wait* - Drivers average idle time at the neighborhood
- *nb_bubble* - Frequencies of App access from riders at the neighborhood

On each record of dwelling at the airport, besides the idle time and whether or not leaving the airport (or getting matched), the following additional information was retrieved:

- *Time control variables*: the hour of a day, the day of a week;
- *Drivers' attributes*: age, gender, driver type, history with the platform, service score.

5.5 Empirical Calibration and Interpretation

This section presents our calibrations for the piecewise survival model by using the above dataset. Denote R as the combinatorial power term of the proportional risks defined in Eq. (5.1). Then, the formula below presents the linear specifications on R that we eventually adopted:

$$\begin{aligned} R = \; & \beta_{day\_period} \cdot x_{day\_period} \\ & + \boldsymbol{\beta_{driver\_attributes}} \cdot \boldsymbol{x_{driver\_attributes}} \\ & + \beta_{canceled} \cdot x_{canceled} \\ & + \sum_{tlead \in \{0min, 30min, 60min\}} (\beta_{ap\_bubble\_tlead} \cdot x_{ap\_bubble\_tlead} + \beta_{nb\_bubble\_tlead} \cdot x_{nb\_bubble\_tlead}) \\ & + \beta_{twait\_stay} \cdot x_{twait\_stay} + \beta_{nb\_wait} \cdot x_{nb\_wait} \end{aligned}$$

where x_{day_period} denotes a binary indicator for the day type (weekday/weekend) as well as the period in a day; $\boldsymbol{x_{driver\_attributes}}$ is a binary vector denoting drivers' categorized attributes (the specific categories will be presented later); $x_{canceled}$ is a binary variable indicating whether the driver has ever received a match but gets canceled at the airport during the current idle period; $x_{ap\_bubble\_tlead}$ and $x_{nb\_bubble\_tlead}$ denote the market condition realized $tlead$ minutes after the considered moment respectively at the airport and neighborhood; x_{twait_stay} and x_{nb_wait} reflect the expected matching time for staying and leaving the airport. We note that the two sets of variables $(\boldsymbol{x_{ap\_bubble\_tlead}}, \boldsymbol{x_{nb\_bubble\_tlead}})$ and $(x_{twait\_stay}, x_{nb\_wait})$ both indicate market conditions, and are fed individually into separate models.

We have calibrated two sets of models respectively for the groups of drivers who arrive in occupied and idle. However, those drivers arriving in idle are more self-selective, while the

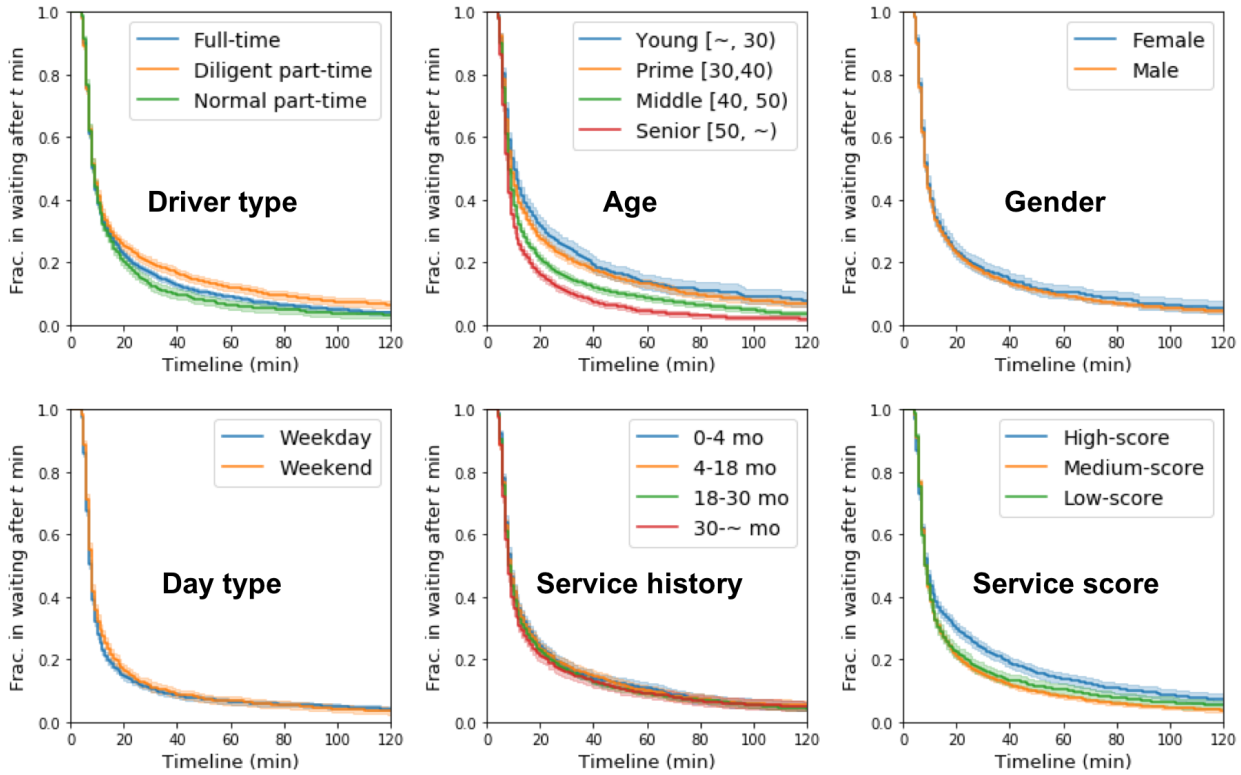


Figure 5.2: Validations for the proportional risk assumption

composition in the occupied group stays closer to the overall driver distributions in the market as the matchings for requests heading to the airport are highly random on drivers. Thus, for the following behavioral interpretations, we focus on the drivers who enter the airport with orders. Before presenting the calibration results, we first validate the assumption of proportional risks. Fig. 5.2 shows the Kaplan-Meier estimates of survivals for different subdivisions of drivers/records. Clearly, among all six sets, none of the two survivals cross significantly with each other. All the survival functions evolve in the same trend with respect to the time. Graphically, this validates that the proportional risk assumption roughly holds on the time controls and driver attributes.

The calibration results are then summarized as follows, with the positive/negative coefficients representing the relative preference/reluctance on leaving:

5.5.1 Drivers in different categories behave differently

Fig. 5.3 compares the coefficient estimates of different categories in four driver attributes. The black dot represents the mean value of each estimate, while the segment around shows the 95% confidence interval. Corresponding to each subfigure, we have the following observations regarding the preferences of drivers in different categories,

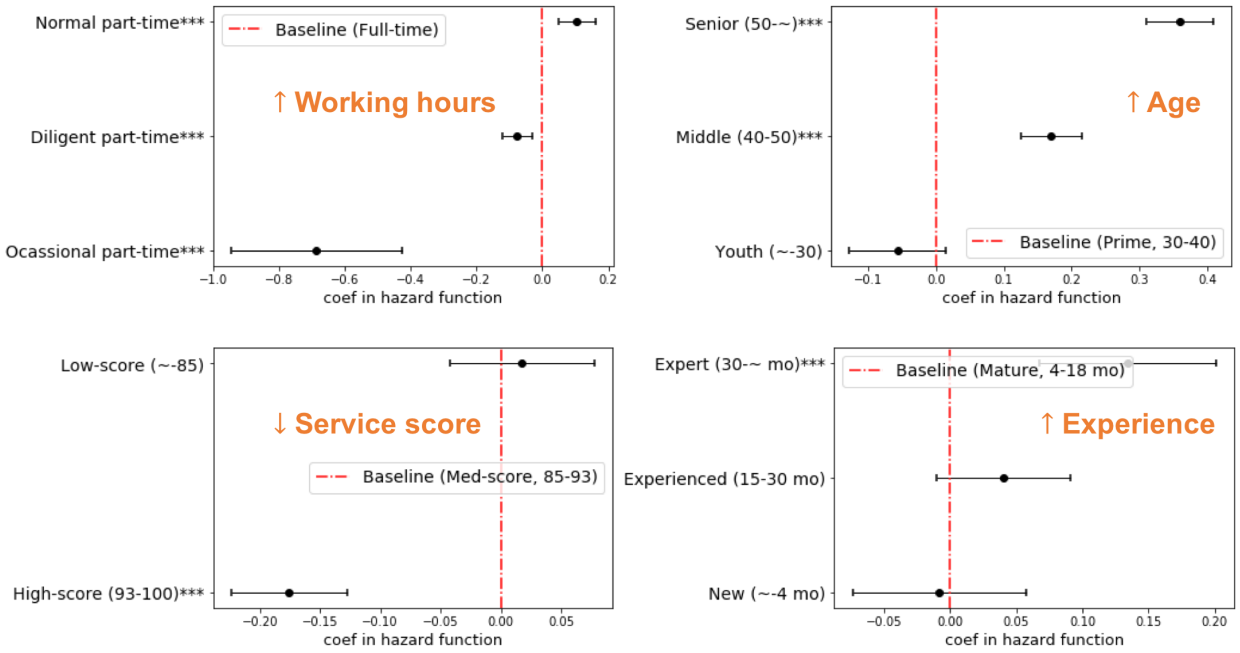


Figure 5.3: Drivers with different attributes differ in hazards

- *Working hours* - Normal part-time drivers are more likely to leave compared to full-time and diligent part-time drivers, while occasional part-time drivers prefer to stay and wait.
- *Age* - More senior drivers compared to the younger counterparts are less patient in waiting.
- *Service score* - Drivers with higher service scores are more reluctant to move, as they are more prioritized in matching.
- *Experience* - Drivers with a longer history of affiliations with Didi Chuxing are generally more impatient in waiting. Perhaps, they know better about the potential waiting time at the airport and are more experienced in finding earning opportunities in urban areas. The perceived costs of leaving may thus be relatively lower for them.

5.5.2 Impacts of the time effects on drivers' survivals

Table 5.1 displays the coefficient estimates on the time fixed effects. The time effects are categorized by the day type (weekday/weekend) as well as the period in a day (morning peak: 04:00-10:00/midday: 10:00-16:00/evening peak: 16:00-22:00/midnight: 22:00-04:00). By comparing the two day-types, we can easily find that drivers have a much higher preference for leaving during the weekdays' peak hours. Also concerning the periods, drivers have relatively higher survival probability for leaving during the morning peak, which then lowers down in the evening. These two observations are fairly consistent with the relative market condition between

Table 5.1: Estimates of the time fixed effects in the survival model

No. events LR test	10,428 1327.884 (df=23)		
Weekday, 04:00-10:00	0.5376*** (0.0339)	Weekend, 04:00-10:00	0.2570*** (0.0379)
Weekday, 10:00-16:00	0 (Base)	Weekend, 10:00-16:00	-0.0721 (0.0486)
Weekday, 16:00-22:00	-0.1425*** (0.0383)	Weekend, 16:00-22:00	-0.1930*** (0.0518)
Weekday, 22:00-04:00	-0.4407*** (0.1323)	Weekend, 22:00-04:00	-0.5928*** (0.1751)

Standard errors in parentheses. ' ' p<1, '.' p<.1, '*' p<.05, '**' p<.01, '***' p<.001

the airport and the neighborhood during different periods (see Fig. 5.4 in the next section), implying that drivers do own some knowledge regarding the general market conditions.

5.5.3 Impacts of order cancellation on drivers' survivals

In the ride-sourcing market, it is common for drivers and customers to cancel orders before they physically meet due to the unsatisfactory over the arrangement. For spot markets where drivers typically take a long time to be matched with a request, cancellation of a matched order may substantially impact their emotions and further influence their behavior. In this study, we apply our survival model to examine whether drivers behave differently after he/she has an order matched but canceled later. Our result shows that there is a significant drop in hazards after cancellations happen, meaning that the cancellation of orders dramatically intensifies drivers' willingness to wait. Such a phenomenon could also be explained by the endowment effect, that receiving an order could strengthen drivers' sense of "owning" a match from the airport, even if the order gets canceled. The choice of leaving the airport will thus carry a higher sense of loss after an order cancellation.

5.5.4 Impacts of the real-time market condition on drivers' survivals

We then evaluate the impacts of the real-time market condition under two different perspectives by using two different sets of covariates.

In the first perspective, we use the App access frequencies at the airport and the neighborhood to characterize market condition, respectively. Besides, two three-variable series are constructed for each time point, to label the condition of the market in 0 min (at the moment), 30 min and 60 min later at the two areas respectively. We use such a structure to examine whether/how drivers

Table 5.2: Coefficient estimates on the frequencies of App access

No. events	10,428	Controls	Yes
LR test	1447.760 (df=26)	Driver F.E.'s	Yes
ap_bubble_0min	-0.0015*	nb_bubble_0min	-0.0001·
	(0.0007)		(0.0001)
ap_bubble_30min	-0.0019**	nb_bubble_30min	0.0002*
	(0.0007)		(0.0001)
ap_bubble_60min	-0.0021***	nb_bubble_60min	0.0002**
	(0.0006)		(0.0001)

Standard errors in parentheses. ' ' p<1, '.' p<.1, '*' p<.05, '**' p<.01, '***' p<.001

respond to the real-time market conditions. Table 5.2 presents the estimates on variables under consideration. Generally, we can see that drivers do notice the market conditions and react in accordance. They become more likely to leave when the neighborhood experiences demand surge, and are more likely to stay when more requests potentially arise at the airport. Specifically, drivers react more consistently to the market variations at the airport, as the corresponding coefficients differ more significantly from zero. We are also surprised to observe that drivers' decisions are more significantly altered by the long-term market condition rather than the short-term. It would thus be interesting to verify if such a long-term "vision" is related to drivers' expectations on the waiting time at the airport.

The second perspective involves two variables that directly bridge drivers' experiences and the real-time market condition. Essentially, we substitute the market condition indicators (*ap_bubble*) and (*nb_bubble*) in the above model respectively with the expected waiting time at the airport (*twait_stay*) and the neighborhood (*nb_wait*) for each moment, and calibrate the coefficients on the new variables. However, a simple inclusion of covariate *twait_stay* will encounter endogeneity issues, as *twait_stay* is dependent on the moving decision of drivers at the airport. We thus instrument *twait_stay* using the frequencies of App access at the airport (*ap_bubble_0min*). Note that *ap_bubble* is a valid instrumental variable (IV) because passengers open the app before they know the supply conditions around. Therefore, there are no feedback effects and *ap_bubble* is uncorrelated with the unknown factors influencing drivers' survivals. The control function approach developed for survival models (Tchetgen et al., 2015) is implemented, and the estimates are displayed in Table 5.3. The result evidences the impacts of the expected waiting time in both areas on drivers' survival probabilities, and again confirms drivers' responsiveness to the real-time market condition.

Table 5.3: Coefficient estimates on the expected waiting time

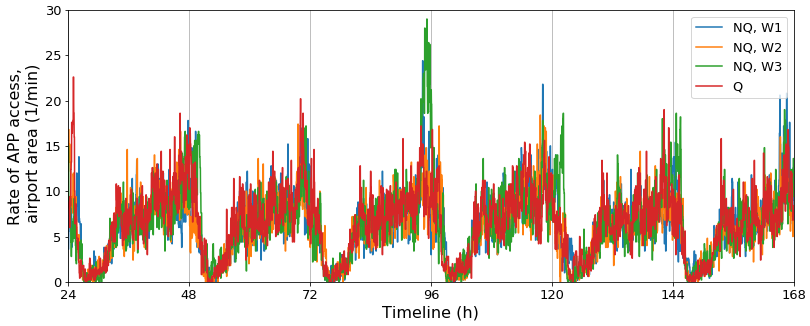
Model	Non-instr.	Instr.
Controls	Yes	Yes
Driver F.E.'s	Yes	Yes
nb_wait	-0.0093***	-0.0220***
	(0.0027)	(0.0035)
twait_stay	0.0029***	0.0217***
	(0.0005)	(0.0033)
No. events	10,428	10,428
LR test	1294.573 (df=22)	1327.884 (df=23)

Standard errors in parentheses. ' ' p<1, '.' p<.1, '*' p<.05, '**' p<.01, '***' p<.001

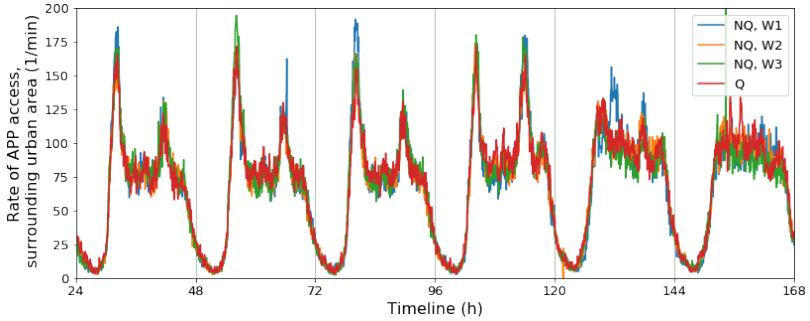
5.6 Drivers' Response to a Switch in Matching Mechanism

Starting from May 21st, 2018, the airport under discussions became the first in China that experimentally implements the "Airport Queuing System" (AQS) for Express service drivers. Before that, idle drivers at the airport are matched "randomly" by the platform with no guarantees on the waiting time. Drivers arrive later could be matched before those coming earlier. The AQS system changed the matching mechanism from "random" to "sequential", under which drivers were updated of their real-time queuing ranks based on their arrival time. Drivers in AQS were guaranteed with a match as long as they wait, and their priorities in matching rise as time progresses. The AQS were only implemented for 6.5 days and got temporarily suspended due to a schedule conflict. So we only got a six-day sample for the system under AQS, but a larger sample from the prior condition (i.e., the three-week dataset used in the previous section). The two samples allow us to investigate drivers' dwelling behavior at the airport both with and without external reference (i.e., the rank in AQS).

We firstly provide an overview of the data under the two periods to acquire some prior knowledge on the system performances under the two matching mechanisms. Fig. 5.4 displays the trends on the App access frequencies for both airport (Fig. 5.4a) and its neighborhood (Fig. 5.4b) areas during the investigated periods. The red line denotes the trend on the six days (Tuesday-Sunday) when the queuing system was implemented, while the other three lines respectively denote the weekly trend under the non-queuing scenario. To be consistent, we trimmed the data on Monday from each of these three weeks under the prior condition. From the figures, we could roughly observe that the customer demand portfolios at both the airport and the area around stay unchanged before and after the implementation of AQS. Furthermore, Fig. 5.5 shows that the rate of drivers arriving in idle at the airport also stay similar throughout the periods, which suggests the limited impact that the AQS imposes on drivers regarding whether cruising to the airport in idle.



(a) Frequencies of App access at the airport



(b) Frequencies of App access at the neighborhood urban area

Figure 5.4: Trends on the App access frequencies at the airport and its neighborhood area

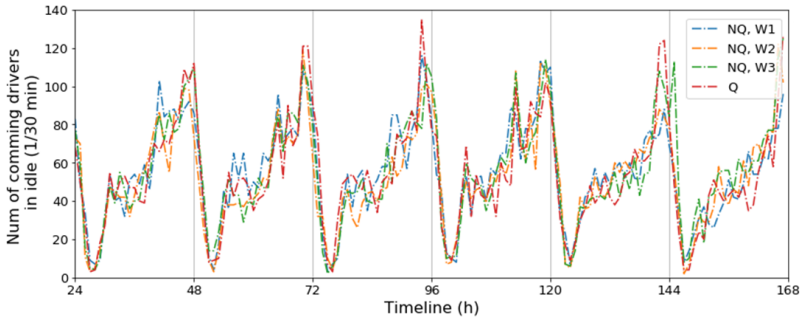


Figure 5.5: Trends on the rate of drivers arriving in idle at the airport

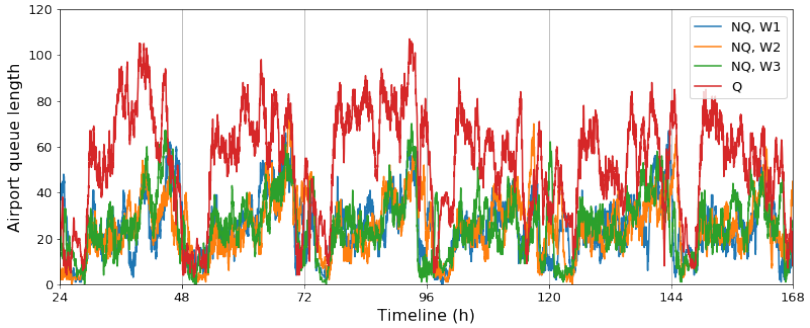
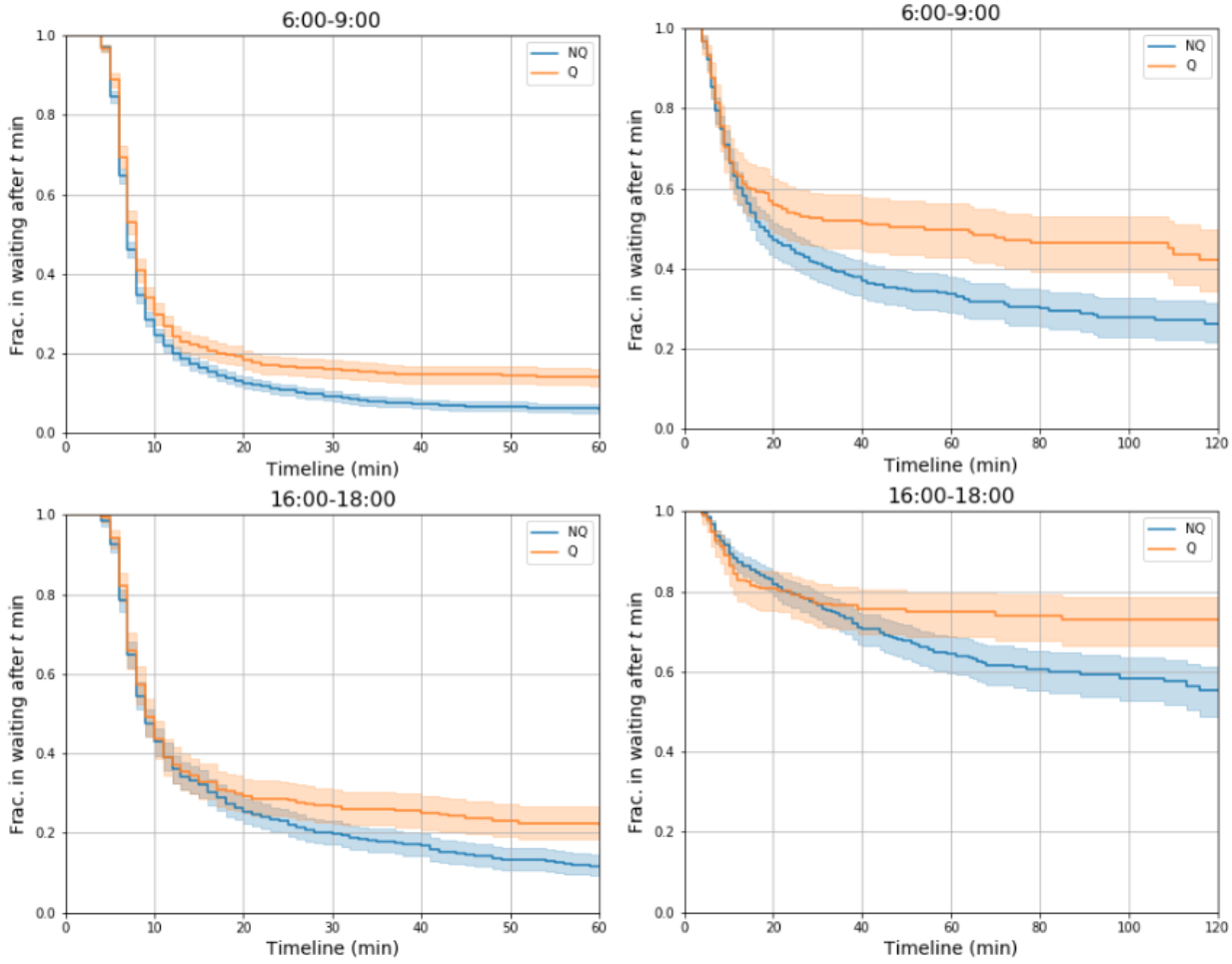


Figure 5.6: Trends on the total number of idle drivers waiting at the airport

Since the arrival rates of occupied drivers are also verified to be stable, we thus conclude that the arrival rates of both demand and supply stay at the roughly same level throughout the two periods.

However, interestingly, the airport queue length plot (see Fig. 5.6) shows that the number of idle drivers dwelling at the airport almost doubled during the daytime after AQS got implemented. Accordingly, the average time for drivers to get matched at the airport also grew significantly under AQS. The reasons should be sought from drivers on how they dwell at the airport. We thus split drivers' waiting periods at the airport with 1 min intervals, and then aggregately plot their survival probabilities by using the Kaplan-Meier estimator. Fig. 5.7 depicts drivers' survivals at the airport during peak hours with/without AQS (i.e., Q versus NQ). The upper two figures correspond to the morning peak, while the two at the bottom show the survivals at the evening peak. Additionally, the survivals are derived separately by different driver groups. As per the initial state on arrivals at the airport (idle/occupied), drivers are categorized into occupied and idle groups, whose survivals



(a) Drivers arriving in occupied

(b) Drivers arriving in idle

Figure 5.7: Comparisons on drivers' survivals at the airport during peak hours with/without AQS

are respectively shown by the left- and right-side figures in Fig. 5.7. Four observations could be easily drawn from the comparisons:

First, it is clear that drivers who arrive in idle have constantly higher survivals over time (more reluctant to leave) compared to their occupied counterparts. One possible reason is that the composition of drivers in the two groups are different, and drivers who cruise to the airport in idle intend to pick up passengers from airport and are thus naturally more patient for waiting. Another possible reason results from the endowment effect that the drivers arriving in idle take the airport as their own choice of receiving the next order while the other occupied party may see the choice as an external assignment (by the platform). Therefore, the idle-type drivers "treasure" the airport waiting more carefully, and are thus more reluctant to change.

Second, in contrast to the morning peak, drivers show higher survivals at the evening peak. Such a difference implies the responsiveness of drivers' decisions to the potential cost associated with staying or leaving. During the morning peak, the airport features a cold spot with fewer orders coming out while the neighborhood area experiences demand surge and is short of supply (see Fig. 5.4). Thus, drivers may perceive lower costs on leaving and higher costs for staying during the morning peak. As for the evening peak, there is comparatively higher customer demand at the airport (see Fig. 5.4a), which lowers down the perceived cost of staying and shapes the choice more appealing to drivers.

Third, the AQS implemented has little effect on discouraging drivers from the long-time queuing at the airport. Except for the right bottom figure in Fig. 5.7 where we do see the fair effect from AQS in prompting drivers with faster decisions, we barely observe positive effects from AQS that improve the system performance metrics such as the average idle time of drivers and the density of idle drivers. Even though the queue length is doubled, drivers arrive as many as before in idle and are more willing to stay and wait. These additional hours of supply (also, wastes in supply) come at the expense of providing a more "reliable" matching.

Last, the AQS implemented significantly strengths drivers' willingness to wait once they decided to stay. The red and blue lines in Fig. 5.7 respectively represent drivers' survivals under the system with and without AQS. As we can observe, the red lines fall above the blue lines under most of the time, and the drivers being idle for more than 20 min under AQS will almost never leave the airport. The reason is that the AQS quantifies and explicitly provides drivers the "gains" they receive through dwelling. The longer time drivers spend in waiting, the higher priorities they will receive in matching. So if they leave the airport before getting matches, all the gains are lost and the efforts already spent become futile. Under such circumstances, drivers' sense of loss aversion is considerably strengthened under AQS, causing them acting more reluctantly in leaving the airport after some time spent in waiting.

The last two observations explain why the driver queue at the airport is lengthened after AQS

is implemented. As the lengthier queue implies greater wastes of supply dwelling in idle, we propose in the next section an enhancement on the AQS that can potentially provide better supply management at spot markets.

5.7 Summary

In this chapter, we study ride-sourcing drivers' dwelling behaviors at spot markets. To accommodate both the individual driver's characteristics and the external market variations, a piecewise survival model with proportional risks is established for the identifications of behaviors and impactful factors. By virtue of massive empirical data from Didi Chuxing, we calibrate the survival model and show that drivers do perceive the market conditions and adjust their actions in response. Also, we validate the temporal correlations among drivers' dwelling decisions. It is found that the longer a driver dwells at the airport, the more reluctant she becomes in abandoning the wait and repositioning to other service areas. Further, by holding drivers into a unified queue, the switch of matching mechanisms from random to sequential substantially strengthens drivers' willingness to dwell and wait for customers. Such a condition, however, intensifies the inefficient use of supply resources.

We believe that the driver queuing system should play a bigger role in supply management rather than being a system just "seemingly more reliable" to drivers. To facilitate the supply management, an effective queuing system needs to address the following two requirements:

1. *Differentiate the "efficient" supply from the "inefficient".*

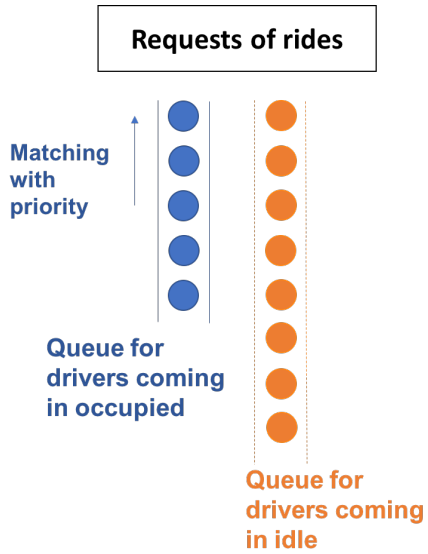


Figure 5.8: A double-queue system

In our context, we evaluate efficiency by the costs of a driver in fulfilling an order from the airport. Obviously, drivers who arrive at the airport occupied are more efficient for picking up customers from the airport compared to those who come in idle, as there are fewer transition costs. Given the customer demand at the airport, it will be smarter management if we could retain a sufficient number of efficient drivers to serve the airport demand while stopping the inefficient drivers from cruising to the airport in idle.

The current AQS did endow the efficient drivers (i.e., those who come to the airport in occupied) with priorities, by letting them start queuing at the beginning of their airport trips (rather than the end after they finish the trips). However, this strategy yields limited effects on encouraging(discouraging) the efficient(inefficient) drivers, as concluded from Fig. 5.7a(Fig. 5.5).

2. *Strategically adapt to the real-time market condition*.
   Higher flexibility is needed in adapting our system operations to the real-time market condition. The AQS implemented did not leave the platform much freedom for management. Ideal management should coordinate between the efficient and inefficient parties, by supplying strong incentives/penalties for encouraging/discouraging the inefficient cruising when necessary.

In light of the above requirements, we propose a double-queue system as a prototype for better design (see Fig. 5.8). There will be two queues set in the system, one express queue (blue) for drivers coming in occupied and one normal queue (orange) for drivers arriving in idle. Suppose the requests of rides generate at a rate of Q, and the drivers from the express and normal lines will be taken at a speed of Q_e and Q_n, respectively. Conceptually, the nature of one-on-one matching gives rise to $Q = Q_e + Q_n$. The assignments on (Q_e, Q_n) generate much greater flexibility for the management. For example, when there is a more-than-sufficient number of drivers in the express line, we could cut off the normal queue and broadcast externally that Q_n will be zero or very small; Conversely, when there is an insufficient number of drivers in the express line, we could tentatively speed up the matching at the normal line and attract more drivers from the outside to catch up the customer demand.

The purpose of the double-queue system is to send stronger signals to efficient drivers to encourage their stay, and then to inefficient drivers to cultivate their adaptations to the market need at spot markets. One meaningful research that could be conducted in the next step is how to allocate Q_e and Q_n appropriately, to balance between the efficiency and reliability under different market conditions.

Government Concerns

The government comes into the play standing by societal or systematic benefits. With the improvements in travel experience in mind though, the government also faces the criticisms for negative system externality imposed by ride-sourcing services. In particular, the traffic manager may need to deal with the congestion due to the extra mileage of RVs. Before taking specific actions, it is critical for the management agencies to first estimate the impacts of empty miles, which is currently a missing ingredient in traffic assignment practices. Chapter 6 thus proposes an enhanced model to encapsulate the empty miles from both cruising and deadheading of RVs. Additionally, a solution algorithm is developed and examined to work efficiently upon realistic-sized networks. In Chapter 7, we then discuss the possible countermeasures that agencies can adopt to mitigate the congestion caused by RVs. Special consideration is paid to investigate the idea of allocating of a certain portion of road space to provide curbside parking spaces to RVs to reduce their cruising. System optimal parking provision schemes are subsequently formulated and analyzed.

CHAPTER 6

Consequence Estimation

General Network Equilibrium Analysis

6.1 Introduction

Conventional trip-based travel demand forecasting models forecast the future-year traffic flow of a traffic network using a four-step process: (i) trip generation, where the total production and attraction of trips from/to each zone are quantified, (ii) trip distribution, where a destination choice model is used to generate the demand for travel between any two given zones in the network, (iii) mode choice, where the number of person-trips using personal vehicles for travel are identified, and converted to vehicular trips based on fixed vehicle-occupancy factors, and (iv) traffic assignment, where user equilibrium assignment models are used to assign vehicular trips to the network, ensuring that no vehicle would be better off by unilaterally changing its route. With ride-sourcing services, vehicular trips are the outcome of the interactions between service operators and passengers, and can no longer be easily determined by applying fixed vehicle-occupancy factors. More specifically, for serving a given passenger demand pattern, the corresponding occupied vehicular trip pattern can be determined similarly as before. Nonetheless, the associated vacant vehicular trips, which result from vacant RVs' behaviors of searching for passengers (hereinafter referred to as cruising trips) and picking up passengers (called as deadheading trips by Ban et al. (2018)), are beyond the reach of traditional analyses. Existing traffic assignment models fail to capture these vacant trips and thus fall short of estimating the impacts of ride-sourcing services on the congestion level of urban traffic networks when the market penetration of ride-sourcing services market is relatively high. We note that a substantial portion of VMTs in the system being vacant or empty is a common feature shared by other shared mobility service systems. It thus becomes increasingly critical for transportation planners to model these vacant trips in their modeling practices. We further note that the state-of-the-art activity-based models also suffers the same limitation as their trip-based predecessors. Therefore, there is a pressing need to enhance existing traffic assignment models to adequately account for vacant trips in shared

mobility systems. Representing one of early efforts in this quest, this chapter develops a static user equilibrium assignment model to capture the network congestion effects of ride-sourcing services.

Ride sourcing shares similarities with traditional taxi services (Zha et al., 2016). A large body of literature has been devoted to network equilibrium analyses of taxi markets. Yang and Wong (1998) made the first attempt to mathematically model the movements of both vacant and occupied taxis in a network context and investigated the equilibrium state of taxi services. It is assumed that once an occupied taxi becomes vacant, it will select a target zone to seek for customers based on a logit model aiming to minimize its expected vacant time. After picking up a customer, the taxi will choose a predetermined shortest route to finish the trip. Their model was then extended by considering the market competition and regulation (Yang et al., 2002), bilateral taxi-customer searching and meeting (Yang et al., 2010), and multiple types of taxis, i.e., street-hailing and e-hailing (He and Shen, 2015). Despite the abundant literature in modeling taxi services, only a few studies have considered congestion due to the routing behaviors of both vacant and occupied taxis. Specifically, Wong et al. (2001) extended the model by Yang and Wong (1998) to further incorporate congestion effects as well as the customer demand elasticity. Specifically, a bi-level framework was proposed to describe the stationary state of taxi movements. At the lower level, given the customer generation and attraction of each zone, a convex mathematical problem delineates the movements of both vacant and occupied taxis in congested road networks. At the upper level, a set of equations describe the relationships among waiting times of taxi and customer, the taxi supply and the customer demand. Recently, Wong et al. (2008) further extended the above model to consider multiple user classes and vehicle modes (e.g., luxury taxis and normal taxis). All these prior studies have laid a solid foundation for our work. As intended, they are applicable to modeling street-hailing taxi systems where matching and meeting occur simultaneously and thus deadheading trips do not exist. In other words, these models consider cruising trips, which only constitute a portion of vacant trips generated by RVs. To facilitate the presentation of the core idea of this chapter, we view street hailing as a form of intra-node matching between customers and vacant vehicles, in which customers can only be matched to vacant vehicles at the same node and thus deadheading is negligible.

More recently, Ban et al. (2018) proposed a traffic assignment model for transportation systems with ride-sourcing services and flow congestion. Particularly, different from the previous models (e.g., Wong et al., 2001, 2008), vacant RVs are assumed to follow the dispatch from a centralized control platform that attempts to maximize the total profit. This assumption can simplify the matching and meeting process between customers and vacant RVs. Consequently, the proposed model only considers deadheading trips while ignoring the cruising portion of vacant VMTs. In practice, after dropping off a customer and before being matched with the next one, idle RVs decide where to go to search for customers. Instead of being dispatched by a control center, idle

RVs decide their cruising-for-customer strategies based on their individual interests.

In existing ride-sourcing systems, cruising and deadheading portions approximately occupy one third and one sixth of an RV's effective working hours respectively in major metropolises. To better account for vacant vehicular miles, we develop a network equilibrium model that considers both the cruising and deadheading trips generated by RVs. To this aim, we firstly extend the modeling framework of Yang and Wong (1998) to describe the network equilibrium state that results from the interactions among occupied and idle RVs as well as the background regular traffic. The network equilibrium model considers intra-node matching between customers and idle RVs, and thus only captures the congestion impacts of cruising trips. We then proceed to consider the inter-node matching between customers and idle RVs, which appears to be a common practice in ride-sourcing markets. Due to the spatial heterogeneity in the ride-sourcing market, e-hailing platforms often match idle RVs from zones with excessive vehicle supply to those experiencing shortage, which yields a substantial number of deadheading VMTs (referred to as "wild goose chase" in Castillo et al. (2018) and further investigated by Zha et al. (2018b) and Xu et al. (2019b)). Our extended model describes the outcome of inter-node matching and captures the congestion effects of both the cruising and deadheading trips generated by RVs.

The remainder of this chapter is organized as follows. Section 6.2 introduces the network equilibrium model with the intra-node matching assumption, which serves as our base model. Section 6.3 proposes an enhanced model to recapitulate the inter-node matching between drivers and riders, and then validates it using the empirical data from Didi Chuxing. Section 6.4 extends the base model by replacing the intra-node matching with the inter-node matching counterparts. A solution algorithm is developed for the extended model, which is implemented to solve numerical examples in Section 6.5. Comparative analyses are subsequently conducted and interpreted. Lastly, Section 6.6 concludes the chapter and points out future research directions.

6.2 Base Model

Consider a road network $G(V, A)$ where V is the set of nodes and A is the set of links in the network. Each node denotes a traffic analysis zone where travel demand is generated from or attracted to, or a change in road geometry/characteristics, while each link represents a road segment connecting two neighboring nodes. For the base model, we assume that the matching radius adopted by the e-hailing platform is relatively small such that customers who request rides from the platform will only be matched to idle RVs nearby (mathematically, at the same node). Consequently, deadheading to pick up customers is negligible. Below we describe the major components of the base model.

6.2.1 Customer demand

Denote the node sets R, S, respectively, as the origins and destinations of customer demands, and let W be the set of origin-destination (OD) pairs. Define β^o and β^i as the out-of-vehicle and in-vehicle value of time (\$/h), respectively. Then, the travel cost C_{rs} between OD pair $(r, s) \in W$ is given by

$$C_{rs} = F_{rs} + \beta^o w_r^c + \beta^i h_{rs}, \quad \forall (r, s) \in W$$

where h_{rs} is the equilibrium or shortest vehicular travel time between node r and s; F_{rs} is the trip fare between the nodes and assumed to follow the structure of $F_{rs} = F_{rs}^0 + \beta^f h_{rs}$, where $\beta^f h_{rs}$ represents the time-based component with $\beta^f (\geqslant 0)$ characterizing the hourly surcharge and F_{rs}^0 denotes a constant for the other time-irrelevant components; and w_r^c is the customer's average waiting time at node r.

Assuming the customer demand Q_{rs} to be a strictly decreasing and convex function of the trip cost C_{rs}, we then have:

$$Q_{rs} = f_{rs}\left(\beta^o w_r^c + (\beta^i + \beta^f) h_{rs}\right), \quad \forall (r, s) \in W \tag{6.1}$$

where $f'_{rs} < 0$ and $f''_{rs} \geqslant 0$. Moreover, $\lim_{w_r^c \to +\infty} w_r^c \cdot f_{rs}(w_r^c | \boldsymbol{h}) \in (0, +\infty)$, implying that there are always finite (small) numbers of customers waiting on each origin node no matter how long they wait.

6.2.2 Idle RV supply

The idle RVs refer to the group of ride-sourcing vehicles that are vacant waiting to be matched. Typically, idle RVs emerge at destination node set S where they drop off customers, and disappear at origin node set R where they get matched to new riders. In this chapter, we assume that, after dropping off a passenger, idle RVs will select and then cruise to a target zone to "search" for customers. Suppose the utility function of an idle RV cruising from node $s \in S$ to node $r \in R$ is prescribed as

$$U_{sr} = \bar{F}_r - \gamma \cdot (\bar{h}_r + h_{sr} + w_r^v), \quad \forall s \in S, r \in R \tag{6.2}$$

where $\bar{F}_r$ and $\bar{h}_r$ are the average fare and service time of the customer trips originating from node r; w_r^v is the idle RV's average waiting time for matching or meeting at node r; and γ denotes

ride-sourcing drivers' value of time (\$/h). Specifically, $\bar{F}_r$ and $\bar{h}_r$ are given as follows,

$$\bar{F}_r = \frac{\sum_{s:(r,s)\in W}(T^o_{rs}+\epsilon^o)F_{rs}}{\sum_{s:(r,s)\in W}(T^o_{rs}+\epsilon^0)}, \quad \bar{h}_r = \frac{\sum_{s:(r,s)\in W}(T^o_{rs}+\epsilon^o)h_{rs}}{\sum_{s:(r,s)\in W}(T^o_{rs}+\epsilon^o)}, \quad \forall r \in R \tag{6.3}$$

where T^o_{rs} is the occupied RV flow that serves customer demand from node r to s; ϵ^0 is a small constant applied to ensure the feasibility of mapping in case $\sum_{s:(r,s)\in W} T^o_{rs}$ in the denominator is zero. Nevertheless, this treatment is primarily for mathematical completeness. Appendix C.3 shows that the zero-flow condition will arise along with an infinite long wait for RVs, i.e. $w^v_r \to +\infty$. In this case, $\bar{F}_r$ and $\bar{h}_r$ weighted respectively from F_{rs} and h_{rs} become negligible among the utility specification (6.2).

Assume that each idle RV cruises towards a node $r \in R$ that maximizes its perceived utility, and the perception error on the utility follows the Gumbel distribution. Then, the portion of idle RVs cruising to r among all those generated at node s, denoted as P_{sr}, is given below:

$$P_{sr} = \frac{\exp(\theta U_{sr})}{\sum_{k\in R}\exp(\theta U_{sk})}, \quad \forall s \in S, r \in R$$

where θ is a constant representing the degree of perceptual dispersion. This leads to the corresponding idle RV flow T^v_{sr}, written as

$$T^v_{sr} = \frac{\exp(\theta U_{sr})}{\sum_{k\in R}\exp(\theta U_{sk})} \cdot \sum_{k:(k,s)\in W} T^o_{ks}, \quad \forall s \in S, r \in R \tag{6.4}$$

We note that the above search model considers only the service opportunities at the target zones, because idle drivers are assumed to reach the zones and then become available for matching there. This, however, could be a strong assumption, because in reality idle drivers may be matched during the transition when they sweep across zones. Relaxing the assumption requires, e.g., the adoption of a sequential search model where idle drivers sequentially update their target zones from the neighboring nodes based on their current locations. A new equilibrium framework that incorporates such a search model is worthy of investigation in the future.

6.2.3 Intra-node matching between hailing customers and idle RVs

As aforementioned, the base model considers only intra-node matching, i.e. customers can only be matched to idle RVs at the same node. Throughout the chapter, we assume one customer is matched with one idle RV, and thus do not consider ride pooling. Since each node features an isotropic zone, an aggregate matching function m_r can be used to characterize the matching frictions between unmatched RVs and customers (e.g., Douglas, 1972). In particular, we adopt the

matching function suggested by Yang and Yang (2011) to capture the competition among drivers and customers over intra-node matching:

$$O_r^m = m_r(N_r^v, N_r^c), \quad \forall r \in R \tag{6.5}$$

where O_r^m represents the realized rate of matching at node r; N_r^v and N_r^c denote the number of idle RVs and hailing customers at node r, respectively.

Under steady states, the variables in the above matching function are also subject to the following relationships:

$$N_r^v = \left(\sum_{s \in S} T_{sr}^v \right) w_r^v \tag{6.6a}$$

$$N_r^c = \left(\sum_{s:(r,s) \in W} Q_{rs} \right) w_r^c \tag{6.6b}$$

$$O_r^m = \sum_{s:(r,s) \in W} T_{rs}^o = \sum_{s:(r,s) \in W} Q_{rs} = \sum_{s \in S} T_{sr}^v \tag{6.6c}$$

6.2.4 Network equilibrium under intra-node matching

Given all the above relations regarding idle and occupied RV movements, we define the network equilibrium state that results from interactions between idle and occupied RVs, and the background regular traffic generated by non-sharing vehicles. Define W^b as the set of OD pairs for the regular traffic, and W^c as the complete set of OD pairs, including those for idle RVs, occupied RVs as well as regular traffic, i.e. $W^c = \{(s,r)|s \in S, r \in R\} \cup W \cup W^b$. Let N denote the total number of RVs serving the network; and let T_{rs}^n denote the OD demand of regular vehicular traffic from node r to s, which in this chapter is assumed to be fixed for all $(r,s) \in W^b$. Then, the equilibrium link flow distribution $\{x_{ij}^{rs}\}$ solves the following system of equalities and inequalities:

Path equilibration between OD pairs

$$\left[t_{ij}(v_{ij}) - \rho_i^{rs} + \rho_j^{rs} \right] x_{ij}^{rs} = 0 \qquad \forall (i,j) \in A, (r,s) \in W^c \tag{6.7a}$$

$$t_{ij}(v_{ij}) - \rho_i^{rs} + \rho_j^{rs} \geqslant 0 \qquad \forall (i,j) \in A, (r,s) \in W^c \tag{6.7b}$$

$$v_{ij} = \sum_{(r,s) \in W^c} x_{ij}^{rs} \qquad \forall (i,j) \in A \tag{6.7c}$$

$$x_{ij}^{rs} \geqslant 0 \qquad \forall (i,j) \in A, (r,s) \in W^c \tag{6.7d}$$

$$T^{rs} = T_{rs}^v + T_{rs}^o + T_{rs}^n \qquad \forall (r,s) \in W^c \tag{6.7e}$$

$$\sum_{i:(i,k)\in A} x_{ik}^{rs} - \sum_{j:(k,j)\in A} x_{kj}^{rs} = \begin{cases} -T^{rs}, & \text{if } k = r \\ T^{rs}, & \text{if } k = s \\ 0, & \text{otherwise} \end{cases} \qquad \forall (r,s) \in W^c \quad (6.7f)$$

$$h_{rs} = \rho_r^{rs} - \rho_s^{rs} \qquad \forall (r,s) \in W^c \quad (6.7g)$$

Customer demands

$$Q_{rs} = f_{rs}\left(\beta^o w_r^c + (\beta^i + \beta^f) h_{rs}\right) \qquad \forall (r,s) \in W \quad (6.7h)$$

$$F_{rs} = F_{rs}^0 + \beta^f h_{rs} \qquad \forall (r,s) \in W \quad (6.7i)$$

Idle RV movements

$$U_{sr} = \bar{F}_r - \gamma \cdot (\bar{h}_r + h_{sr} + w_r^v) \qquad \forall s \in S, r \in R \quad (6.7j)$$

$$\bar{F}_r = \frac{\sum_{s:(r,s)\in W} (T_{rs}^o + \epsilon^o) F_{rs}}{\sum_{s:(r,s)\in W} (T_{rs}^o + \epsilon^o)} \qquad \forall r \in R \quad (6.7k)$$

$$\bar{h}_r = \frac{\sum_{s:(r,s)\in W} (T_{rs}^o + \epsilon^o) h_{rs}}{\sum_{s:(r,s)\in W} (T_{rs}^o + \epsilon^o)} \qquad \forall r \in R \quad (6.7l)$$

$$T_{sr}^v = \begin{cases} \frac{\exp(\theta U_{sr})}{\sum_{k\in R} \exp(\theta U_{sk})} \cdot \sum_{k:(k,s)\in W} T_{ks}^o, & \forall s \in S, r \in R \\ 0, \ \ \forall (s,r) \in W^c \backslash \{(s,r) | s \in S, r \in R\} \end{cases} \quad (6.7m)$$

Intra-node matching

$$\sum_{s:(r,s)\in W} T_{rs}^o = M_r \left(\left(\sum_{s\in S} T_{sr}^v \right) w_r^v, \ \left(\sum_{s:(r,s)\in W} Q_{rs} \right) w_r^c \right) \qquad \forall r \in R \quad (6.7n)$$

$$\sum_{s\in S} T_{sr}^v = \sum_{s:(r,s)\in W} Q_{rs} \qquad \forall r \in R \quad (6.7o)$$

$$T_{rs}^o = \begin{cases} Q_{rs} & \forall (r,s) \in W \\ 0 & \forall (r,s) \in W^c \backslash W \end{cases} \quad (6.7p)$$

RV fleet conservation

$$\sum_{(r,s)\in W} T_{rs}^o h_{rs} + \sum_{(s,r):s\in S, r\in R} T_{sr}^v \cdot (h_{sr} + w_r^v) = N \quad (6.7q)$$

where $\{\rho_k^{rs}\}$ are auxiliary variables, and $\{t_{ij}\}_{(i,j)\in A}$ denote link performance functions that increase monotonically on the corresponding link flows. Note that the number of RVs N is assumed to be fixed, known to the transportation planning agencies. However, if drivers' opportunity wage is

available, it is straightforward to relax this assumption by introducing a function that relates the fleet size to the effective income of drivers. A similar treatment can also be done for the background traffic demand to capture its elasticity on the travel time. The solution algorithm introduced in Section 6.4.4 can be readily adapted to incorporate these extensions. In addition, the above model assumes that drivers always take the routes with minimum travel time. In practice, drivers may be concerned about other costs such as fuel consumption and wearing that are primarily distance-based. In this case, a generalized cost function can be introduced to encapsulate the distance-based cost component for route choice. Consequently, when deciding where to search for customers, RV drivers will consider the generalized travel cost.

The existence of equilibria for the above nonlinear complementarity system (6.7) is proved in Appendix C.3, as the special case of the system under inter-node matching, to be introduced in the following section.

6.3 Modeling Inter-node Matching Between Customers and Idle RVs

The base model captures the matching of customers and drivers within a small neighborhood. However, due to the spatial heterogeneity in the ride-sourcing markets, platforms frequently match idle RVs from one area with excessive supply to customers at another area experiencing supply shortage. As a consequence, RVs are often matched to customers who are several miles away. Modifications should be made to handle this type of long-distance spatial matching, which appears to be a common practice by ride-sourcing platforms.

One straightforward remedy is to enlarge each node to cover a relatively large area with internally balanced demand and supply. However, such an evading strategy may compromise the representativeness and accuracy of the established model, as the aggregate matching function is likely to be biased if intra-node heterogeneity is left out. More importantly, as the node "grows" larger, the intra-node traffic becomes substantial and comparable to the inter-node traffic, defeating the original intent of conducting a network equilibrium analysis.

We thus resort to explicit modeling of the long-distance matching among different nodes, while keeping the node itself to be isotropic and at the neighborhood scale. For passengers at node $r \in R$, we use $M^c(r)$ to represent the set of nodes whose idle RVs can be potentially matched to them. Let $L = \cup_{r \in R} M^c(r)$. Note that these matching sets are not mutually exclusive. Reversely, there is also another matching set $M^v(l) \subseteq R$ for idle RVs at each $l \in L$, representing the set of nodes whose passengers can be matched with idle RVs at l. In this chapter, we assume that all matching sets $\{M^c(r)\}$ and $\{M^v(l)\}$ are exogenously predetermined, and then use T^m_{lr} to denote the rate of

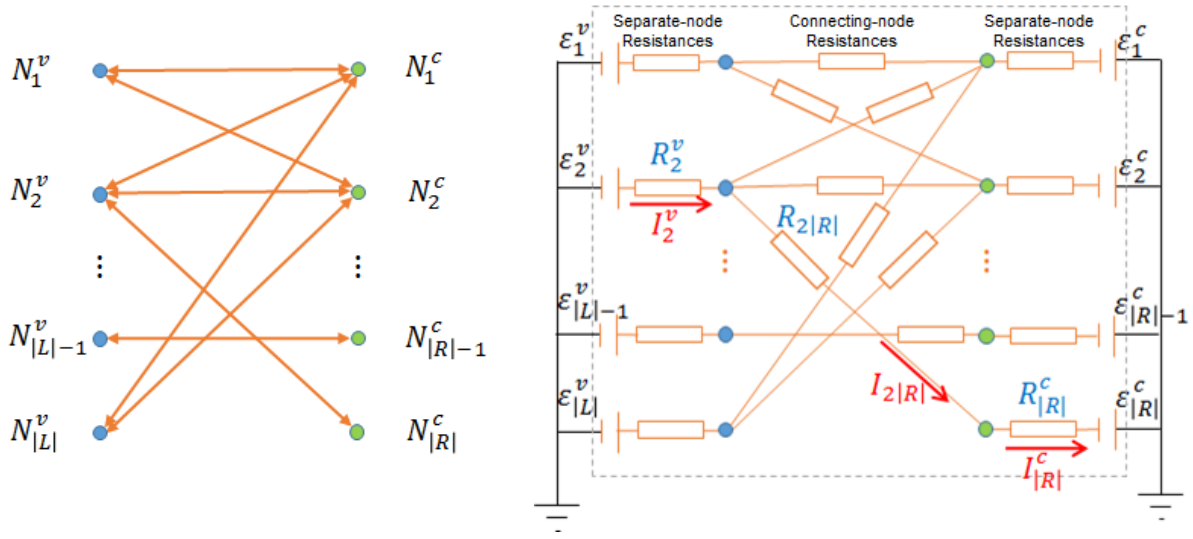
RVs matched from node $l \in L$ to $r \in M^v(l)$[1].

6.3.1 Challenges in modeling inter-node matching

Previously, the intra-node matching flow is estimated from the number of hailing customers and idle vehicles at the same node, using a single-output matching function. With inter-node matching, the idle vehicles/hailing customers at each node can possibly be matched to customers/RVs at a set of nodes. Consequently, a multi-output matching function must be developed to delineate aggregately the matching process. Fig. 6.1a outlines a conceptual instance for inter-node matching, where the blue circles on the left and the green circles on the right represent the pools of idle RVs and hailing customers at different nodes, respectively. Let N_l^v and N_r^c denote the average number of entities respectively at node $l \in L$ and $r \in R$ under the steady state. Then, as shown in Fig. 6.1a, the RV accumulations N_1^v at node 1 are "digested" by flows T_{11}^m and T_{12}^m, while the customer accumulations N_1^c are matched to generate flows of T_{11}^m and T_{12}^m, etc. Therefore, when determining T_{11}^m, the knowledge on N_1^v and N_1^c is not sufficient. We also must know flows such as T_{12}^m and T_{21}^m, which may further depend on $\{T_{2r}^m\}$ and $\{T_{l2}^m\}$, etc. Propagating by nodes and links, the inter-dependencies are likely to integrate over the whole set of matching flows $\{T_{lr}^m\}_{l\in L, r\in R}$ throughout a well-connected network.

Another challenge arises from the matching priority issue. Taking Fig. 6.1a as an example, there are two possible matching outcomes for RVs in N_1^v, either node 1 or 2. Suppose customer

[1]Intra-node matching flows in this framework are essentially indicated by $\{T_{rr}^m\}, \forall r \in R$.



(a) (b)

Figure 6.1: The analogy of (a) inter-node matching flows to (b) currents in an electric circuit

requests from node 1 can be served by RVs from the same node with less pick-up time, compared to the counterparts at node 2. Then, idle RVs at node 1 will be prioritized to match with closer customers, i.e., the customers at node 1 in this case. The modeling of such a priority is nontrivial, because the comparative nature of priorities essentially gives rise to asymmetries for the paired nodes as well as the interdependencies between pairs.

6.3.2 Inter-node matching functions

Our inter-node matching function is built on an analogy to electric circuits. For each inter-node matching graph (e.g., Fig. 6.1a), a corresponding electric circuit can be constructed by duplicating the connections with resistances (a.k.a., connecting-node resistance, see Fig. 6.1b). Additionally, each node outstretches a separate resistance (a.k.a., separate-node resistance) as well as an external power supply, and connects in parallel to constitute a circuit. Denote ε_i^t, R_i^t and I_i^t as the power voltage, the resistance and the current on each node's individual branch,respectively, where $\{(t,i)|t \in \{v,c\}; i \in L, \text{if } t = v; i \in R, \text{if } t = c\}$. Denote R_{lr} and I_{lr} as the resistance and the current on connections between the paired nodes l and r, where $l \in L$ and $r \in R$. Then, the currents on the graph are subject to the following relationships as per Kirchhoff's circuit laws:

$$(\varepsilon_l^v - R_l^v \cdot I_l^v) - (R_r^c \cdot I_r^c - \varepsilon_r^c) = R_{lr} \cdot I_{lr} \qquad \forall (l,r) \in \{(l,r)|r \in M^v(l), l \in L\} \tag{6.8a}$$

$$I_l^v = \sum_{r \in M^v(l)} I_{lr} \qquad \forall l \in L \tag{6.8b}$$

$$I_r^c = \sum_{l \in M^c(r)} I_{lr} \qquad \forall r \in R \tag{6.8c}$$

The two terms on the left-hand side of Eq. (6.8a) represent the potentials at the blue and green nodes, while the right-hand side calculates the potential difference in terms of Ohm's law. The latter two equations Eq. (6.8b) and (6.8c) essentially represent current conservation at nodes. We hypothesize that such an electric-circuit framework mimics or approximates the inter-node matching process and its outcomes. The current on each branch of the circuit is analogous to the matching flow between the corresponding paired nodes, and the voltage of each power supply measures the accumulation of entities at the corresponding node. For each individual branch, higher voltage/accumulation will yield larger current/flow. Further, the resistances in the electric circuit characterize the time that RVs/customers spend during each process. Specifically, the separate-node resistance represents the waiting time of RVs/customers at each node, while the connecting-node resistance quantifies the RVs' deadheading or pick-up time for customers in between. The interdependencies of currents/flows thus transmit rotationally through the two types of resistances/residence time and propagate systematically to the whole graph.

In light by the above analogy, we construct the inter-node matching function following a similar principle. Denote ϕ_l^v and ϕ_r^c as the matching potentials of RVs and customers that wait at different nodes, respectively, and define them as the following functions[2],

$$\begin{aligned} \phi_l^v &= \log\left(\Phi^v\left(T_l^v, N_l^v\right)\right) && \forall l \in L \\ \phi_r^c &= -\log\left(\Phi^c\left(T_r^c, N_r^c\right)\right) && \forall r \in R \end{aligned}$$

where the potential function Φ (including Φ^v and Φ^c) ranges in $(0, +\infty)$, decreasing on the cumulative flow T and increasing on the accumulation N, i.e., $\frac{\partial \Phi}{\partial T} < 0$ and $\frac{\partial \Phi}{\partial N} > 0$. Note that the waiting time w does not appear in Φ's variable argument list, because in this case, w can be directly written as a function of N over T. In addition, we define the potential difference on each pair of connected nodes as the flow and travel time in between, i.e.,

$$\delta\phi_{lr} = \log\left(\Delta(T_{lr}^m, h_{lr})\right) \qquad \forall (l,r) \in \{(l,r) | r \in M^v(l), l \in L\}$$

where the function Δ also ranges from $(0, +\infty)$, and $\frac{\partial \Delta}{\partial T^m} > 0$, $\frac{\partial \Delta}{\partial h} > 0$.

Then, by setting the accumulations $\{N_l^v\}$, $\{N_r^c\}$, and node-transfer time $\{h_{lr}\}$ as given parameters, the resultant matching flow pattern $\{T_{lr}\}$ in line with Eq. (6.8) solves the following equation system[3],

$$\Phi^v\left(T_l^v, N_l^v\right) \cdot \Phi^c\left(T_r^c, N_r^c\right) = \Delta(T_{lr}^m, h_{lr}) \qquad \forall (l,r) \in \{(l,r) | r \in M^v(l), l \in L\} \tag{6.9a}$$

$$T_l^v = \sum_{r \in M^v(l)} T_{lr}^m \qquad \forall l \in L \tag{6.9b}$$

$$T_r^c = \sum_{l \in M^c(r)} T_{lr}^m \qquad \forall r \in R \tag{6.9c}$$

We assume the functions $\Phi(T, N)$, including Φ^v and Φ^c, as well as $\Delta(T, h)$ additionally satisfy the following properties (with implications clarified below), given any N and h with positive and finite values:

- Domain of definition - both $\Phi(T, N)$ and $\Delta(T, h)$ are continuous functions defined on $T \in (0, +\infty)$.

  For each pair of nodes with one located in the matching range of the other, there will be positive flows matched in between.

[2]The logarithm associated with the potentials (also, the potential differences defined later) seems unnecessary, but it facilitates making a connection with a Cobb-Douglas-type matching function.

[3]Eq. (6.9a) results from the internal relations among potentials that $\phi_l^v - \phi_r^c = \delta\phi_{lr}$. The effectiveness of such a physical construct is validated empirically in Section 6.3.3.

- Boundary conditions - $\lim_{T\to 0^+} \Phi(T,N) = +\infty$ and $\lim_{T\to +\infty} \Phi(T,N) = 0$;

  These two conditions can be obtained conceptually, as the accumulation N is stuck in the matching process when $T \to 0^+$ and dissipates instantaneously for $T \to +\infty$.

  $\lim_{T\to 0^+} \Delta(T,h) = 0$ and $\lim_{T\to +\infty} \Delta(T,h) = +\infty$.

  As per the Ohm's law, less potential differences are associated with less matching flows through the time impedance.

- Limiting behavior - there exist $p > 0$ and $q > 0$ such that[4]

$$\Phi(T,N) = \begin{cases} \Theta(T^{-p}), \text{ as } T \to 0^+ \\ \Theta(T^{-q}), \text{ as } T \to +\infty \end{cases}$$

  This implies a diminishing marginal rate of substitution in the matching process. We let Φ possibly converge/grow on T with different speed when approaching $0^+/+\infty$, respectively.

Based on these general properties on Φ and Δ, the existence and uniqueness of solutions $\{T_{lr}^m\}$ to system (6.9) are proven in Appendix C.2. Also, it is worth to point out that when the exogenous matching range of each node confines to itself, the derived inter-node matching functions degenerate to an intra-node counterpart, i.e.,

$$\Phi^v\left(T_{rr}^m, N_r^v\right) \cdot \Phi^c\left(T_{rr}^m, N_r^c\right) = \Delta(T_{rr}^m, h_{rr})$$

which incorporates the specification of Yang and Yang (2011) as a special case.

6.3.3 Empirical validation of the inter-node matching functions

We now apply an empirical dataset from Didi Chuxing to demonstrate the effectiveness of the proposed inter-node matching function. The dataset contains service states of all customers and drivers in the central district of Changsha, China (see Fig. 6.2 for a map of the district) during the whole year of 2019. The region covers an area of about 60 km^2 and consists of 50 standard hexagonal zonal partitions. We aggregate the data by one-hour interval between 7:00 to 22:00 each day to produce samples required by the proposed matching functions. Each sample then records the number of rider-driver matches and the average pickup time between two zones, as well as the total drivers and riders' flows and accumulations at each zone, respectively. Fig. 6.3 visualizes the inter-zonal matching flows in this area. The left and right columns respectively list the 50 hexagonal zones where waiting customers and idle drivers stay, while the connections in between visualize the

[4]Big Θ pertains to one of the Bachmann-Landau notations. By writing $f(n) = \Theta(g(n))$, it means f is asymptotically bounded both above and below by g.

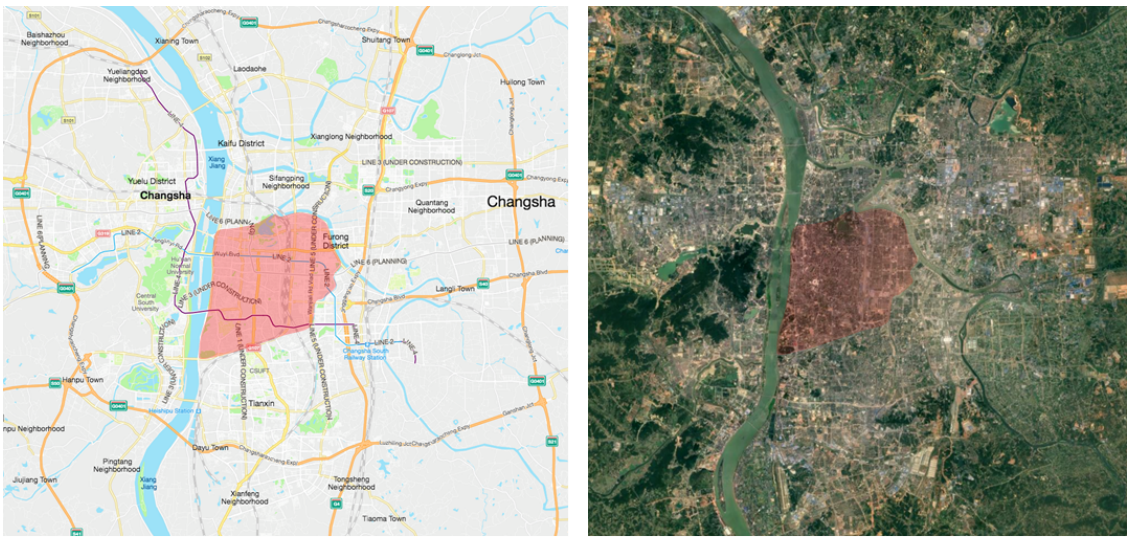


Figure 6.2: Maps for the target district in Changsha, China

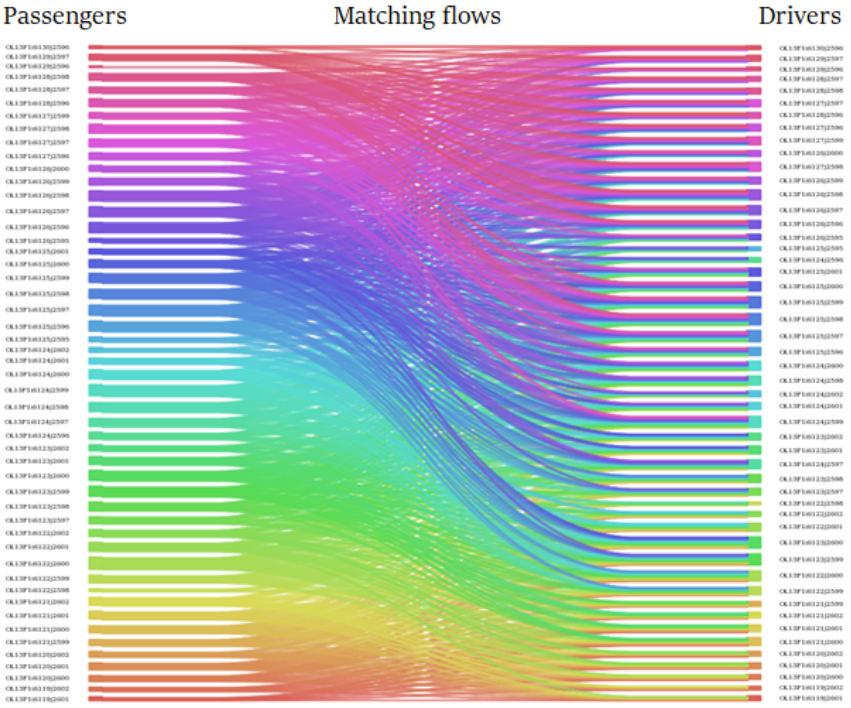


Figure 6.3: Inter-zonal matching between customers and drivers. The left and right columns respectively list the set of zones where riders and drivers stay before matched. The connection of two zones visualizes the relative number of matches realized in between. Thicker linkages denote more pairs of riders and drivers are matched.

relative numbers of matches. As shown in the figure, a significant portion of matches takes place inter-zonally under the 660-meter-hexagon partitions, which an intra-zonal matching function may fall short of modeling. The empirical analysis below further demonstrates the necessity of an inter-zonal matching model and validates the effectiveness of the one we propose.

Let the potential functions Φ^v, Φ^c and the potential difference function Δ in Eq. (6.9a) all take the product forms as follows,

$$\begin{aligned}
\Phi^v(T_l^v, N_l^v) &= T_l^{v q_T^v} \cdot N_l^{v q_N^v} \\
\Phi^c(T_r^c, N_r^c) &= T_r^{c q_T^c} \cdot N_r^{c q_N^c} \\
\Delta(T_{lr}^m, h_{lr}) &= \eta_{lr}^{-1} \cdot T_{lr}^m \cdot F(h_{lr})^{-1}
\end{aligned}$$

where $\boldsymbol{\eta}$ and $\boldsymbol{q}$ are parameters; $F(\cdot)$ specifies a univariate function of the inter-zonal travel time. Substituting Φ^v, Φ^c, and Δ in Eq. (6.9a) with their correspondences and then taking logarithms yield,

$$\log T_{lr}^m = \log \eta_{lr} + \log F(h_{lr}) + q_T^v \cdot \log T_l^v + q_N^v \cdot \log N_l^v + q_T^c \cdot \log T_r^c + q_N^c \cdot \log N_r^c$$

which is invited as the basis for our empirical validation.

The following model is constructed for regression analysis,

$$\begin{aligned}
\log T_{lr,t}^m = {} & \alpha_{lr} + \beta_t + \gamma_{lr,t} + q_N^v \cdot \log N_{l,t}^v + q_N^c \cdot \log N_{r,t}^c \\
& + q_T^v \cdot \log T_{l,t}^v + q_T^c \cdot \log T_{r,t}^c + q_h' \cdot \log h_{lr,t} + q_h'' \cdot h_{lr,t} + q_h''' \cdot h_{lr,t}^2 \\
& + \mathbb{1}_{\text{intra}} \times (d + d_N^v \cdot \log N_{l,t}^v + d_N^c \cdot \log N_{r,t}^c + d_T^v \cdot \log T_{l,t}^v + d_T^c \cdot \log T_{r,t}^c) + \epsilon_{lr,t}
\end{aligned} \tag{6.10}$$

where the subscript t denotes the timestamp of each sample; the inter-zonal (α_{lr}) and time (β_t) fixed effects are incorporated to control for influences that are constant either over time or across zonal pairs, respectively. We spell the time fixed effects as a series of indicators for hour of week, week of year, as well as national holidays; $\gamma_{lr,t}$ is a set of controls that vary over both space and time; meanwhile, the real-time weather information, including the temperature, humidity, precipitation intensity, PM 2.5, wind speed, and sky condition in the region, is also incorporated to control the latent impacts on the system. The effects of spatial distance h_{lr} in matching are captured using three terms, respectively being $\log h_{lr}$, h_{lr} and h_{lr}^2. Special considerations are given to the within-zonal matching to allow certain differentiation from the matching between two different zones.

Based on the construction (6.10), we calibrate five degenerated models (see Table 6.1) for comparison. The model M1 captures only the fixed effects from space, time, and weather, etc., without using state variables from the ride-sourcing system; The model M2 additionally involves

		M1	M2	M3	M4	M5
Fixed effects & controls		✓	✓	✓	✓	
$\{\boldsymbol{N^v}, \boldsymbol{N^c}\}$			✓	✓	✓	✓
$\{\boldsymbol{T^v}, \boldsymbol{T^c}, \boldsymbol{F(h)}\}$				✓	✓	✓
$\{\boldsymbol{T^v}, \boldsymbol{T^c}, \boldsymbol{F(h)}\}$				✓	✓	✓
$\nVdash_{\text{intra}} \times \{\mathbf{1}, \boldsymbol{N^v}, \boldsymbol{N^c}, \boldsymbol{T^v}, \boldsymbol{T^c}\}$					✓	✓
Degrees of freedom		144	156	145	165	12
Sample size		76995	76995	76995	76995	77016
Explained variance	$1 - \text{Var}(\boldsymbol{y} - \hat{\boldsymbol{y}})/\text{Var}(\boldsymbol{y})$	0.47	0.67	0.76	0.94	0.95
R-squared score	$1 - \sum_{i=1}^{n}(y_i - \hat{y}_i)^2 / \sum_{i=1}^{n}(y_i - \bar{y})$	0.34	0.60	0.73	0.93	0.94
Mean absolute error	$\sum_{i=1}^{n} \lvert y_i - \hat{y}_i \rvert / n$	16.36	13.09	10.84	5.59	5.07
Mean squared error	$\sum_{i=1}^{n}(y_i - \hat{y}_i)^2 / n$	612.1	371.2	253.8	61.5	51.0
Correlation	$\text{cov}(\boldsymbol{y}, \hat{\boldsymbol{y}})/(\sigma_y \sigma_{\hat{y}})$	0.69	0.82	0.87	0.97	0.97

Table 6.1: Estimation of five regression models for inter-zonal matching. Truncated regressions are performed on the samples with hourly matching volumes higher than 20 between two zonal pairs. Results are presented using five in-sample performance metrics by comparing the predicted ($\hat{\boldsymbol{y}}/\{\hat{y}_i\}$) and observed ($\boldsymbol{y}/\{y_i\}$) numbers of inter-zonal matches over the sample (of size n). The item $\bar{y}$ denotes the sample mean of $\boldsymbol{y}$.

the number of idle drivers and customers accumulated at each zone, adapting the existing aggregate matching function of street-hailing taxi systems; The model M3 characterizes a basic one that comes into our inter-node matching category, while the model M4 provides extra freedom for within-zonal matching; The last model M5 is a parsimonious inter-node matching model that abandons all the controls and retain only the essential system state variables. Comparisons across the five models help disclose the completeness of different sets of factors in determining the inter-zonal matching.

Additionally, it is worth noting that among the samples of all the effective zonal pairs[5], only 2% contains more than 20 matches within the hourly slot. Since our model takes a logarithmic scaling, the overwhelmingly large-sized samples with extremely low numbers of matches substantially dictates the regression. The dominance of small-number samples then gives rise to large estimation bias at the high-number regime when we map the results back to its original scale. To resolve such an issue, we truncate the samples by trimming those with an hourly number of matches $T^m_{lr,t}$ less than 20, and then apply the rest to calibrate the inter-node matching models using truncated regression.

The bottom five rows in Table 6.1 present a set of performance metrics that evaluate the five models over in-sample (i.e., the truncated sample) predictions for inter-zonal matching volumes. As clearly shown, the models are significantly improved as we enrich the variable

[5]Quantitatively, an effective zonal pair is determined as the one having more than 300 passenger-driver matches throughout the year of 2019.

specifications from model M1 to M4. Specifically, the inter-node matching models with a complete set of variables specified consistently outperform the existing intra-node model developed for street-hailing taxis, which is surely better than the one without considering market information. Meanwhile, the last model M5, which keeps only the most core set of covariates, provides a prediction level equivalent to the most sophisticated model M4. This implies that the structure of our model adheres well to the intrinsic physics of the matching process managed by the platform, regardless of its variations in time, space and weather, etc.

The above results validate the representativeness of our model for inter-zonal matching at the high-volume regime. But as mentioned, about 98% of the samples contain only very few matches over the hourly periods. Fig. 6.4 compares the predicted versus observed numbers of matches over the whole sample from the effective zonal pairs, by using the calibrated models above. Overall, the set of figures suggests highly-consistent results with the in-sample analysis, assuring the generalization of our proposed model. Specifically, the model M2, which is originally applied

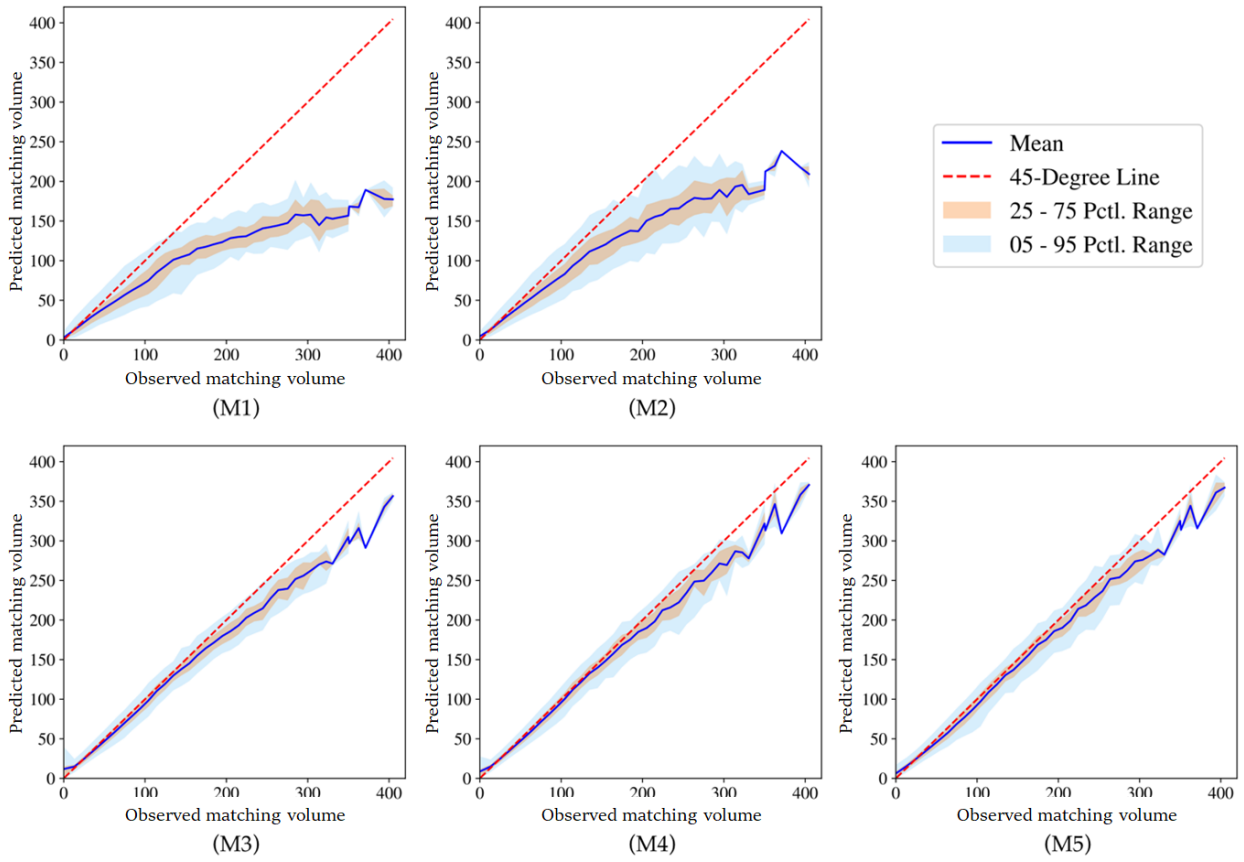


Figure 6.4: Predicted versus observed number of matches over the non-truncated full sample. The blue line shows the mean predictions of inter-zonal matching volumes over each level observed, while the red and blue shaded areas respectively graph the [25, 75] and [5, 95] percentile ranges of predictions. The red dotted line presents a 45-degree line for reference.

for the taxi market, largely underestimates the number of matches in the high-volumes regime. The inter-node matching model M3 fixes such an underestimation significantly, while model M4 with additional differentiation for intra-zonal matching provides the highest consistency between the predicted and observed matching volumes. Again, the parsimonious model M5 is able to achieve a similar level of prediction precision as model M4.

In summary, the empirical validation proves the potential of our modeling framework as a scalable recapitulation for inter-node matching. However, it is worth noting that the calibrated models above are yet not ready for parametric interpretation, as the parameters may not take the physical meaning as they should be in the model construction. The covariates are not fully identified under the current setting, due to the nature of endogeneity amongst the accumulations and flows of drivers and riders. Notwithstanding, the comparisons between the predicted and the observed matching volumes do endorse our model specifications as well as the comprehensiveness of factors that we consider for inter-node matching. For future implementations, we could conduct designated experiments to inference the causal relationship between the inter-zonal matching rates and the system states at different zones, so as to systematically calibrate the matching functions.

6.4 Network Equilibrium under Inter-node Matching Condition

With the enhanced matching functions proposed in Section 6.3, we proceed to relax the intra-node matching assumption in the base model to consider the inter-node matching between customers and idle RVs. The modules of the customer demand and the idle RVs' search movements are first adapted accordingly to the new context.

6.4.1 Customer demand under inter-node pickups

Customers' travel costs now explicitly embody the inter-node pickup time:

$$C_{rs} = g_{rs} + \beta^o w_r^c + \beta^m \check{h}_r + (\beta^i + \beta^f) h_{rs}, \quad \forall (r,s) \in W$$

where β^m denotes customers' value of the pickup time (\$/h) and $\check{h}_r$ is the average inter-node pickup time for customers on node r:

$$\check{h}_r = \frac{\sum_{l \in M^c(r)} (T_{lr}^m + \epsilon^m) h_{lr}}{\sum_{l \in M^c(r)} (T_{lr}^m + \epsilon^m)}, \quad \forall r \in R \tag{6.11}$$

The small constant ϵ^m features another treatment defined to ensure the feasibility of mapping (6.11) when all the flow T^m_{lr}s are zero. We note that the use of ϵ^m here is also innocuous, because the zero-flow condition essentially arises when w^c_r approaches infinity, under which any $\check{h}_r$ valued from the convex hull formed by $\{h_{lr}\}$ becomes nil in comparison.

The customer demand Q_{rs} is then subject to a function reformulated as follows:

$$Q_{rs} = f_{rs}\left(\beta^o w^c_r + \beta^m \check{h}_r + (\beta^i + \beta^f) h_{rs}\right), \quad \forall (r,s) \in W$$

6.4.2 Idle RVs' search target zones

Although the fundamental behavioral consideration remains the same, the modeling of idle RV movements also must be updated, as drivers now choose their search target nodes based on all the potential matching outcomes, including matches to passengers at other nodes in the matching set of a target node. Therefore, the utility function of an idle RV cruising from node $s \in S$ to any node $l \in L$ is re-specified as

$$U_{sl} = \hat{F}_l - \gamma \cdot (\hat{h}_l + h_{sl} + w^v_l), \quad \forall s \in S, l \in L$$

where $\hat{F}_l$ and $\hat{h}_l$ are the average fare and service time of RVs that get matched at node l. Note that the service time now consists of two parts corresponding to deadheading to pick up the matched passenger and then delivering them. We can expand $\bar{F}_l$ and $\bar{h}_l$ respectively as follows,

$$\hat{F}_l = \frac{\sum_{r \in M^v(l)} (T^m_{lr} + \epsilon^m) \bar{F}_r}{\sum_{r \in M^v(l)} (T^m_{lr} + \epsilon^m)}, \quad \hat{h}_l = \frac{\sum_{r \in M^v(l)} (T^m_{lr} + \epsilon^m)(h_{lr} + \bar{h}_r)}{\sum_{r \in M^v(l)} (T^m_{lr} + \epsilon^m)}, \quad \forall l \in L$$

where $\bar{F}_r$ and $\bar{h}_r$ still follow the previous definitions in Eq. (6.3). By assuming that drivers' perception errors on utilities follow the Gumbel distribution, we then write the idle RV flow T^v_{sl} under the inter-node matching scenario as

$$T^v_{sl} = \frac{\exp(\theta U_{sl})}{\sum_{k \in L} \exp(\theta U_{sk})} \cdot \sum_{k:(k,s) \in W} T^o_{ks}, \quad \forall s \in S, l \in L$$

6.4.3 Equilibrium condition

Based on the above modifications derived for inter-node matching, we rewrite the network equilibrium conditions (6.7) to reflect the impacts of inter-node matching. In particular, we replace the intra-node matching function with the inter-node counterpart and account for the inter-node pickups in the cost specifications of customers and drivers' behavioral components. The framework

(6.12) thus characterizes a general equilibrium under inter-node matching that integrates the ride-sourcing market and network equilibration as well as the interplay between them. The complete set of OD pairs W^c now includes those for idle RV trips, deadheading RV trips, occupied RV trips as well as trips made by regular traffic, i.e.,

$$W^c = \{(s,l)|s \in S, l \in L\} \cup \{(l,r)|l \in L, r \in M^v(l)\} \cup W \cup W^b$$

Then, the equilibrium link flow distribution $\{x_{ij}^{rs}\}$ satisfies the following conditions:

Path equilibration between OD pairs

$$\left[t_{ij}(v_{ij}) - \rho_i^{rs} + \rho_j^{rs}\right] x_{ij}^{rs} = 0 \qquad \forall (i,j) \in A, (r,s) \in W^c \qquad \text{(6.12a)}$$

$$t_{ij}(v_{ij}) - \rho_i^{rs} + \rho_j^{rs} \geqslant 0 \qquad \forall (i,j) \in A, (r,s) \in W^c \qquad \text{(6.12b)}$$

$$v_{ij} = \sum_{(r,s) \in W^c} x_{ij}^{rs} \qquad \forall (i,j) \in A \qquad \text{(6.12c)}$$

$$x_{ij}^{rs} \geqslant 0 \qquad \forall (i,j) \in A, (r,s) \in W^c \qquad \text{(6.12d)}$$

$$T^{rs} = T_{rs}^v + T_{rs}^m + T_{rs}^o + T_{rs}^n \qquad \forall (r,s) \in W^c \qquad \text{(6.12e)}$$

$$\sum_{i:(i,k) \in A} x_{ik}^{rs} - \sum_{j:(k,j) \in A} x_{kj}^{rs} = \begin{cases} -T^{rs}, & \text{if } k = r \\ T^{rs}, & \text{if } k = s \\ 0, & \text{otherwise} \end{cases} \qquad \forall (r,s) \in W^c \qquad \text{(6.12f)}$$

$$h_{rs} = \rho_r^{rs} - \rho_s^{rs} \qquad \forall (r,s) \in W^c \qquad \text{(6.12g)}$$

Customer demands

$$Q_{rs} = f_{rs}\left(\beta^o w_r^c + \beta^m \check{h}_r + (\beta^i + \beta^f) h_{rs}\right) \qquad \forall (r,s) \in W \qquad \text{(6.12h)}$$

$$\check{h}_r = \frac{\sum_{l \in M^c(r)} (T_{lr}^m + \epsilon^m) h_{lr}}{\sum_{l \in M^c(r)} (T_{lr}^m + \epsilon^m)} \qquad \forall r \in R \qquad \text{(6.12i)}$$

$$F_{rs} = F_{rs}^0 + \beta^f h_{rs} \qquad \forall (r,s) \in W \qquad \text{(6.12j)}$$

Idle RV movements

$$U_{sl} = \hat{F}_l - \gamma \cdot (\hat{h}_l + h_{sl} + w_l^v) \qquad \forall s \in S, l \in L \qquad \text{(6.12k)}$$

$$\hat{F}_l = \frac{\sum_{r \in M^v(l)} (T_{lr}^m + \epsilon^m) \bar{F}_r}{\sum_{r \in M^v(l)} (T_{lr}^m + \epsilon^m)} \qquad \forall l \in L \qquad \text{(6.12l)}$$

$$\hat{h}_l = \frac{\sum_{r \in M^v(l)} (T_{lr}^m + \epsilon^m)(h_{lr} + \bar{h}_r)}{\sum_{r \in M^v(l)} (T_{lr}^m + \epsilon^m)} \qquad \forall l \in L \qquad \text{(6.12m)}$$

$$\bar{F}_r = \frac{\sum_{s:(r,s)\in W}(T^o_{rs}+\epsilon^o)F_{rs}}{\sum_{s:(r,s)\in W}(T^o_{rs}+\epsilon^o)} \qquad \forall r \in R \quad (6.12n)$$

$$\bar{h}_r = \frac{\sum_{s:(r,s)\in W}(T^o_{rs}+\epsilon^o)h_{rs}}{\sum_{s:(r,s)\in W}(T^o_{rs}+\epsilon^o)} \qquad \forall r \in R \quad (6.12o)$$

$$T^v_{sl} = \begin{cases} \frac{\exp(\theta U_{sl})}{\sum_{k\in L}\exp(\theta U_{sk})} \cdot \sum_{k:(k,s)\in W} T^o_{ks}, & \forall s \in S, l \in L \\ 0, & \forall (s,l) \in W^c \backslash \{(s,l)|s \in S, l \in L\} \end{cases} \quad (6.12p)$$

$$T^o_{rs} = \begin{cases} Q_{rs} & \forall (r,s) \in W \\ 0 & \forall (r,s) \in W^c \backslash W \end{cases} \quad (6.12q)$$

Inter-node matching

$$\Phi^v\left(T^v_l, w^v_l \cdot T^v_l\right) \cdot \Phi^c\left(T^c_r, w^c_r \cdot T^c_r\right) = \Delta\left(T^m_{lr}, h_{lr}\right) \qquad \forall l \in L, r \in M^t(l) \quad (6.12r)$$

$$T^v_l = \sum_{r\in M^v(l)} T^m_{lr} \qquad \forall l \in L \quad (6.12s)$$

$$T^c_r = \sum_{l\in M^c(r)} T^m_{lr} \qquad \forall r \in R \quad (6.12t)$$

$$\sum_{r\in M^v(l)} T^m_{lr} = \sum_{s\in S} T^v_{sl} \qquad \forall l \in L \quad (6.12u)$$

$$\sum_{l\in M^c(r)} T^m_{lr} = \sum_{s:(r,s)\in W} T^o_{rs} \qquad \forall r \in R \quad (6.12v)$$

$$T^m_{lr} = 0, \quad \forall (l,r) \in W^c \backslash \{(l,r)|l \in L, r \in M^v(l)\} \quad (6.12w)$$

RV fleet conservation

$$\sum_{(r,s)\in W^c} (T^o_{rs} + T^v_{rs} + T^m_{rs}) \cdot h_{rs} + \sum_{(s,l):s\in S, l\in L} T^v_{sl} \cdot w^v_l = N \quad (6.12x)$$

Again, $\{\rho^{rs}_k\}$ are auxiliary variables. The existence of an equilibrium solution to the above system (6.12) is proved in Appendix C.3[6]. As system (6.12) degenerates into (6.7), when the matching range of each node shrinks to only cover itself, the proof readily guarantees the existence of an equilibrium solution for the intra-node matching system. The equilibrium solution is likely non-unique for the system considered. Even for an ideal isotropic market without traffic congestion, it has been proved that multiple market equilibria can exist (see, e.g., Zha et al., 2018b).

[6]We suggest readers review Section 6.4.4 before reading the proof in the appendix, as some mapping systems defined in the former are referred in the proof.

6.4.4 Solution procedure

An iterative algorithm in the same vein as the one proposed by Yang and Wong (1998) is developed to solve the network equilibrium system (6.12a-x). Before presenting the solution procedure, we first reformulate some conditions into mathematical problems that can easily be solved through commercial solvers.

Specifically, given $\{\boldsymbol{T^{rs}}\}$, Eqs. (6.12a-f) can be equivalently reformulated as a convex program below:

$$\text{(PE)} \quad \min_{\boldsymbol{x},\boldsymbol{v}} \sum_{(i,j)\in A} \int_0^{v_{ij}} t_{ij}(\varpi) d\varpi$$
$$\text{s.t.} \quad \text{(6.12c)-(6.12f)}$$

Furthermore, treating $\{\boldsymbol{T^m}, \boldsymbol{T^o}, \boldsymbol{h}\}$ as exogenous variables, the idle RV movements (6.12k-q) can be captured by using the following mathematical program:

$$\text{(IRVM)} \quad \min_{\boldsymbol{T^v}} \sum_{s\in S}\sum_{l\in L}\left[\left(-\hat{F}_l + \gamma(\hat{h}_l + h_{sl})\right) T^v_{sl} + \frac{1}{\theta} T^v_{sl}\left(\ln T^v_{sl} - 1\right)\right]$$

$$\text{s.t.} \quad \sum_{s\in S} T^v_{sl} = \sum_{r\in M^v(l)} T^m_{lr} \qquad \forall l \in L \qquad \text{(6.13a)}$$

$$\sum_{l\in L} T^v_{sl} = \sum_{r:(r,s)\in W} T^o_{rs} \qquad \forall s \in S \qquad \text{(6.13b)}$$

To verify, we examine the KKT conditions of IRVM as follows:

(6.13a-b)

$$-\hat{F}_l + \gamma(\hat{h}_l + h_{sl}) + \frac{1}{\theta}\ln T^v_{sl} + \beta_l + \tau_s = 0 \qquad \forall l \in L, s \in S \qquad (6.14)$$

where $\boldsymbol{\beta}$ and $\boldsymbol{\tau}$ are Lagrangian multipliers associated with (6.13a) and (6.13b). From (6.14), we have

$$-\hat{F}_l + \gamma(\hat{h}_l + h_{sl}) + \frac{1}{\theta}\ln T^v_{sl} + \beta_l = -\hat{F}_k + \gamma(\hat{h}_k + h_{sk}) + \frac{1}{\theta}\ln T^v_{sk} + \beta_k$$

which is equivalent to

$$\frac{T^v_{sl}}{T^v_{sk}} = \frac{\exp\left\{\theta\left[\hat{F}_l - \gamma\left(\hat{h}_l + h_{sl} + \frac{\beta_l}{\gamma}\right)\right]\right\}}{\exp\left\{\theta\left[\hat{F}_k - \gamma\left(\hat{h}_k + h_{sk} + \frac{\beta_k}{\gamma}\right)\right]\right\}} \qquad \forall l,k \in L, s \in S$$

and further yields

$$\frac{T_{sl}^v}{\sum_{r:(r,s)\in W} T_{rs}^o} = \frac{\exp\left\{\theta\left[\hat{F}_l - \gamma\left(\hat{h}_l + h_{sl} + \frac{\beta_l}{\gamma}\right)\right]\right\}}{\sum_{k\in L}\exp\left\{\theta\left[\hat{F}_k - \gamma\left(\hat{h}_k + h_{sk} + \frac{\beta_k}{\gamma}\right)\right]\right\}} \qquad \forall l \in L, s \in S$$

From the above formula, we can interpret the term β_l/γ as the RVs' waiting time at node l, i.e., $w_l^v = \beta_l/\gamma$ for all $l \in L$. As $\boldsymbol{\beta}/\gamma$ is not unique, we can always add a constant η to $\boldsymbol{\beta}/\gamma$, such that the above equation still holds.

As per (6.12x), we have

$$\sum_{(r,s)\in W^c} (T_{rs}^v + T_{rs}^m + T_{rs}^o)\cdot h_{rs} + \sum_{(s,l):s\in S, l\in L}\left(\frac{\beta_l}{\gamma} + \eta\right)\cdot T_{sl}^v = N$$

which gives rise to

$$\eta = \frac{N - \sum_{(r,s)\in W^c}(T_{rs}^v + T_{rs}^m + T_{rs}^o)\cdot h_{rs} - \sum_{(s,l):s\in S,l\in L} T_{sl}^v \cdot \beta_l/\gamma}{\sum_{(s,l):s\in S,l\in L} T_{sl}^v}$$

and accordingly,

$$w_l^v = \frac{\beta_l}{\gamma} + \eta$$

We caution that the above steps cannot guarantee the non-negativity of $\boldsymbol{w^v}$. Additional treatment is thus taken in the solution procedure below (Step 4) to ensure the feasibility of equilibrium produced.

Given the above deductions, we develop the following solution procedure to solve the system (6.12) (see Fig. 6.5 for the corresponding flow chart):

1. Initialize $\boldsymbol{N^c}$, $\boldsymbol{N^v}$ and $\boldsymbol{h}$.

2. Obtain $\hat{\boldsymbol{T}}^{\boldsymbol{m}}$ by solving (6.12r)-(6.12t).

3. Deduce $\tilde{\boldsymbol{w}}^{\boldsymbol{c}}$ by solving the following equation on each source node $r \in R$:

$$\sum_{s:(r,s)\in W} f_{rs}\left(\beta^o\tilde{w}_r^c + \beta^m\check{h}_r + (\beta^i+\beta^f)h_{rs}\right) = \sum_{l\in M^c(r)}\hat{T}_{lr}^m \qquad \forall r \in R$$

   Then, retrieve $\{\hat{\boldsymbol{T}}^{\boldsymbol{o}}, \hat{\boldsymbol{w}}^{\boldsymbol{c}}, \hat{\boldsymbol{N}}^{\boldsymbol{c}}\}$ by letting $\hat{T}_{rs}^o = f_{rs}\left(\beta^o\tilde{w}_r^c + \beta^m\check{h}_r + (\beta^i+\beta^f)h_{rs}\right)$ for all $(r,s) \in W$ and $\hat{w}_r^c = \max\{0, \tilde{w}_r^c\}$, $\hat{N}_r^c = \hat{w}_r^c \cdot \sum_{s:(r,s)\in W}\hat{T}_{rs}^o$ for all $r \in R$.

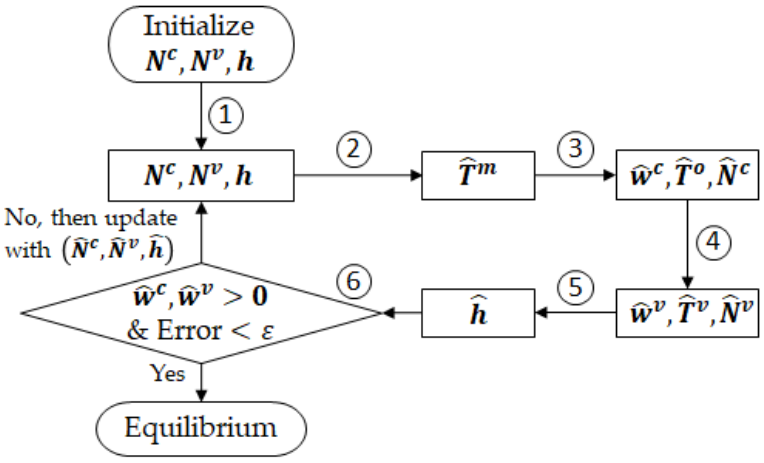


Figure 6.5: Solution procedures

4. Obtain $\{\hat{\boldsymbol{T}}^{\boldsymbol{v}}, \hat{\boldsymbol{w}}^{\boldsymbol{v}}\}$ by solving IRVM, then update $\hat{w}_l^v = \max\{0, \hat{w}_l^v\}$ and $\hat{N}_l^v = \min\{N, \hat{w}_l^v \cdot \sum_{s\in S} \hat{T}_{sl}^v\}$ for all $l \in L$.

5. Obtain $\{\hat{\boldsymbol{x}}, \hat{\boldsymbol{h}}\}$ by solving PE.

6. If $\hat{\boldsymbol{w}}^{\boldsymbol{c}}, \hat{\boldsymbol{w}}^{\boldsymbol{v}} > \mathbf{0}$ and $\kappa_c \|\boldsymbol{N}^{\boldsymbol{c}} - \hat{\boldsymbol{N}}^{\boldsymbol{c}}\| + \kappa_v \|\boldsymbol{N}^{\boldsymbol{v}} - \hat{\boldsymbol{N}}^{\boldsymbol{v}}\| + \kappa_h \|\boldsymbol{h} - \hat{\boldsymbol{h}}\| < \varepsilon$, stop, and the obtained $\{\hat{\boldsymbol{T}}^{\boldsymbol{m}}, \hat{\boldsymbol{T}}^{\boldsymbol{o}}, \hat{\boldsymbol{w}}^{\boldsymbol{c}}, \hat{\boldsymbol{T}}^{\boldsymbol{v}}, \hat{\boldsymbol{w}}^{\boldsymbol{v}}, \hat{\boldsymbol{N}}^{\boldsymbol{v}}, \hat{\boldsymbol{N}}^{\boldsymbol{c}}, \hat{\boldsymbol{x}}, \hat{\boldsymbol{h}}\}$ is the equilibrium solution; otherwise, update $(\boldsymbol{N}^{\boldsymbol{c}}, \boldsymbol{N}^{\boldsymbol{v}}, \boldsymbol{h})$ using its linear combination with $(\hat{\boldsymbol{N}}^{\boldsymbol{c}}, \hat{\boldsymbol{N}}^{\boldsymbol{v}}, \hat{\boldsymbol{h}})$ and repeat the process.

where $\boldsymbol{\kappa}$ in the last step characterizes a set of weights and ε is a given tolerance. Note that the proof of equilibrium existence in Appendix C.3 rationalizes the above procedure design. Basically, the system of equations (6.12) is transformed into a self-mapping on the set of variables $\{\boldsymbol{N}^{\boldsymbol{c}}, \boldsymbol{N}^{\boldsymbol{v}}, \boldsymbol{h}\}$. Deriving the system equilibrium then becomes the problem of finding a fixed-point solution, where the procedure terminates when the conditions before and after the mapping become sufficiently close. The convergence properties of the procedure are not yet theoretically investigated. But for our extensive numerical experiments in Section 6.5.2, we have always been successful in achieving convergence on the large-scale Friedrichshain network.

6.5 Numerical Experiments

This section presents numerical experiments with two networks of different scale. We first use a small-scale Nguyen-Dupius network to illustrate the necessity of considering the inter-node matching. We show that the intra-node model may otherwise yield significant biases in evaluating the vacancy/empty miles generated by RVs of an e-hailing system. We further investigate how RVs' vacancy can be impacted by different market structures on supply and demand. The second set of

experiments is conducted on the Friedrichshain (Berlin) network to showcase how the network and market conditions react to various scales of trip demand and RV fleet supply.

6.5.1 Experiments on the Nguyen-Dupius network

The Nguyen-Dupius network (Nguyen and Dupuis, 1984) consists of 13 nodes, 38 links, and 6 OD pairs (see Fig. 6.6). The link performance function is assumed to be a linear one as follows:

$$t_{ij}(v_{ij}) = a_{ij} + b_{ij} \cdot v_{ij} \qquad \forall (i,j) \in A$$

The parameters $\{a_{ij}\}$ and $\{b_{ij}\}$ as well as the regular background vehicular OD demand $\{T^n_{rs}\}$ for the network are provided respectively in Table C.2a and C.2b, Appendix C.4.

Let there be ride-sourcing demands between the two node sets $\{1, 4, 5\}$ and $\{2, 3\}$ located in the corners of the network (see the filled nodes in Fig. 6.6), and then specify the customer demand function as

$$f_{rs}(C_{rs}) = Q^0_{rs} \cdot \exp\left(-\hat{\theta} C_{rs}\right) + \epsilon^f / C_{rs} \qquad \forall (r,s) \in W$$

where Q^0_{rs} in the first term indicates the potential customer demand from node r to s (see Table

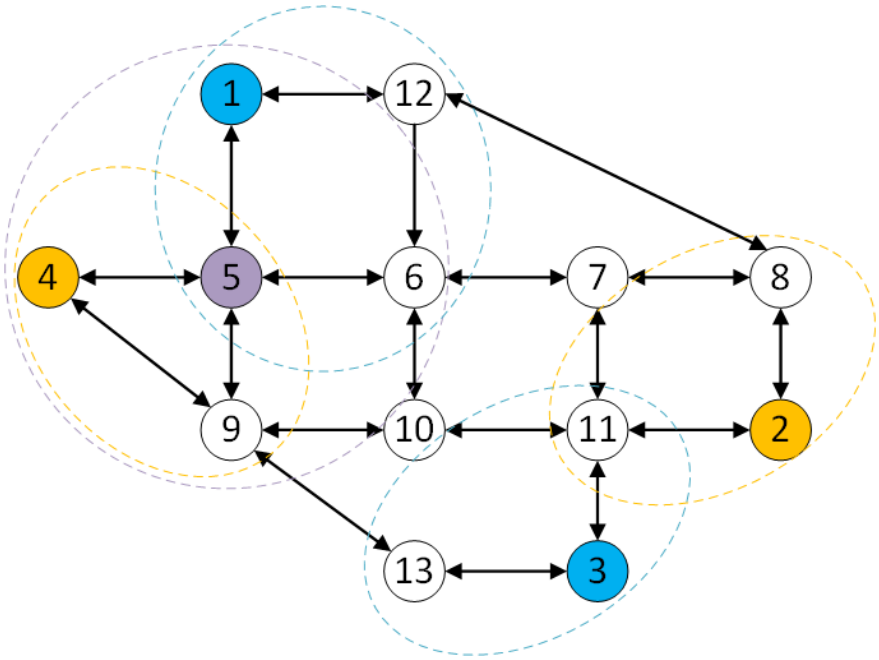
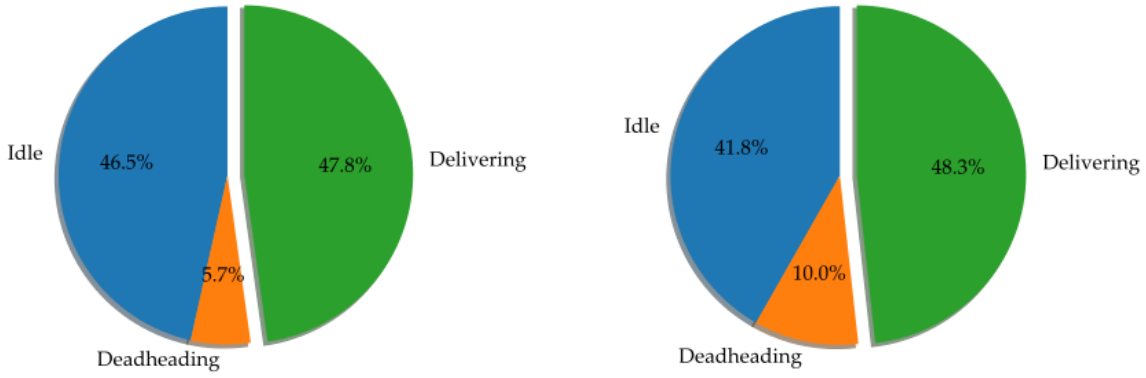


Figure 6.6: The Nguyen-Dupius network. Origin/destination nodes are highlighted by the filled-in color. The matching set of each origin node under intra-node matching includes itself, while the inter-node counterpart covers the nodes in the circle consistently colored as the origin.



(a) Intra-node matching

(b) Inter-node matching

Figure 6.7: Allocation of RVs' service states under intra- and inter-node matching

C.2c); $\hat{\theta}$ features travelers' sensitivity over the cost; ϵ^f in the second term is a small constant. The reciprocal term, which imposes marginal effects under normal conditions, is constructed purposely to ensure the extreme condition $\lim_{w_r^c \to +\infty} w_r^c \cdot f_{rs}(w_r^c|\boldsymbol{h}) \in (0, +\infty)$.

For inter-node matching, we further specify the potential functions Φ^v, Φ^c and the potential difference function Δ in Eq. (6.12p) as follows:

$$\Phi^v\left(T_l^v, N_l^v\right) = \left(T_l^v\right)^{-q_T^v} \cdot \left(N_l^v\right)^{q_N^v}$$
$$\Phi^c\left(T_r^c, N_r^c\right) = \left(T_r^c\right)^{-q_T^c} \cdot \left(N_r^c\right)^{q_N^c}$$
$$\Delta(T_{lr}^m, h_{lr}) = \eta \cdot T_{lr}^m \cdot (h_{lr})^{q_h}$$

where η, q_h and $\{q_j^i\}_{i\in\{v,c\}, j\in\{T,N\}}$ are parameters. Note that all parameter values used in the experiments are summarized in Table C.2d, Appendix C.4 for reference.

Necessity of the inter-node matching model

The matching sets under intra-node and inter-node matching are summarized below:

Matching range	$M^c(1)$	$M^c(4)$	$M^c(5)$	$M^c(2)$	$M^c(3)$
Intra-node	$\{1\}$	$\{4\}$	$\{5\}$	$\{2\}$	$\{3\}$
Inter-node	$\{1, 5, 6, 12\}$	$\{4, 5, 9\}$	$\{1, 4, 5, 6, 9\}$	$\{2, 8, 11\}$	$\{3, 11, 13\}$

Assuming 2,250 RVs in service, equilibrium solutions are then obtained by applying the procedure developed in Section 6.4.4.

We first compare in Fig. 6.7 the allocations of RVs' service states under the intra-node and inter-node matching. The pie charts show the percent of RVs that are idle, deadheading, and delivering. Note that in intra-node matching, a small, constant intra-node pickup time is considered as deadheading whenever a match is made. The RVs' occupancy appears similar in the two cases,

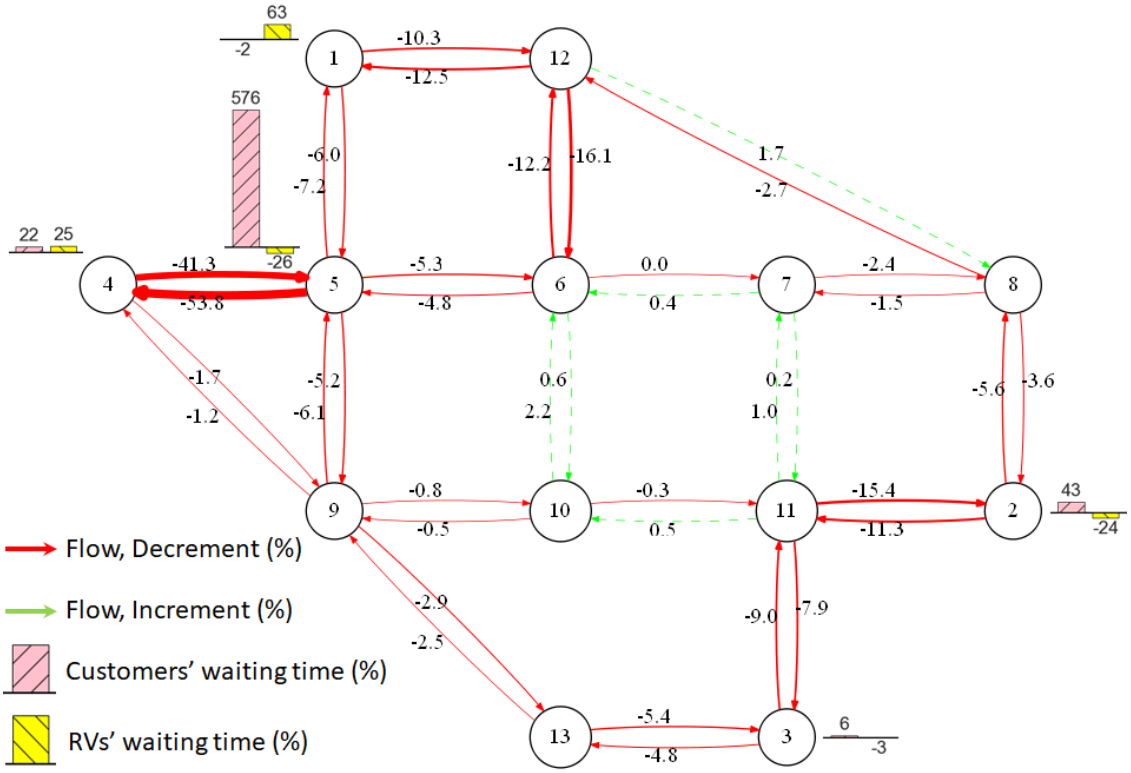


Figure 6.8: Comparisons of flow distributions as well as customers and RVs' waiting time under intra- and inter-node matching

while the state allocation for vacancy differs drastically. In contrast to the intra-node matching counterpart, RVs under inter-node matching on average experience much less idleness but spend significantly more time on pickup. Such a contrast implies a critical trade-off between waiting (idling) and deadheading of RVs by varying the matching radius, which concerns not only drivers but also ride-sourcing platforms (see Xu et al., 2019b).

Taking the inter-node equilibrium solution as the baseline, we then compare in detail the differences between these two solutions. Fig. 6.8 displays the relative changes on the link flows as well as RVs and customers' waiting time. The red and green arrows indicate percentage decreases and increases of flows on different links, respectively. The exact percent difference is provided next to each arrow and shown graphically by its thickness. There is a significantly higher amount of flow on links around the two sets of source nodes under inter-node matching (e.g., link $5 \to 4$ and $11 \to 2$). Further, the red and yellow columns in Fig. 6.8 present the percentage changes on RVs' and customers' waiting time, respectively. Since inter-node matching enlarges the matching range, thereby reducing the meeting frictions between drivers and customers, both parties' waiting time in matching may decrease compared to the intra-node matching counterparts (see, e.g., node 4). Counterexamples, however, arise when the matching sets of two source nodes grow large enough

to overlap with each other; for example, the pair of nodes 1 and 5, where node 1 exhibits more significant excess of supply compared to node 5. As a consequence, as we enlarge the matching range to have node 1 and 5 overlap substantially in their matching sets, the supply excess condition on the two nodes neutralizes, which yields significantly shorter waiting time for customers on node 5 and RVs on node 1.

The above differences suggest that for e-hailing systems with large matching radius, an intra-node matching model may drastically bias the estimates of market and network conditions due to the lack of capability of modeling vehicles' deadheading.

Impacts of market attributes on RVs' vacancy

Next, we investigate the impacts of different market attributes on RVs' vacancy. Specifically, this subsection focuses on two system attributes: the level of RVs' fleet supply and the spatial symmetry of customers' demand pattern. The contrasts in supply level are constructed using two fleet sizes, 2,250 and 1,750, which respectively correspond to the condition of sufficiency and shortage. In addition, we adjust the potential demand portions between the two clusters of origins (1, 4, 5) and (2, 3) (see Fig. 6.6 for specific locations) to differentiate the symmetry of demand patterns. The nominal example introduced above characterizes a symmetric case with relatively balanced ride-hailing demand between these two clusters, while an asymmetric instance is newly generated by intensifying the demand imbalance (see Table C.2c for the specific setting).

Fig. 6.9 summarizes the allocations of RVs' service states under the four combinations of supply-demand attributes. We obtain the following observation via comparison. In all cases, RVs dedicate around half of their service time to delivery and spend the other half on cruising or deadheading, producing the so-call empty miles. More specifically, compared to the sufficient supply scenario, RVs experience less idleness and higher occupancy under supply shortage. The contrasts appear more noticeable under symmetric demand patterns (see Figs. (6.9a) v.s. (6.9c) and (6.9b) v.s. (6.9d)). Symmetric demand patterns allow idle RVs to get a match quicker after dropping a passenger, yielding a higher occupancy rate. Indeed, demand symmetry increases vehicular occupancy in both insufficient and sufficient supply cases.

6.5.2 Experiments on the Friedrichshain network

The above small-scale Nguyen-Dupius network example has demonstrated the need for constructing the inter-node matching model and the capability of the model on capturing ride-sourcing market conditions. In this section, we present a second set of experiments with a network based on Friedrichshain, Berlin (see Fig. 6.10 for an overview) to demonstrate how the model can be applied to a realistic network to estimate the network and market conditions. The Friedrichshain network contains 224 nodes, 523 links, and 506 OD pairs. Due to the

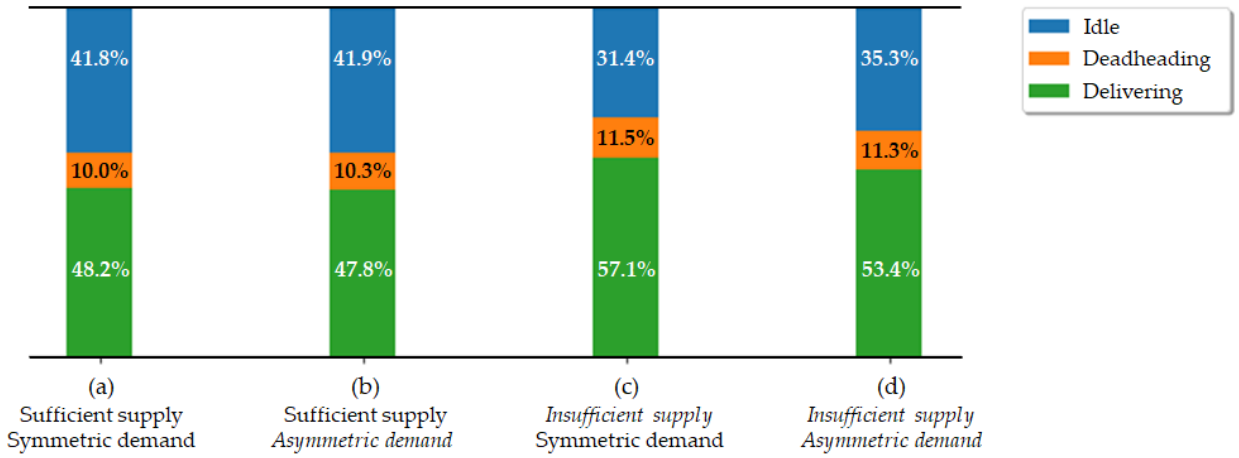


Figure 6.9: Allocations of RVs' service states under different supply-demand contexts

space limitation, other network data are omitted in this dissertation, but are made available at https://limos.engin.umich.edu/research/ride sourcing/.

Our experiments mainly involve two variables: the demand for ride-sourcing services and the total RV supply. We combine these two variables to generate 100 instances to see how the model predicts the network and market conditions. For all these numerical examples, we have successfully achieved convergence when applying the proposed procedure. The convergence is

Figure 6.10: The Friedrichshain Network, containing 224 nodes and 523 links.

typically reached within several hundreds of iterations. Each iteration takes about 10 seconds to execute in our GAMS code on a desktop with Intel i5-6600 CPU@3.30Ghz and 8GB memory. Two measures are of particular interest: average network running speed to represent the congestion level and the percent time being vacant to reflect the ride-sourcing market. Denote $\boldsymbol{Q^0}$ as the base level of customer demand for ride-sourcing services. We then multiply $\boldsymbol{Q^0}$ by a demand index α to yield different levels of trip requests. Ten scenarios are considered by varying α from 1 to 10. Additionally, for each level of α, we consider ten different RV fleet sizes N to replicate different levels of labor supply. Fig. 6.11 summarizes the average network running speed and the average empty time ratio of RVs of these 100 instances.

The average network running speed in Fig. 6.11a is calculated as the miles traveled by all the vehicles over the total time they spend on road. As seen from the graph, more demand or supply generally increases congestion. However, when the level of demand (supply) is high, the marginal contribution of additional supply (demand) to congestion is more significant. Fig. 6.11b presents the empty time ratio of RVs, which denotes the percentage of time on average RVs spend in idle or deadheading. Intuitively, we observe that the ratio decreases with the demand level while increasing with the supply level. In contrast to the speed measure, the empty time ratio of RVs shows the greatest sensitivity under the low level of demand and high level of supply.

Note that even with severe supply shortage (the lower left corner of Fig. 6.11b), the empty time ratio stays above 40%. By allowing a certain level of empty ratio, the ride-sourcing system can endure fewer frictions and operate more smoothly. Thus, it would be meaningful to investigate the optimal empty ratio that maximizes the social welfare of the whole system.

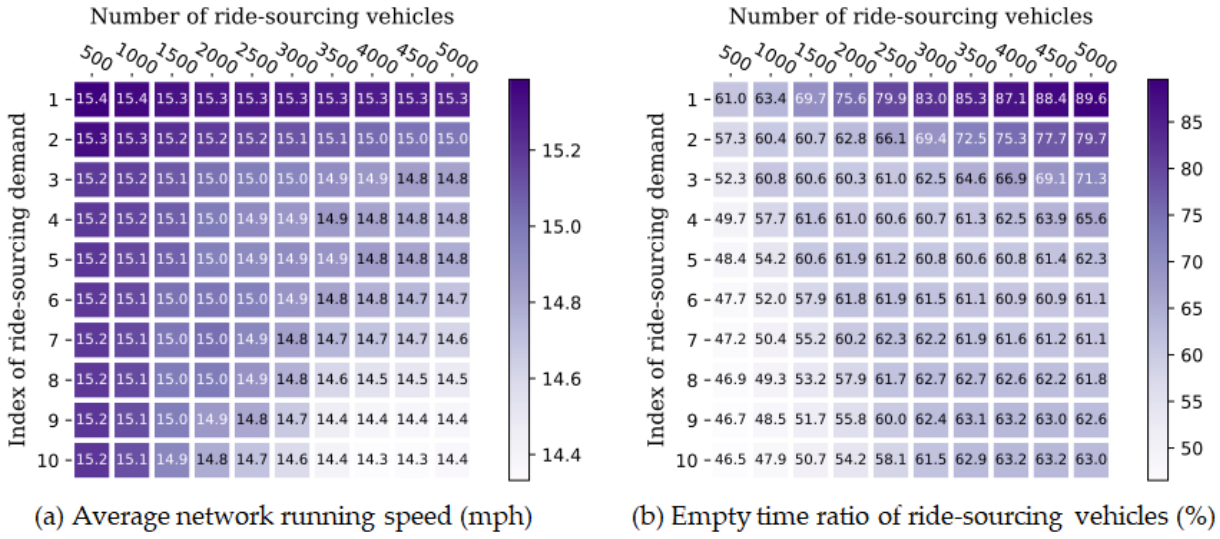


Figure 6.11: Summary of the ride-sourcing system's state statistics

6.6 Summary

This chapter proposes an equilibrium traffic assignment model for urban transportation networks where a large portion of travel demands are served by ride-sourcing services. In such networks, besides the occupied vehicular trips transporting travelers from their origins to destinations, massive traffic is contributed by vacant RV trips that originate from the end of one customer trip to the start of the next. Different from existing models, our model takes into account two major types of vacant trips generated by RVs, i.e. cruising and deadheading. We first present a basic model by considering intra-node matching between customers and idle RVs. The model characterizes a straightforward extension from Yang and Wong (1998) and is only capable of capturing the congestion effect of cruising trips. To cope with distant matching, a common practice adopted by ride-sourcing platforms, we extend the basic model to incorporate inter-node matching. Using an analogy to electricity circuits, we propose inter-node matching function to handle the spatial interaction between neighboring zones in the matching process. Such a specification, as empirically validated using the real-world data from DiDi, endows us with higher flexibility in representing matching strategies of platforms and depicting the movements of vacant RVs. An iterative solution procedure is proposed to solve for the resulting network equilibrium. The numerical examples on the small-scale Nguyen-Dupius network and the realistic Friedrichshain network demonstrate the feasibility of applying our model to predict the performance of a ride-sourcing system and capture its impact on network traffic conditions.

For future research, we plan to develop a guidance for transportation planning agencies on specifying the values of the parameters in the the proposed inter-node matching function. In fact, as we mention in Section 6.3.3, empirically calibrating the matching function may encounter the endogeneity problem, because the specified covariates are deeply connected in a real system. Special model constructions are thus needed to help fully identify the impacts of different market factors. We hope that the proposed modeling framework can help government agencies with their policy-making in regulating or managing a ride-sourcing system.

CHAPTER 7

Consequence Mitigation

Parking Provision for Ride-Sourcing Vehicles

7.1 Introduction

As ride-sourcing services become increasingly popular in metropolitan areas, the primary focus of this chapter is about the congestion effect of the service. Although their aggregate congestion externality remains controversial (Anderson, 2014; Li et al., 2016; Nie, 2017), the cruising of vacant ride-sourcing vehicles (RVs) generates additional traffic demand that may worsen traffic conditions. Since there are only a limited number of spaces in downtown areas for RVs to park, vacant RVs can only cruise on roads while waiting to be matched with the next customer. In a road network filled with RVs, e.g., the San Francisco downtown area served by many Uber or Lyft vehicles, such cruising contributes additional traffic demand and slows down other traffic, making already-congested streets even worse. Such a phenomenon is expected to become much more severe when emerging shared-use mobility services play a major role in meeting travel needs in metropolitan areas. One intuitive solution is to segment a certain portion of road space to provide on-street parking for RVs to reduce cruising; however, such a plan also reduces road capacity and possibly yields more congestion. The research question is therefore how to find appropriate balance to determine the optimal provision of parking spaces for a given road network.

To answer such a "big picture" question or make a "sketchy decision" at a highly aggregated level (Daganzo et al., 2012), this chapter develops a macroscopic modeling framework that uses only a few parameters but attempts to capture the major underlying mechanisms of the ride-sourcing system under parking provision. Considering a hypothetical matching mechanism adopted by the platform, we further materialize the framework and then apply it to study the interactions between the ride-sourcing system and parking provision under various market structures. The next section first reviews relevant studies and differentiates this paper from other existing endeavors in modeling curbside parking for passenger cars. Section 7.3 establishes a conceptual framework, integrating components concerning each aspect of a ride-sourcing system.

In Section 7.4, we employ a deductive approach to further materialize the framework based on a simplified matching mechanism between RVs and customers. Section 7.5 then applies the derived model to formulate the optimal parking provision strategies as mathematical programs. Numerical experiments are generated to compare the properties and performance of ride-sourcing systems and the corresponding parking provision strategies under two different market structures. Section 7.6 concludes the chapter and discusses future research directions.

7.2 Literature Review

To the best of our knowledge, none of the previous studies has explicitly considered the idea of providing on-street parking spaces for RVs. Two relevant facilities that may come to mind are taxi stands and on-street parking for private cars. Therefore, the rest of this section briefly reviews the state-of-the-art regarding each facility and then emphasizes how the system therein differs from the concept we currently discuss.

Taxi stands are prescribed places where taxis can form queues and serve customers in a first-come-first-serve fashion (Salanova et al., 2014). They are usually established in hotspot areas such as railway stations, airports and hotel or mall entrances, to reduce the uncertainty associated with the meeting between taxis and customers (Salanova et al., 2011). In the literature, taxi stands are traditionally modeled with the queueing theory, by using either single-ended queues where customers arrive randomly and are served sequentially by taxis (e.g. Manski and Wright, 1976) or double-ended queues that capture random arrivals of both sides (e.g., Kashyap, 1966). Based on a double-ended queue framework with a limited capacity for waiting taxis, Matsushima and Kobayashi (2006, 2010) subsequently investigated the endogenous formation and equilibria of markets with single and multiple taxi stands. They found that the market efficiency can potentially be hampered if the capacity of waiting taxis at a taxi stand is not restricted. Apart from the aforementioned structural models, aggregate models were proposed to delineate taxi drivers' choices between waiting at taxi stands and serving relatively uncertain customer demands via either radio dispatch (Schroeter, 1983) or street search (Yang et al., 2010). Recently, another group of studies empirically assessed the relation between supply and passenger demand at taxi stands and provided policy implications on the number and locations of taxi stands in urban regions (e.g. Cooper et al., 2010; Wong et al., 2014a). However, none of previous studies has considered the impact of taxi stands on traffic congestion. Little is known about how the cruising of taxis would be affected by a newly developed stand in a downtown area.

Different from taxi stands, parking spaces in this paper are provided for RVs to reduce cruising and the costs associated with it. Passengers do not come to those parking spaces to meet RVs. Instead, all matchings are made by the online platform. Staying parked or cruising, an RV

experiences little difference in the probability of being matched with a nearby customer, as the matching is made with a sufficiently large radius. Consequently, a ride-sourcing driver would likely seek for parking after dropping off a passenger. In contrast, taxis may not take taxi stands even if they are available, because street hailing may possibly yield higher profits (Wong et al., 2014a).

In the investigation of curbside parking for private cars, a crucial consideration is the cruising of private cars when searching for parking. Following the work of Arnott and Inci (2006), researchers have developed various theoretical frameworks (e.g., Arnott and Rowse, 2009; Geroliminis, 2015; Cao and Menendez, 2015; Liu and Geroliminis, 2016 to model the impact of cruising on traffic congestion and propose potential solutions to it. The most relevant to our discussions are Arnott and Inci (2006, 2010) and Arnott et al. (2015). The former discussed the equilibrium and stability of a road traffic system in which surface space can be allocated between road and parking. The latter investigated the optimal pricing and capacity of downtown on-street parking, with or without garage parking as a supplementary choice. While we account for the same cruising effect, the parking system of interest exhibits different characteristics. First of all, different from parking provision for private cars where pricing can be introduced to regulate parking demand, curbside parking spaces discussed in this chapter are provided for free to RVs, as a tool to reduce congestion and encourage shared-use mobility. The congestion impact of these parking spaces are much more complicated, as the ride-sourcing system pertains to a two-sided market built on the interactions between RVs, customers and the background traffic. Secondly, parking is a must for private cars but optional for RVs. Actually, the relations between travel demand and parking demand are opposites in the parking systems for private cars or RVs. In the former, larger travel demand implies larger parking demand. On the contrary, a higher level of passenger demand likely yields less waiting time for RVs, thereby reducing their parking needs.

As shared-use mobility services will play a more important role in serving mobility needs in metropolitan areas, we envision that parking provision for shared-use vehicles will become a viable policy option to cope with the issue of unproductive empty miles. The methodology we developed here will then help government agencies with such critical policy making.

7.3 Conceptual Modeling Framework

We consider an e-hailing ride-sourcing system operating under the provision of dedicated parking spaces to RVs. In the system, after serving a customer, an RV starts searching for an empty parking space (shown as Fig. 7.1). The process terminates with the RV either successfully finding a parking space or being matched to a waiting customer by an online platform. More specifically, if a new dispatch order arrives before the RV finding an open parking space, it stops parking search and

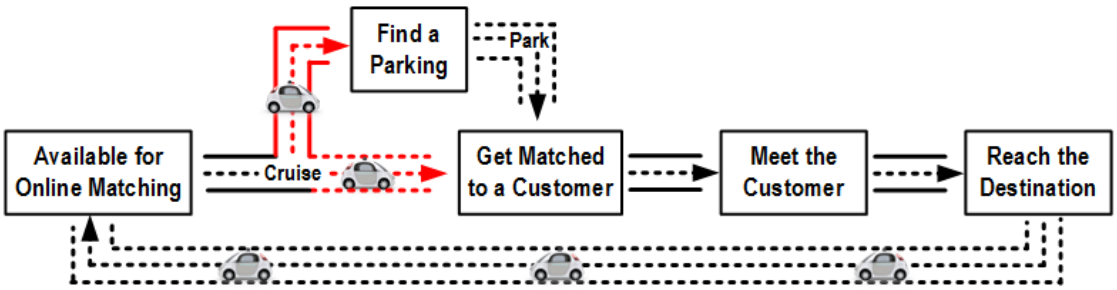


Figure 7.1: Workflow of an RV with e-hailing

starts serving the customer; otherwise, the RV will cruise until finding a vacant space, and then park there and wait for the next dispatch.

To facilitate the presentation, we decompose the framework into five components including three processes and two characteristic functions. The three processes capture respectively the aggregate matching between waiting RVs and customers, the meeting of matched RVs and customers, and the search for empty parking spaces by cruising RVs. The first characteristic function is the customer demand function relating the customer hourly arrival rate to the average cost perceived for each trip. The second characteristic function delineates the relationship between the system's average speed and density under a network or macroscopic fundamental diagram (e.g., Mahmassani et al., 1984; Geroliminis and Daganzo, 2008). The following presents each component.

7.3.1 Matching process of RVs versus customers

An aggregate matching function can be used to characterize the online matching frictions between unmatched RVs and customers (e.g., Douglas, 1972). To capture the competition among drivers and customers, we apply the following matching function suggested by Yang and Yang (2011):

$$m^{c-t} = M^c(J^u, N^u)$$

where m^{c-t} represents the rate of matchings per hour; J^u and N^u denote the number of unmatched customers and RVs, respectively.

7.3.2 Meeting process of RVs versus customers

After the online match, the time before the physical meeting, denoted as w^m, is identical for the matched RV and customer. Intuitively, the average meeting distance r^m is a function of the densities of both unmatched RVs and customers. Assuming the total area served by RVs is

constant, we have:

$$r^m = D^m\left(J^u, N^u\right)$$

where the function D^m depends on the specific matching mechanism adopted by the online platform. Denote v as the average network running speed. Then, the average meeting time can be written as

$$w^m = \frac{D^m\left(J^u, N^u\right)}{v}$$

7.3.3 Search process of RVs versus vacant parking spaces

The search process of cruising RVs for vacant parking spaces can be aggregately viewed as another matching process, and we thus have:

$$m^{p-t} = M^p\left(K^v, N^c, v, w^t\right)$$

where m^{p-t} is the rate of matchings per hour; K^v is the number of vacant parking spaces; N^c is the number of cruising RVs; and w^t is average waiting time for an RV to be matched with a customer. Different from online matching, the friction for a cruising RV to meet with a vacant parking space also depends on travel time between them (and thus the network running speed v) (Liu and Geroliminis, 2016). Moreover, the parking search process can be terminated if the RV gets matched first and thus the matching rate of parking m^{p-t} will depend on how quickly the online customer matching can take place, i.e., w^t.

7.3.4 Customer demand function

Let Q denote customer demand per hour and w^c be the average waiting time for customers getting matched; L_t, F and C are the travel distance, fare and total monetary cost for each trip, respectively. Define β and γ respectively as the out-of-vehicle and in-vehicle value of time (\$/h). Then, the total monetary cost C is given as

$$C = F + \beta \cdot \left(w^c + w^m\right) + \gamma \cdot \frac{L_t}{v}$$

Assume the customer demand Q is a monotonically decreasing function of the trip cost C. Then, we have

$$Q = f\left(F + \beta \cdot \left(w^c + w^m\right) + \gamma \cdot L_t \cdot v^{-1}\right)$$

where $f' < 0$.

7.3.5 Speed versus density

Instead of assuming a constant travel time, we apply a macroscopic or network fundamental diagram to relate the average running speed with the average RV density within the network. As demonstrated by Geroliminis and Daganzo (2008) using traffic data from the Yokohama network, the plot of average network speed versus density follows closely a negatively correlated curve. Consider an isotropic[1] area served by RVs, we obtain the following function V specified by the macroscopic or network fundamental diagram:

$$v = V\left(\frac{d^t}{k_j}\right) \tag{7.1}$$

where d^t represents the average running vehicular density of the network, and k_j is the jam density. Assuming the occupied and cruising vehicles do not differentiate in speed, we then further write the average vehicular density d^t of the network as:

$$d^t = \frac{N - N^p + N^b}{A_u - A^p} \tag{7.2}$$

where N, N^p and N^b are three variables representing the total number of RVs, parked RVs, and background vehicles, respectively; the constant A_u indicates the total utilizable area for roads and curbside parking, while the variable A^p represents the area used for parking spaces. The numerator and the denominator of Eq. (7.2)'s right-hand side respectively represent the total number of running vehicles within the network as well as the effective area of road space. To further derive the framework, we make the following two assumptions:

Assumption 7.1. *The background traffic demand is a monotonically increasing function of the network average speed.*

Assumption 7.2. *The jam density is identical to the density in a fully-occupied parking lot.*

Denote Q^b as the background traffic demand. We then materialize Assumption 7.1 with the following demand function g:

$$Q^b = g\left(\frac{L_b}{v}\right)$$

where $g' < 0$, and L_b is the average travel distance of the background traffic. By replacing N^b and A^p with their equivalences, we rewrite Eq. (7.2) as:

$$d^t = \frac{k_j\left(N - N^p + g\left(\frac{L_b}{v}\right) \cdot \frac{L_b}{v}\right)}{k_j A_u - K} \tag{7.3}$$

[1]The definition for "isotropic" simply follows Arnott and Inci (2010), which implies the density of all components is uniformly distributed over the network throughout the time.

where K is the total number of parking spaces. Thus, substituting Eq. (7.3) into (7.1) yields:

$$v = V\left(\frac{N - N^p + g\left(\frac{L_b}{v}\right) \cdot \frac{L_b}{v}}{k_j A_u - K}\right)$$

7.3.6 Stationary system states under parking provision

Below we summarize all relationships obtained so far:

$$m^{c-t} = M^c\left(J^u, N^u\right) \tag{7.4a}$$

$$w^m = \frac{D^m\left(J^u, N^u\right)}{v} \tag{7.4b}$$

$$m^{p-t} = M^p\left(K^v, N^c, v, w^t\right) \tag{7.4c}$$

$$Q = f\left(F + \beta \cdot \left(w^c + w^m\right) + \gamma \cdot \frac{L_t}{v}\right) \tag{7.4d}$$

$$v = V\left(\frac{N - N^p + g\left(\frac{L_b}{v}\right) \cdot \frac{L_b}{v}}{k_j A_u - K}\right) \tag{7.4e}$$

In this chapter, we focus on the stationary state of the system condition under which the following three relations always hold: $m^{c-t} = Q,\ J^u = w^c Q$ and $N^u = w^t Q$. Replacing $m^{c-t},\ J^u$ and N^u in Eq. (7.4a,b) with their corresponding equivalences yields:

$$Q = M^c\left(w^c Q,\ w^t Q\right)$$
$$w^m \cdot v = D^m\left(w^c Q,\ w^t Q\right)$$

To simplify the framework, we further make the following assumption:

Assumption 7.3. *The states of unmatched RVs and customers do not affect their probability of getting matched.*

This assumption implies that no matter being cruising or parked, unmatched RVs will have the same probability of being matched; customers, walking or motionless, will be matched with equal probability. This assumption appears reasonable because the matching radius utilized by the online platform is usually large enough so that RVs and customers are comparably stationary. Moreover, the area considered in this model is assumed to be spatially isotropic in terms of both customer demand and parking provision, under which RVs do not have intentions for chasing hotspots. Assumption 7.3 then leads to the following relationship:

$$\frac{m^{p-t}}{m^{c-t}} = \frac{N^u - N^c}{N^u} \tag{7.5}$$

Substituting Eq. (7.5) into (7.4c) and replacing $m^{c-t},\ N^u$ with their equivalences yield:

$$Q - \frac{N^c}{w^t} = M^p\left(K^v, N^c, v, w^t\right)$$

Further, the conservation of fleet size requires

$$N = Q \cdot \left(w^t + w^m + \frac{L_t}{v}\right)$$
$$N^u = Qw^t = N^c + N^p$$

and the conservation of parking spaces gives

$$K = K^v + N^p$$

Hence, by replacing K^v and N^c in Eq. (7.4c) with their equivalences and summarizing all the independent relations, we obtain the following conceptual modeling system:

$$Q = M^c\left(w^c Q,\ w^t Q\right) \tag{7.6a}$$
$$w^m \cdot v = D^m\left(w^c Q,\ w^t Q\right) \tag{7.6b}$$
$$\frac{N^p}{w^t} = M^p\left(K - N^p, Qw^t - N^p, v, w^t\right) \tag{7.6c}$$
$$N = Q \cdot \left(w^t + w^m + \frac{L_t}{v}\right) \tag{7.6d}$$
$$Q = f\left(F + \beta \cdot (w^c + w^m) + \gamma \cdot \frac{L_t}{v}\right) \tag{7.6e}$$
$$v = V\left(\frac{N - N^p + g\left(\frac{L_b}{v}\right) \cdot \frac{L_b}{v}}{k_j A_u - K}\right) \tag{7.6f}$$

There are nine unknowns in total, i.e., the customer and RV waiting time for getting matched w^c and w^t, the meeting time after being matched w^m, the number of parked RVs N^p, the customer demand Q, the network average speed v, the fare F, the total fleet size N and the total number of parking spaces K. If the latter three variables are specified by stakeholders, e.g., a government agency determines the parking provision K while a ride-sourcing service provider decides its price F and fleet size N, the solution to the nonlinear system (7.6a-f) prescribes an estimate of the performance of the ride-sourcing system and road network.

7.4 Deductive Framework Realization

The conceptual framework established in Section 7.3 is flexible enough to model systems under various operating conditions. To facilitate the investigation of the interactions between the e-hailing ride-sourcing and parking systems, a simplified matching mechanism of the platform is considered in this section, which allows us to analytically establish the relationships between variables in the matching processes. Subsequently, a deductive realization of the conceptual framework is derived, and will be applied to examine various scenarios in later sections.

7.4.1 Online matching of RVs versus customers

The online matching algorithm can be simplified as matching a requesting customer to his or her closest RV within the matching radius (Xu et al., 2018). To delineate the process, we discretize the whole analysis period (e.g., one hour) into time steps of equal length τ and consider the following simplified process:

- At the beginning of each time step, every unmatched customer and new arrival sends a request to his/her closest unmatched RV through the platform, and remains stationary at the current location;
- The platform loops through all the unmatched RVs. For each RV, it checks whether the RV has been requested by any customers. If yes, it dispatches the RV to its closest requesting customer; otherwise, it does nothing;
- The matched pairs disappear from the spatial matching network or pool while those unmatched remain until the next step.

The process essentially implies that at each time step, the platform matches all the mutually closest pairs of RVs and customers at that moment, while those pairs in which at least one side is not "optimal" for the other remain unmatched until the next step.

For each unmatched RV, we define its dominant zone as a neighboring area where the distance from any point inside to the central RV is shorter than the distance to any other unmatched RVs on the map (see Fig. 7.2). Let J_d^u and N_d^u denote the density of unmatched customers and RVs respectively. It is assumed that the distribution of RVs is uniform. Therefore, the area of each RV's dominant zone is approximately equal to N_d^{u-1}. On the other hand, the distribution of unmatched customers is assumed to follow the spatial Poisson point process (Chiu et al., 2013). Thus, the probability of the existence of n unmatched customers inside the dominant zone of an RV is

$$P\{n\} = \frac{\exp\left(-\frac{J_d^u}{N_d^u}\right) \cdot \left(\frac{J_d^u}{N_d^u}\right)^n}{n!}, n = 0, 1, 2, \cdots$$

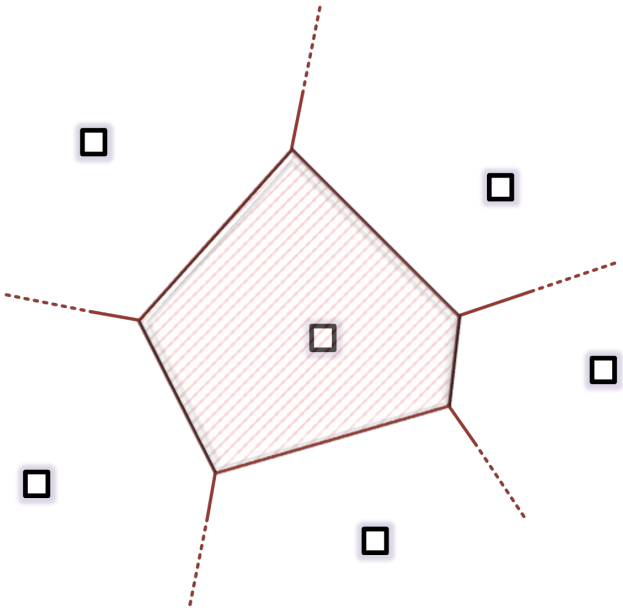


Figure 7.2: Dominant zone of an unmatched RV. The figure displays a portion of the spatial matching network, where squares and lines respectively represent unmatched RVs and boundaries of their dominant zones. The shadowed area in the middle denotes the dominant zone of the central RV.

This leads to the following probability of an RV getting matched at each time step, denoted as P_m:

$$P_m = 1 - P\{0\} = 1 - \exp\left(-\frac{J_d^u}{N_d^u}\right)$$

Given the length τ of each matching step, the RVs' expected waiting time for being matched is

$$w^t = \frac{\tau}{1 - \exp\left(-\frac{J_d^u}{N_d^u}\right)} \tag{7.7}$$

Since the relationship $J_d^u/N_d^u = w^c/w^t$ always holds under a stationary process, we then substitute it into Eq. 7.7 and obtain

$$w^c = \frac{\tau}{\frac{N_d^u}{J_d^u} \cdot \left[1 - \exp\left(-\frac{J_d^u}{N_d^u}\right)\right]}$$

Hence, both average matching time functions $w^c(J_d^u, N_d^u)$ and $w^t(J_d^u, N_d^u)$ are homogenous of degree zero and depend only on the ratio J_d^u/N_d^u. If $J_d^u \gg N_d^u$, we have $w^t \cong \tau$ and $w^c \cong J_d^u \cdot \tau / N_d^u$. This means that when the size of the customers dominates the RV fleet size, RVs can be matched almost immediately after they become unoccupied, while customers have to wait for serval steps before they get matched, and vice versa. A detailed explanation for the relationship between w^t and w^c is given in Appendix D.2.

7.4.2 Meeting distance of matched RVs versus customers

Using the same matching mechanism defined in Section 7.4.1, we derive the form of meeting distance function $D^m\left(J^u, N^u\right)$ in this section. For an unmatched RV, we denote by x the distance of the closest unmatched customer with a cumulative distribution function $H\left(\cdot\right)$ and density function $h\left(\cdot\right)$. Since the customers' location follows the spatial Poisson point process, we obtain $H\left(\cdot\right)$ and $h\left(\cdot\right)$ as follows:

$$H\left(x\right)=1-P\left\{0\right\}=1-\exp\left(-\pi x^2 J_d^u\right),\ x\geq 0$$

$$h\left(x\right)=2\pi x J_d^u\cdot\exp\left(-\pi x^2 J_d^u\right),\ x\geq 0$$

As previously assumed, the unmatched RVs are uniformly distributed. This enables us to further approximate the dominant zone of each unmatched RV as a disc, whose radius r is $1/\sqrt{\pi\cdot N_d^u}$. Then, the matching process of an RV can thus be simplified as matching the closest customer within distance r. So, the expected meeting distance is given as

$$\begin{aligned} r^m\left(J_d^u,\, N_d^u\right) \cong & \mu\cdot H(r)^{-1}\cdot\left[\int_0^r x\cdot h\left(x\right)\mathrm{d}x\right] \\ = & \frac{\mu}{1-\exp\left(-\pi r^2 J_d^u\right)}\cdot\left[\frac{1}{2\sqrt{J_d^u}}\cdot\mathrm{erf}\left(\sqrt{\pi J_d^u}\cdot r\right)-r\cdot\exp\left(-\pi r^2 J_d^u\right)\right] \\ = & \frac{\mu}{1-\exp\left(-\frac{J_d^u}{N_d^u}\right)}\cdot\left[\frac{1}{2\sqrt{J_d^u}}\cdot\mathrm{erf}\left(\sqrt{\frac{J_d^u}{N_d^u}}\right)-\frac{1}{\sqrt{\pi N_d^u}}\cdot\exp\left(-\frac{J_d^u}{N_d^u}\right)\right] \end{aligned} \tag{7.8}$$

where $\mathrm{erf}(x)=\frac{2}{\sqrt{\pi}}\int_0^x e^{-t^2}\mathrm{d}t$ is the Gauss error function, and $\mu\approx\sqrt{\frac{\pi}{2}}$ is a scaling parameter that adjusts the Euclidean distance to the Manhattan distance (Arnott, 1996). A detailed extreme condition analysis of formula (7.8) can be found in Appendix D.3.

7.4.3 Probability for cruising RVs finding vacant parking spaces

The search process of a cruising RV for a vacant parking space ends with the RV finding either a parking space or a customer. As the occurrences of two outcomes are independent, we can derive the distribution function for each outcome separately.

As mentioned previously, the matching is carried out at interval τ, and the probability of an unmatched RV getting matched at each step is $1-\exp\left(-J_d^u/N_d^u\right)$. This implies that the process of an RV getting matched is a discrete-time Markov process. Given τ being small enough, we approximate the process in continuous time. Let t_c be the time for an RV being matched to a customer, and $G\left(t_c\right)$ be the corresponding cumulative distribution function. Then, $G(t_c)$ can be

written as

$$G\left(t_c\right)=1-e^{-\frac{t_c}{w^t}}$$

Similarly, following (Arnott and Rowse, 1999), the process of an RV finding a parking space is also Markovian, and the time duration t_p is subject to an exponential distribution as well, i.e.,

$$p\left(t_p\right)=\frac{1}{w^f}\cdot e^{-\frac{t_p}{w^f}}$$

where $p\left(t_p\right)$ denotes the probability density function and w^f is the expected time duration.

Let p^p denote the probability of an unmatched RV finding a parking space before being matched with a customer. Then, p^p is derived as follows

$$p^p=\int_0^{+\infty} p\left(t\right)\cdot\left[1-G\left(t\right)\right]\cdot\mathrm{d}t=\frac{w^t}{w^f+w^t}$$

Thus, we have

$$\frac{N^p}{w^t}=Q^p=Q\cdot p^p=\frac{Qw^t}{w^f+w^t} \tag{7.9}$$

Let L_r denote the total length of the road system. Then, the average distance between two adjacent empty parking spaces is L_r/K^v. Thus, the expected time for finding a parking space w^f can be approximated as

$$w^f\cong\frac{L_r}{\left(K-N^p\right)\cdot v} \tag{7.10}$$

Combining Eqs. (7.9) and (7.10) yields

$$Q\cdot {w^t}^2=N^p\cdot\left(w^t+\frac{L_r}{\left(K-N^p\right)\cdot v}\right)$$

7.4.4 The deductive modeling system

As for the demand function Eq. (7.6e), we employ the following form used by Yang and Yang (2011):

$$Q=f\left(F+\beta\left(w^c+w^m\right)+\gamma\frac{L_t}{v}\right)=Q^0\cdot\exp\left[-\kappa\left(F+\beta\left(w^c+w^m\right)+\gamma\frac{L_t}{v}\right)\right]$$

where Q^0 indicates the potential hourly customer demand and κ represents the cost elasticity.

Regarding the speed function Eq. (7.6f), we adopt a Greenshields-type function (Greenshields et al., 1935) to approximate the macroscopic or network fundamental diagram for the network

considered in this paper, i.e.,

$$v{=}v_f{\cdot}\left(1{-}\frac{N{-}N^p{+}g\left(\frac{L_b}{v}\right){\cdot}\frac{L_b}{v}}{k_jA_u{-}K}\right)$$

where the background traffic demand function is again specified in an exponential form given as:

$$g\left(\frac{L_b}{v}\right)=Q_b^0{\cdot}\exp\left(-\kappa_b\cdot\frac{L_b}{v}\right)$$

where we use Q_b^0 to indicate the potential hourly demand for background traffic and κ_b to represent the travel time elasticity. Additionally, J_d^u and N_d^u are subject to the following relationships,

$$J_d^u{=}\frac{J^u}{A_w}{=}\frac{Qw^c}{A_w}$$

$$N_d^u{=}\frac{N^u}{A_w}{=}\frac{Qw^t}{A_w}$$

where A_w denotes the total area served by RVs.

By combining all the formulas derived above in this section and replacing J_d^u and N_d^u with their corresponding equivalences, we obtain an analytically tractable model for describing the e-hailing ride-sourcing system under parking provision:

$$w^t{=}\frac{\tau}{1{-}\exp\left(-\frac{w^c}{w^t}\right)} \tag{7.11a}$$

$$w^m{\cdot}v{=}\frac{\mu}{1{-}\exp\left(-\frac{w^c}{w^t}\right)}{\cdot}\left[\frac{1}{2}\sqrt{\frac{A_w}{Qw^c}}{\cdot}\mathrm{erf}\left(\sqrt{\frac{w^c}{w^t}}\right)-\frac{1}{\sqrt{\pi}}\sqrt{\frac{A_w}{Qw^t}}\cdot\exp\left(-\frac{w^c}{w^t}\right)\right] \tag{7.11b}$$

$$Q{\cdot}{w^t}^2{=}N^p{\cdot}\left(w^t{+}\frac{L_r}{(K{-}N^p){\cdot}v}\right) \tag{7.11c}$$

$$N{=}Q{\cdot}\left(w^t{+}w^m{+}\frac{L_t}{v}\right) \tag{7.11d}$$

$$Q{=}Q^0{\cdot}\exp\left[-\kappa\left(F{+}\beta\left(w^c{+}w^m\right){+}\gamma\frac{L_t}{v}\right)\right] \tag{7.11e}$$

$$v{=}v_f{\cdot}\left(1{-}\frac{N{-}N^p{+}Q_b^0{\cdot}\exp\left(-\kappa_b\cdot\frac{L_b}{v}\right)\cdot\frac{L_b}{v}}{k_jA_u{-}K}\right) \tag{7.11f}$$

The existence of equilibrium solutions is proved for the above nonlinear system in Appendix D.4. We note that the solution is generally not unique. Multiple solutions may appear in both non-congested and congested regimes. However, those equilibrium solutions falling deeply into the congested regime are dynamically unstable and usually unable to sustain (Small et al., 2007).

Hence, in this paper, we confine the system to solve only within the non-congested and slightly-congested regimes. Our numerical experiments suggest that the resultant solutions are always unique for different combinations of (F, N, K) set with rational values.

7.4.5 Numerical example for impacts of parking provision

In this numerical example, we consider a Manhattan-like area of 20 square miles served by 50,000 RVs. The average trip distance of ride-sourcing orders is 2.5 miles, and each trip on average costs \$15. Other parameters used in the example can be found in Appendix D.1. We then vary the number of parking spaces K in the area to investigate its impacts on system performance. Figure 7.3 displays the changes of multiple performance measures, including the RV's waiting time for

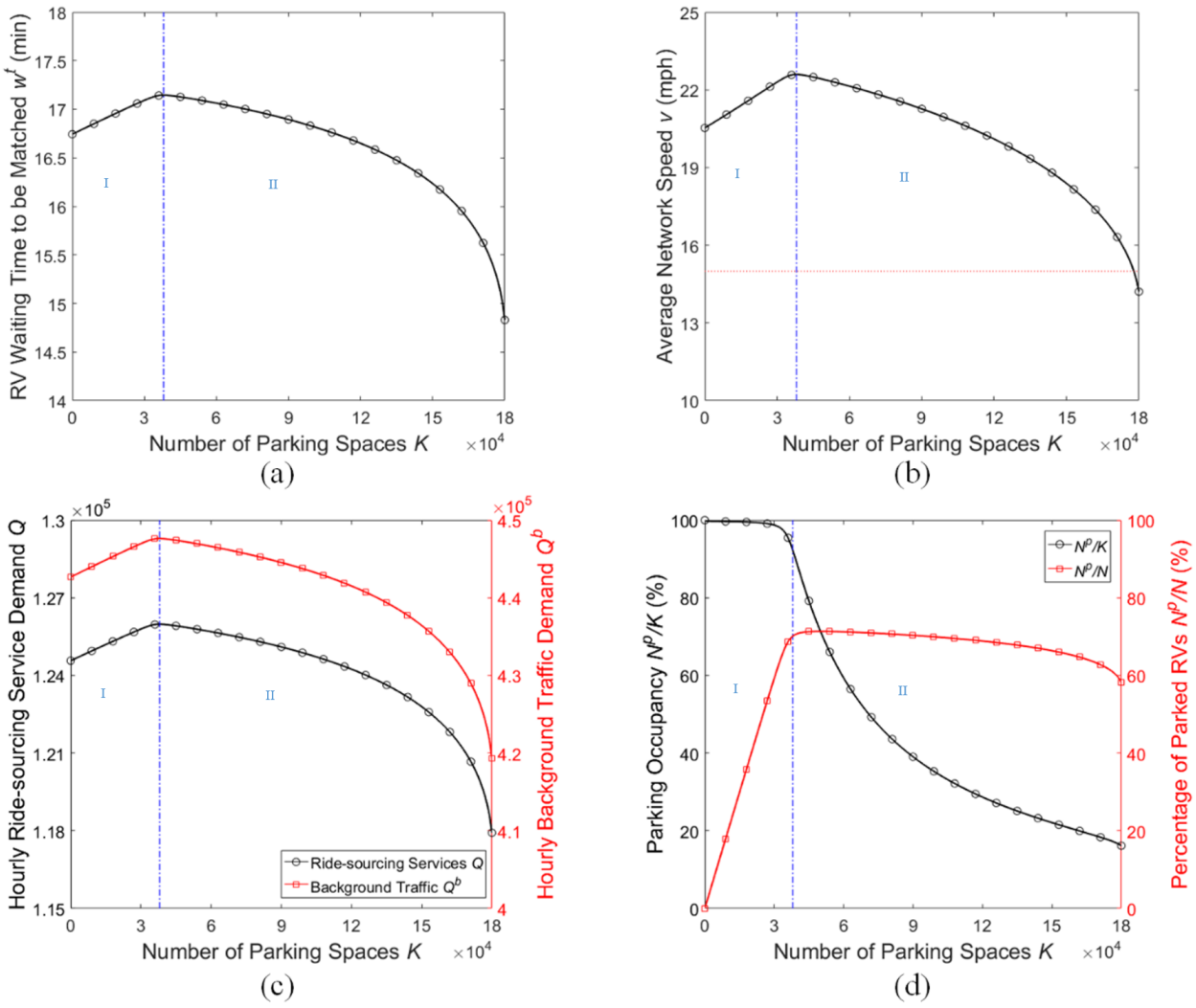


Figure 7.3: Impacts of the parking provision on the ride-sourcing system

matching w^t (Fig. 7.3a), the network average speed v (Fig. 7.3b), the realized hourly ride-sourcing and background traffic demand Q and Q^b (Fig. 7.3c), the parking space occupancy N^p/K and the instantaneous percentage of parked RVs N^p/N (Fig. 7.3d), respectively.

The figures present a clear trade-off between the reduction of cruising and capacity drop caused by the allocation of road space to parking. As shown in Figures (7.3b) and (7.3c), the average speed v and customer serving rate Q increase initially and peak at the parking supply K being approximately 3.8×10^4, then decrease. This suggests that $K= 3.8\times10^4$ is a threshold where the system performance improves with respect to the increase in parking provision on the left side and gets worse on the right side. More specifically, we divide the parking provision by $K= 3.8\times10^4$ into two areas, respectively indexed as I and II in Figure 7.3. In area I, the system is short of parking spaces, with almost all of the provided parking spaces occupied by waiting RVs (Fig. 7.3d). The number of parked RVs thus grows linearly with respect to the number of parking provision (Fig. 7.3d). Consequently, the reinforcement of parking provision leads to the rising of network speed and the reduction of service time per customer, which in turn results in more RVs staying vacant. Therefore, the RV's waiting time for being matched w^t increases in area I (Fig. 7.3a). In area II, by contrast, the additional provision of parking spaces can no longer attract more RVs to park (Fig. 7.3d) but consume the precious road space and leads to more congestion (Fig. 7.3b). As a result, the service time per customer increases, and the customer demand and the number of vacant RVs decrease accordingly (Fig. 7.3c and 7.3d). Thus, the oversupply of parking spaces only facilitates the deterioration of the traffic condition, which causes the unnecessary waste of labor supply in the ride-sourcing system. Interestingly, Figure 7.3c shows that under a fixed fare and RV supply, parking provision imposes the same effect on the demand for ride-sourcing services and the background traffic. While better traffic condition will undoubtedly benefit the background traffic, reduced trip-serving time for a ride-sourcing system also means less travel cost for customers and higher demand to serve for drivers.

7.5 Time-Varying Optimal Parking Provision

Given that travel demand highly fluctuates over a day, it is reasonable to require the parking provision to adjust accordingly. The implementation of time-varying curbside parking provision can be similar to a reversible lane: put a traffic light at the head of the lane, which turns red during off-peak hours to declare the lane as a parking lane for RVs, and becomes off during peak hours to give the lane back to the running traffic. Assuming traffic patterns are stable across days, and each day can be divided into I periods with equal length such that the traffic demand within each period becomes relatively stable. Let Q_i^0 and $Q_{b,i}^0$ respectively denote the potential hourly demand of ride-souring service and background traffic at period $i \in I$, which differ from period to period.

We then investigate the time-varying optimal parking provision for ride-sourcing systems under different market structures.

7.5.1 Mathematical formulations for optimal strategies

Ride-sourcing companies do not own the fleets themselves, but offer online matching services for those RVs/drivers registered to their platform. This results in two significant features of the ride-sourcing system: 1) the number of RVs in service is unrestricted. As long as there is still room for profit, more RVs can enter the market to provide service, if a sufficient labor supply is assumed; 2) the working time of RV drivers is not restricted. Since drivers in this case are self-employed, they have the flexibility of deciding their working time. Let N_i denote the total number of RVs in service at period i. Then, owing to these features, the fleet size in service N_i varies over periods and is endogenously determined depending on the real-time market condition. Nevertheless, the commission fee charged by the ride-sourcing company empowers it to indirectly control the number of vehicles in service.

Based on the modeling system 7.6, we now investigate the optimal parking provision for the ride-sourcing market. Two market structures are compared in this paper: 1) a system optimum case, where the government controls the fare $\{F_i\}$, the RV fleet size $\{N_i\}$ and the parking spaces $\{K_i\}$ to maximize the social benefit; 2) a monopoly case, where a private ride-sourcing company controls the fare $\{F_i\}$ and the fleet size $\{N_i\}$ to maximize its revenue, while the government only determines the size of parking spaces $\{K_i\}$ to maximize the social benefit. We use subscript i to specify the period each variable applies to. Denote F_i^c as the commission fee that drivers pay to the ride-sourcing company for each order or transaction fulfilled at period i. Then, the total hourly profit for the ride-sourcing company and RV drivers, respectively denoted as Π_i^c and Π_i^d, can be calculated as

$$\Pi_i^c = (F_i^c - c_p) \cdot Q_i \tag{7.12a}$$

$$\Pi_i^d = (F_i - F_i^c) \cdot Q_i - (c_v N_i - c_s N_i^p) \tag{7.12b}$$

where c_p is the unit-order operation cost of the e-hailing platform; c_v represents drivers' hourly cost of operating a RV; c_s denotes the hourly savings for a single parked RV. Thus, the term $c_v N_i - c_s N_i^p$ on the right-hand-side of Eq. (7.12b) represents the total hourly expenses for RV drivers operating their vehicles. Combined with the travelers' benefits, the hourly social benefits S_i for period i can be formulated as:

$$S_i = \int_0^{Q_i} f^{-1}\left(\omega | Q_i^0\right) \mathrm{d}\omega - Q_i \cdot \left[\beta\left(w_i^c + w_i^m\right) + \gamma \frac{L_t}{v_i}\right] - (c_p Q_i + c_v N_i - c_s N_i^p)$$

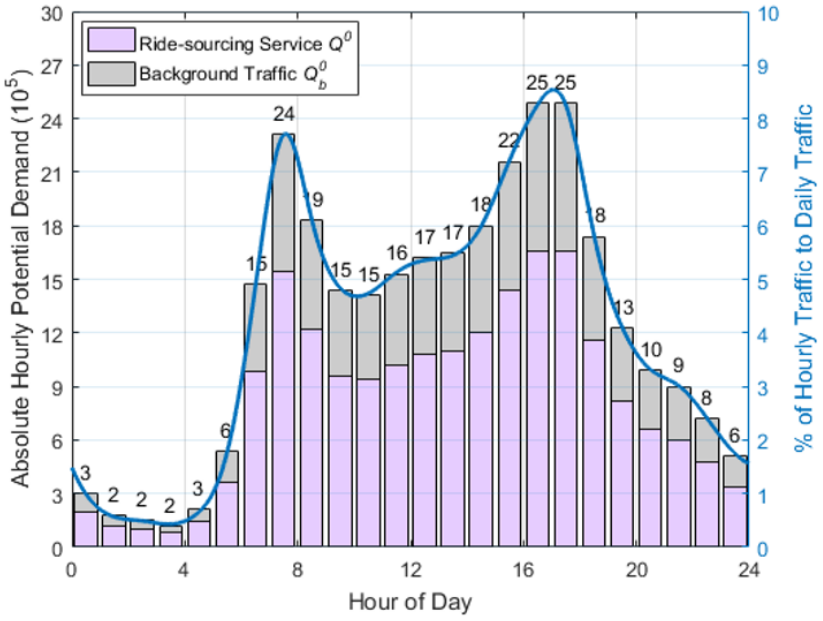


Figure 7.4: Within-day hourly potential demand variations of ride-sourcing service Q^0 and background traffic Q_b^0

$$+ c_b \left[\int_0^{Q_i^b} g^{-1}\left(\omega | Q_{b,i}^0\right) \mathrm{d}\omega - Q_i^b \cdot \frac{L_b}{v_i} \right]$$

where c_b denotes the value of time in the background traffic. The first integral term represents consumers' surplus; the second and third terms respectively indicate the total hourly time costs for customers and operational expenses of the ride-sourcing system; the last term represents the total benefit of travelers in background traffic.

Assume the labor market of drivers is perfectly competitive and each N_i is realized at a level such that $\Pi_i^d = 0$. Thus, the system optimum case of the ride-sourcing market solves the nonlinear optimization model below for each $i \in I$:

$$\max \left\{ S_i \,\middle|\, F_i, N_i, K_i \geq 0;\ \Pi_i^d = 0;\ \text{Eq. (7.6) for period } i \right\} \tag{7.13}$$

And the monopoly case solves the following bi-level program:

$$\max \left\{ S_i \,\middle|\, K_i \geq 0;\ (F_i^*, N_i^*) = \arg\max \left\{ F_i, F_i^c \geq 0;\ \Pi_i^d = 0;\ \text{Eq. (7.6) for period } i \right\} \right\} \tag{7.14}$$

In the above two problems, there are no correlations across periods, and all variables are self-contained by relationships within their corresponding period. Therefore, the problems can be decomposed into each period and solved individually.

7.5.2 Numerical experiments

Consider a Manhattan-like area of 20 square miles served by 50,000 RVs. We assume the potential daily travel demand for ride-sourcing service and background traffic are respectively 20 and 10 million trips, and then split them by hours using the same percentage data (see the blue line in Fig. 7.4) provided by Sloboden et al. (2012). The hourly potential demands are displayed as bars in Fig. 7.4, which present a typical double-peak pattern that peaks at about 8:00 and 17:00. Based on this variable demand pattern, we solve mathematical programs (7.13) and (7.14) to estimate the ride-sourcing system performance under different scenarios.

We present in Fig. 7.5 the daily varying trajectories of six variables, including three determinant variables (F, K, N) and three dependent variables (w^t, v, Π). It can be clearly seen that the two scenarios – the system optimum (SO) and the monopoly (MP) – display distinct trajectories in non-peak hours. Compared to SO, Fig. 7.5a shows that the ride-sourcing company under MP tends to charge a higher trip fare F and commission fee F^c, which yields fewer customers and drivers. This fact is also supported by Figure 7.5e and 7.5b, which demonstrate that although the number of RVs in service N is lower under MP, their waiting time for online matching w^t instead increases. Because the customer demand under MP shrinks even more than the vehicle supply in comparison to SO, the customer's waiting time for online matching w^c decreases. Moreover, since the network average speed v is mostly higher under MP (see Fig. 7.5d), the service time is also much less. Consequently, the customers in MP enjoy a better level of service but with a higher fare, while the longer waiting time of RVs in MP yields a greater need for parking spaces. Evidently, Fig. 7.5c shows that more parking spaces should be provided by the government under MP.

Then, we turn to the distinctions between peak and non-peak hours, which respectively imply high-demand and low-demand scenarios. In response to demand variations, Fig. 7.5a indicates that the company under both SO and MP raises the fare and commission fee at peak hours, then lowers them to encourage ridership when demand drops. The system during peak hours behaves like a seller's market where RVs receive an instant match while customers endure longer waiting times; conversely, at non-peak hours, the system switches to a buyer's market and RVs must wait (Fig. 7.5b). Such a transition can also be witnessed through Fig. 7.5c, where the supply of parking spaces considerably increases for non-peak hours, especially after midnight. The number of RVs should typically increase with respect to the surge of potential demand. However, Fig. 7.5e shows that less RVs are needed during peak hours as compared to the numbers before and after the peak. This counterintuitive phenomenon is primarily due to the congestion caused by the demand surge of background traffic, under which both SO and MP benefit from controlling the market size.

Generally speaking, the strategies taken by the company respectively under SO and MP show more differences in lower demand scenarios at non-peak hours. The reason is that at a higher demand level, although with different purposes, both SO and MP require the company to charge

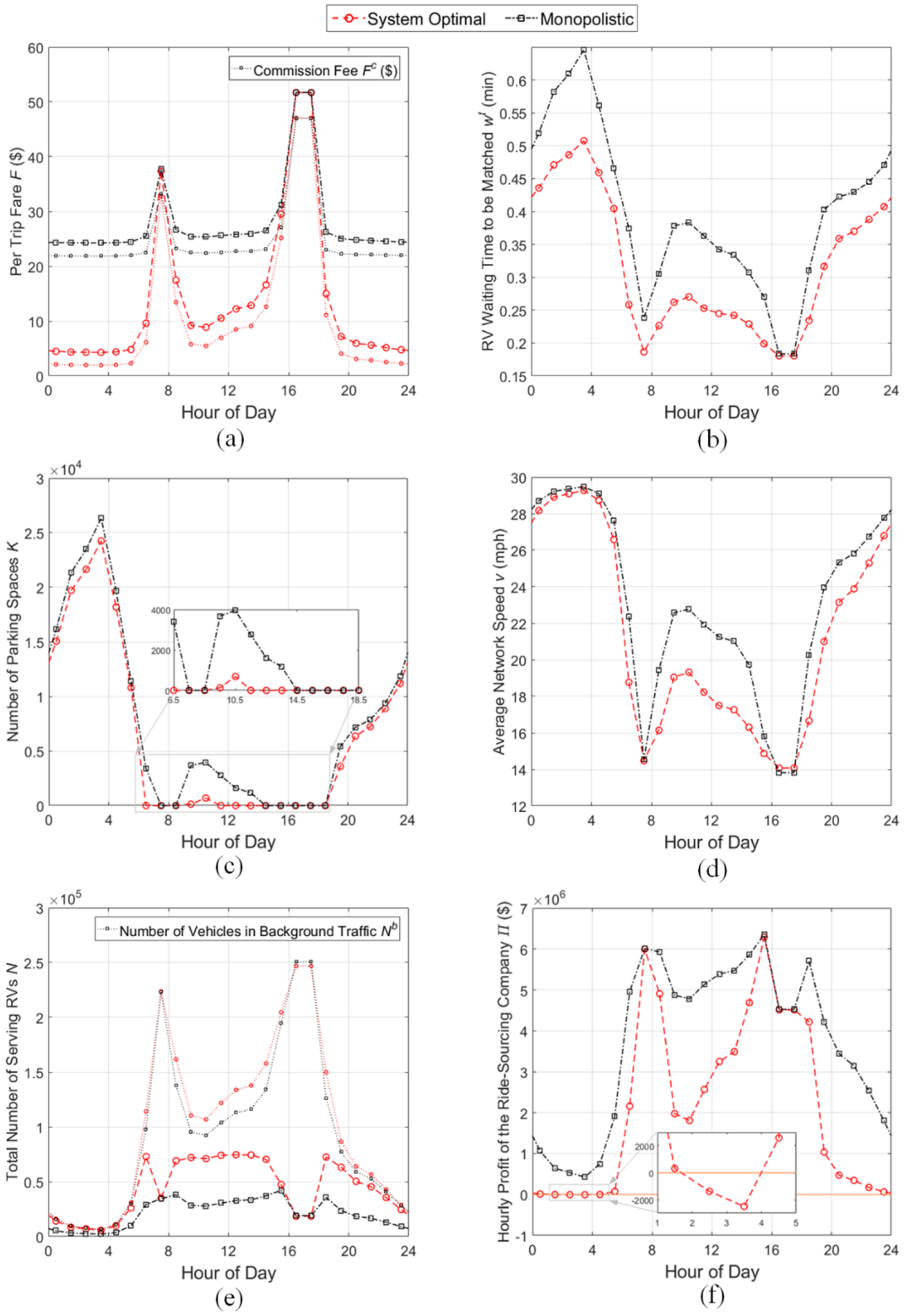


Figure 7.5: Time-varying optimal parking provision for the ride-sourcing system

higher prices. The primary purpose of SO is to suppress excessive demand and reduce customers' long waiting time, but MP is intended to exploit the high demand to enhance profit for the company. Such a contraction of purposes becomes evident at the lower demand level and gives rise to two opposite action directions. While the company under MP still tends to charge higher prices to secure its profit, the system under SO prefers to charge lower prices and use the excessive transport capacity to serve additional customers. Hence, Fig. 7.5f shows that the profitability of the company under the two cases almost overlap during peak hours but diverges significantly at other time.

Furthermore, it is worthy of additional note that in Fig. 7.5f the ride-sourcing services under SO maintains profitable during most time of a day, expect for a few hours after midnight when the demand is extremely low. In other words, when the customer demand is sufficiently high, the ride-souring company can sustain itself. This observation is consistent with the finding in the literature. If the congestion effect is not considered, in a regular taxi market where the taxi trip production is increasing returns to scale, the first-best pricing only covers the cost when a taxi is occupied and thus drivers' profit is negative (e.g., Arnott, 1996). However, when congestion is severe, the first-best pricing may yield positive profit (Yang et al., 2005).

7.6 Summary

The popularization of ride-sourcing services brings not only convenience to customers, but also new burden for road networks owing to the substantial cruising of RVs for customers. This chapter thus has investigated the idea of providing parking spaces for those vacant e-hailing RVs, allowing them to park while waiting for online dispatches. A macroscopic conceptual framework was firstly established to aggregately delineate the e-hailing ride-sourcing system and describe its interactions with the parking provision. To quantitatively measure the system performance, we derived a deductive model based on the conceptual framework considering a simplified matching mechanism of the platform. Two market structures as well as their correspondingly optimal parking management strategies were considered and formulated based on the deductive model. We then conducted numerical experiments to comprehensively delineate the impacts of parking provision and compare the ride-sourcing systems under different scenarios. Here, we highlight three major findings regarding the parking provision strategies for the government:

1. For a given road system, a time-varying optimal parking provision can balance between reduction of cruising for RVs and capacity drop of the road system.

2. For non-peak hours, the optimal number of parking spaces is inversely proportional to the instantaneous customer demand. No space is needed for peak hours with demand surge.

3. Even though there are fewer RVs in service under the monopoly scenario, a larger supply of parking spaces should be provided as compared to the system optimum case.

Our future research will improve and extend the proposed modeling frameworks in the following aspects. First, we assume in this chapter that the ride-sourcing vehicles are uniformly distributed. It would be more realistic to consider other less restrictive stochastic distributions of RVs. Second, the framework can be enhanced if hotspots are captured. The network in this chapter is assumed to be spatially isotropic, while the road network in real life is usually inhomogeneous such that vacant RVs tend to gather around hotspots, cruising there for parking and waiting for customers. Third, the correlations of labor supply and traffic condition between periods are neglected in the current framework. This might be captured by modeling the cross-period traffic and introducing a new dimension of temporal equilibrium on labor supply in a manner similar to Zha et al. (2018a). Fourth, conventional taxis now also embrace the e-hailing technology. If we consider taxis that use both e-hailing and street-hailing and also search for parking spaces, the government's parking management strategy will likely be different. Lastly in the foreseeable future, the advent of autonomous vehicles will sustainably stimulate the adoption of ride-sharing and car-sharing (Kornhauser et al., 2013; Fagnant and Kockelman, 2015; Martinez and Crist, 2015; Krueger et al., 2016). It is thus intriguing to investigate how the government should respond to such transformation in parking provision for those shared automated vehicles. All these extensions are expected to substantially enhance the analytical capability of the modeling framework. However, the challenge remains to be balancing the modeling tractability and behavioral realism.

CHAPTER 8

Conclusions and Future Research

8.1 Research Summary

As a symbolic icon for emerging mobility in recent years, ride-sourcing services, such as Uber, Lyft, and Didi Chuxing, are becoming increasingly important in meeting the travel needs in urban areas. Despite the great success achieved in business, these services have also created many controversies, which arose extensive attention both from academia and among the society. This dissertation focuses on the empty miles produced by ride-sourcing vehicles, which, if not appropriately taken care of, can cause an enormous waste of the precious supply resource and worsen the already severe traffic congestion in urban areas. However, it is not straightforward to address the empty miles directly by boosting customer demand or controlling the fleet size, given the vast complexities behind the system operations. By virtue of multidisciplinary methodologies and comprehensive data from Didi Chuxing, this dissertation methodically examines the interactions among the primary four stakeholders in ride-sourcing systems (including riders, drivers, the platform, as well as the public-sector regulator). We identified, corroborated, and estimated the consequential issues that may lead to devastating empty miles, then proposed countermeasures to resolve them to sustain the network system efficiency. Below summarizes our main discussions and results in view of each major market player:

Riders' Mode/Service Choices [Chapter 2 & 3]

Similar to other mobility services, the demand in ride-sourcing markets is driven by multi-level rider choices over different travel modes (e.g. driving or taking public transportation) and service modes (e.g., riding solo or pooling with others), which can be modeled in standard ways by applying utility theories. Major differences come from the composition of riders' cost utilities. In contrast to the street-hailing counterpart, riders using app-hailing services, after they get matched online, need to spend extra time waiting for drivers to show up physically. The fact that riders now trade off between the triple-tuple costs, i.e. the costs in matching, pickup and delivery stages,

yields brand-new system characteristics, which further contribute to the variations of platforms' managerial strategies under different market structures. We analyzed riders' value of time in different service stages and found that riders weigh the matching delay much higher than the pickup delay, probably due to their aversion on uncertainties. Consequently, platforms would manage to ensure the quick matching (locking) of riders, which as a side effect may lead to devastating empty miles for drivers in pickups. For monopolistic markets, we proved that the platform can undertake the expanse of delaying riders in matching to exchange for shorter pickup time, reduce the empty miles traveled by vehicles, and achieve higher system efficiency.

Drivers' Customer Search Behavior [Chapter 4 & 5]

The flexibility and diversity of freelance drivers' working patterns pose challenges for ride-sourcing platforms in effectively managing the workforce. To pave the basis for proper labor/fleet management, we built structural models and calibrated the behavior of drivers in customer search using comprehensive data from DiDi. Specifically, we distinguish between two types of spatial markets. One is urban areas with service opportunities scattered over the space, where drivers after becoming idle cruise around in search of customers. We developed a cross-nested dynamic discrete choice model to disentangle the continuity in opportunities along different search paths, and also to differentiate the latent behavior of drivers in local search. A clear pattern was identified for drivers in chasing the hotspots, which may cause long-time emptiness in cruising and the deceptive supply shortage at the near-hotspot areas. The other type of market is the islanded areas, such as airports, railway stations, and shopping malls, etc., where a large number of idle drivers accumulate. We applied survival analysis and found the longer time drivers dwell in those areas the more reluctant they are in leaving. More importantly, the results reveal that the provision of exhaustive market information by the platform does not necessarily improve system performance, but may arouse drivers' irrational behaviors causing greater supply wastes in idleness. Efficient information-sharing mechanisms are well-needed, especially between drivers and the platform.

Platform's Operation Strategies [Chapter 2 & 3]

The ride-sourcing service produces a two-sided market, where the platform as a mediator takes strategies to "coordinate" supply/demand and attain its business goals. Among all instruments held by the platform, the most fundamental and influential ones are matching and pricing. In this dissertation, we modeled the complexities underlying these instruments, identified the inefficient operations, and provided countermeasures to restore system efficiency and improve the level of service for both riders and drivers. We first focused on the isotropic market and identified a matching failure that can cause requesting riders to match with idle drivers very far away. The failure gives rise to severe supply wastes in devastating pickups, especially during demand

surge when the system is thirsty for supply resources. We proved that reducing matching radius can resolve the issue and the vehicle utilization sustains if the matching radius can be adjusted adaptively in a certain way. We then turned to the anisotropic markets and identified another type of matching failure, which is due to the indiscriminate matching of riders despite their spatial attributes but can be addressed using price or matching differentiation. A generalized fluid-based system model was developed to tackle the needs for implementable management strategies in a realistic large-scale market. An empirical study was further conducted to showcase the tractability of our fluid-based model as well as the effectiveness of spatial pricing in restoring the system throughput in anisotropic markets.

Market Regulations [Chapter 3, 6 & 7]

The regulator comes into the play standing by societal or systematic benefits, holding the responsibility of governing the negative system externalities imposed by ride-sourcing services. We found that although allowing higher pricing flexibility facilitates market clearing, the platform could set price more than necessary, which compromises the interest of riders and drivers potentially. To avoid the platform from taking unreasonably high markups, the government may need to regulate the pricing power of the platform. Besides, the traffic manager may need to deal with the congestion due to the extra mileage taken by ride-sourcing vehicles, especially the empty miles from the end of one delivery trip to the start of the next. In the dissertation, we proposed a general network equilibrium model to encapsulate the production of vacant miles, and customized a solution algorithm to efficiently estimate the system states and consequences upon realistic-sized networks. Analytical models were further prescribed to assist agencies in mitigating the induced congestion, by effectively allocating road spaces to curbside parking to reduce cruising miles and enhance the overall social benefits.

8.2 Future Research

This dissertation focuses on the empty miles of ride-sourcing services. However, addressing the empty miles (or empty time/vacancy, in general) of shared-use fleets is a common challenge across shared mobility services. Numerous questions, problems, and proposals remain to be investigated. Future research may be pursued through the following directions, which embody a blend of theoretical and practical efforts.

Behavioral Analytics and Mechanism Design

An effective information sharing between the platform and service providers could reduce the fleet vacancy, while improving the overall system efficiency. However, despite an unprecedented

amount of information produced, in terms of both quantity and quality, it has been less efficiently processed, analyzed, and shared partially due to the gap between ideal theories and noisy realities. It is a challenging but meaningful task to bring closer the empirical analytics of user behaviors with the designs of strategies and mechanisms.

Mobility System Integration

The transportation ecosystem is multimodal, involving driving, ride/car-pooling, transit, cycling, and even drone services shortly. Shared mobility services, as increasingly important components of the ecosystem, should be seen and managed as a beneficial supplement rather than in isolation. Same for addressing the empty miles. We could prescribe optimal service selection and configuration to improve multimodal connections, promote system integration, and strengthen the competitiveness and sustainability of the whole system.

Implications of Vehicle Automation

In the foreseeable future, the emerging connected and autonomous vehicle technology will further revolutionize urban and rural mobility and facilitate the shift from car ownership to sharing/subscription. The encounter of vehicle automation and shared mobility has the potential to spur service innovations and catalyze new business models that may fundamentally reshape our ways of transporting crowds and logistics. It is thus critical to understand the impacts and implications of these services and provide guidance on their development and deployment.

APPENDIX A

Appendices for Chapter 2

A.1 Nomenclature

Table A.1: Notation list of sets, variables, parameters, and functions

Notation	Description
Variables	
S	System supply of trips
S_e^∞	Supply function of an e-hailing system without restricting the matching radius
S_e^r	Supply function of an e-hailing system under a restricted matching radius
S_s	Supply function of a street-hailing system
w^c	Average matching time of requests
w^m	Average pickup time of requests
C	Service costs of riders
N^v	Density of idle vehicles
N^r	Density of waiting customers
Q	Trip flux
A	Matching block size
λ^c	Poisson arrival flux of requests
λ^t	Poisson arrival flux of idle drivers
N^c	Customers' tolerance on the request queue length
N^t	Drivers' tolerance on the driver queue length
π_i	Steady state probability of the queue being at state i
Ω	State space of the double-ended queue
ρ	Flux ratio of λ^t over λ^c
p_a	Percentage of abandonment
$\sigma_{w^c}^2$	Queuing time variance for riders
$\hat{\sigma}_{w^c}^2$	Continuous approximation of the queuing time variance for riders

Functions	
W_e	Riders' cost function for the e-hailing mode
W_s	Riders' cost function for the street-hailing mode
M	Bilateral meeting function
$d(i)$	Riders' average distance to the closest vehicle (normalized)
Parameters	
F	Average fare of requests
N	Density of all vehicles
t	Average delivery time of customer trips
w_0^c	Customers' tolerance for the expected waiting time on matching
w_0^t	Drivers' tolerance for the expected waiting time on matching
v	Average network running speed
λ_e^t	External vehicle flux excluding the cruising flux of idle drivers
ρ_e	Flux ratio of λ_e^t over λ^c

A.2 Proof of theorems in Section 2.4.3 - Effects of matching blocks (radii) on the supply curve

A.2.1 Proof for Theorem 2.1

Define a function $f(\lambda^t|\lambda^c)$ on λ^t as $f(\lambda^t|\lambda^c) = \lambda^t \cdot [w^m(\lambda^t|\lambda^c, A) + t]$. Then, suppose there are two potential demands $\lambda_l^c,\ \lambda_h^c$ with $\lambda_l^c < \lambda_h^c$. Let λ_i^t be the equilibrated driver flux respectively realized under λ_i^c, $i \in \{l, h\}$ and we thus have $f(\lambda_i^t|\lambda_i^c) = N$, $\forall i \in \{l, h\}$. The fact that w^m is decreasing on ρ and increasing on λ^c yields

$$f(\lambda_l^t|\lambda_h^c) > f(\lambda_l^t|\lambda_l^c) = f(\lambda_h^t|\lambda_h^c) \tag{A.1}$$

Given λ^c, the first-order derivative of f pertains to the following equation:

$$\frac{f'}{w^m} = 1 + \frac{t}{w^m} + \epsilon_{\lambda^t}^{w^m}$$

where $\epsilon_{\lambda^t}^{w^m}$ denotes the elasticity of w^m over λ^t. As previously mentioned, $|\epsilon_{\lambda^t}^{w^m}|$ can be extremely small under the block size favored by the platform. This leads to a positive f' on λ^t over the range of interest. Therefore, combining $f' > 0$ with Eq. (A.1) yields $\lambda_l^t > \lambda_h^t$.

Denote C_i as the average cost of riders under λ_i^c, $i \in \{l, h\}$. We next prove $C_h > C_l$ by separately showing that the same relationship holds for both $\{w^c\}$ and $\{w^m\}$. The proof for w^m is straightforward:

$$w^m(\lambda_h^t|\lambda_h^c) = \frac{N}{\lambda_h^t} - t > \frac{N}{\lambda_l^t} - t = w^m(\lambda_l^t|\lambda_l^c)$$

As for w^c, we firstly prove $\rho_h < \rho_l$ by contradiction. If $\rho_h > \rho_l$, then from the monotonicity of w^m on ρ and N^c, we have $w^m(\rho_h, \lambda_h^t) < w^m(\rho_l, \lambda_l^t)$. Obviously, it contradicts with the conclusion above, and suggests $\rho_h < \rho_l$. Note that Eq. (2.9) of function w^c can be simplified as follows,

$$w^c(\rho, \lambda^c) = w_0^c \cdot \left(1 - \frac{1}{Aw_0^c\lambda^c \cdot (1-\rho)}\right)$$

As $\lambda_h^c > \lambda_l^c$ and $\rho_h < \rho_l$, we obtain $w^c(\lambda_h^t|\lambda_h^c) > w^m(\lambda_l^t|\lambda_l^c)$. Combining with the counterpart from w^m, we prove $C_h > C_l$.

Given $C_h > C_l$ and $\lambda_h^t < \lambda_l^t$, we conclude that the supply curve is still backward bending in the domain of supply shortage.

A.2.2 Proof for Theorem 2.2

Suppose two systems under two different block sizes, respectively denoted as A_l and A_h with $A_l < A_h$, have realized the same vehicle flux λ^t. We thus have

$$w^m(\lambda_l^c|A_l) = w^m(\lambda_h^c|A_h)$$

where λ_i^c respectively denotes the potential demand in the system of A_i, $i \in \{l, h\}$.

We prove in contradiction by assuming that $\lambda_l^c < \lambda_h^c$, and it follows $\rho_l > \rho_h$. As $N_l^c = A_l\lambda^t w_0^c < A_h\lambda^t w_0^c = N_h^c$, based on the monotonicity of w^m on ρ and N^c, $w^m(\rho_h, N_h^c) > w^m(\rho_l, N_l^c)$ will always hold. Obviously, it contradicts with the above equality, suggesting $\lambda_l^c > \lambda_h^c$ and $\rho_l < \rho_h$. Again, applying these relations to the w^c function gives rise to $w^c(\rho_l, \lambda_l^c) > w^c(\rho_h, \lambda_h^c)$ and therefore $C_l > C_h$.

Since the cost is always greater in a smaller zone for the same realized vehicle flux, the theorem is proved.

A.2.3 Proof for Theorem 2.3

Suppose a ride-hailing system that currently operates under the condition of (λ_l^c, A_l) experiences a demand surge to λ_h^c. If the block size remains to be A_l, as already proven, the trip supply worsens while the cost of riders increases. Below we prove the existence of A_h with $A_h < A_l$ such that $C_h > C_l$ and $\lambda_l^t > \lambda_h^t$, where C_i and λ_i^t denote the equilibrated system state variables respectively realized under (λ_i^c, A_i), $i \in \{l, h\}$.

Firstly, we show that there exists a block size A_m that gives rise to a trip flux of λ_l^t under the request flux of λ_h^c. The fleet conservation then yields the following relationship for $w^m(\rho, N^c)$:

$$w^m\left(\lambda_l^t/\lambda_l^c,\ A_l\lambda_l^t \cdot w_0^c\right) = w^m\left(\lambda_l^t/\lambda_h^c,\ A_m\lambda_l^t \cdot w_0^c\right) \tag{A.2}$$

If we specify A_m to be the value of A_l, the right-hand-side of Eq. (A.2) will be greater than the left. And if A_m is valued 0, the right will be equal to zero. As w^m is a continuous function on A_m, there exists an A_m less than A_l that satisfies the condition. By virtue of Theorem 2.2, the per-rider's cost C_m realized under (λ_h^c, A_m) will be greater than C_l, which is achieved under the condition of (λ_l^c, A_l).

Secondly, we prove that given the request flux λ^c, the equilibrated vehicle flux λ^t increases by shrinking the block size. Taking the total derivative of Eq. (2.11a) on λ^t and A yields,

$$\frac{\mathrm{d}\lambda^t}{\mathrm{d}A} = \frac{\frac{\lambda^t}{w^m} \cdot \frac{\partial w^m}{\partial A}}{1 + \frac{t}{w^m} + \epsilon_{\lambda^t}^{w^m}}$$

Since the elasticity $\epsilon_{\lambda^t}^{w^m}$ features a small absolute value and $\frac{\partial w^m}{\partial A} > 0$, we thus obtain that $\frac{\mathrm{d}\lambda^t}{\mathrm{d}A} > 0$.

Shrink the block size from A_m by δA, and specify C_h and λ_h^t as the equilibrated system state variables realized under $(\lambda_h^c, A_m - \delta A)$. From $\frac{\mathrm{d}\lambda^t}{\mathrm{d}A} > 0$, we have $\lambda_h^t > \lambda_l^t$. Also, by specifying δA with a small value, we can always retrieve C_h such that $C_m - C_h$ is less than $C_m - C_l$. Since $A_l > A_m > A_m - \delta A$, we prove the existence of A_h.

A.3 Proof of Theorem 2.4 in Section 2.4.4 - The monotonicity of riders' costs on N^t

The average queuing time w^c and pickup time w^m of riders characterize the following functions on (ρ, N^c):

$$w^c(\rho, N^c) = \frac{1}{\rho_e A \lambda^t} \cdot \sum_{i=1}^{N^c} i \cdot \pi_{1-i} = \frac{1-\rho_e}{\rho_e A \lambda^c} \cdot \sum_{i=1}^{N^c} i \cdot \rho^{N^c - i} \tag{A.3a}$$

$$w^m(\rho, N^c) = \frac{\sqrt{A}}{v\rho_e} \cdot \sum_{i=-N^c+1}^{N^t} d^a(i) \cdot \pi_i \tag{A.3b}$$

where $N^c = A w_0^c \lambda^c \rho$ and $d^a(i)$ features a deceasing function converted from $d(i)$, i.e.

$$d^a(i) = \begin{cases} d(1), & i \in \{-N^c+1, -N^c+2, \cdots, 0\} \\ d(i), & i \in \{1, 2, \cdots, N^c\} \end{cases}$$

We firstly prove the monotonicity of ρ over N^t:

Proposition A.1. *Flux ratio ρ decreases monotonically on drivers' queuing tolerance N^t.*

Proof. The variables ρ and N^t satisfy an equation below:

$$1 - \rho_e = \frac{1-\rho}{1-\rho^{N^t+N^c+1}}$$

Define a function $f(\rho, N)$ as $f(\rho, N) = \frac{1-\rho}{1-\rho^N}$. It can be easily verified that both $\frac{\partial f}{\partial \rho}$ and $\frac{\partial f}{\partial N}$ are negative when $N > 1$. Then, taking a total derivative of $1 - \rho_e = f(\rho, N^t + N^c + 1)$ to N^t yields

$$0 = \frac{\partial f}{\partial \rho} \cdot \rho' + \frac{\partial f}{\partial N} \cdot (A w_0^c \lambda^c \cdot \rho' + 1)$$

which further gives rise to $\rho' < 0$. □

Suppose drivers' tolerance over the queue length increases from N_l^t to N_h^t. Denote ρ_i, N_i^c respectively as the flux ratio and the capacity of request queue realized under N_i^t, $i \in \{l, h\}$. Then, based on the monotonicity given by Proposition A.1, we have $\rho_l > \rho_h$ and $N_l^c > N_h^c$.

Proposition A.2. *Riders' average queuing time monotonically decreases on N^t.*

Proof. We prove $w^c(\rho_l, N_l^c) > w^c(\rho_h, N_h^c)$ by taking ρ and N^c as two independent variables. Firstly, based on the formula (A.3a), it is clear that $w^c(\rho_l, N_l^c) > w^c(\rho_h, N_l^c)$. And the inequality

$\sum_{i=1}^{N_l^c} i \cdot \rho^{N_l^c - i} > \sum_{i=N_l^c - N_h^c + 1}^{N_l^c} i \cdot \rho^{N_l^c - i} > \sum_{i=1}^{N_h^c} i \cdot \rho^{N_h^c - i}$ further gives rise to $w^c(\rho_h, N_l^c) > w^c(\rho_h, N_h^c)$. Combining both stages yields $w^c(\rho_l, N_l^c) > w^c(\rho_h, N_h^c)$. □

Proposition A.3. *Riders' average pickup time monotonically decreases on* N^t.

Proof. We prove the proposition by respectively comparing $w^m(\rho_l, N_l^c)$ and $w^m(\rho_h, N_h^c)$ under two complementary conditions. When $N_l^c + N_l^t \geqslant N_h^c + N_h^t$,

$$\begin{aligned} w^m(\rho_l, N_l^c) &= \frac{\sqrt{A}(1-\rho_e)}{v\rho_e} \cdot \sum_{i=-N_l^c+1}^{N_l^t} d^a(i) \cdot \rho_l^{N_l^c+i} \\ &> \frac{\sqrt{A}(1-\rho_e)}{v\rho_e} \cdot \sum_{i=-N_l^c+1}^{N_h^t-(N_l^c-N_h^c)} d^a(i) \cdot \rho_h^{N_l^c+i} \\ &\geqslant \frac{\sqrt{A}(1-\rho_e)}{v\rho_e} \cdot \sum_{i=-N_h^c+1}^{N_h^t} d^a(i) \cdot \rho_h^{N_h^c+i} \\ &= w^m(\rho_h, N_h^c) \end{aligned}$$

and when $N_l^c + N_l^t < N_h^c + N_h^t$,

$$\begin{aligned} w^m(\rho_l, N_l^c) &= \frac{\sqrt{A}(1-\rho_e)}{v\rho_e} \cdot \sum_{i=-N_l^c+1}^{N_l^t} d^a(i) \cdot \rho_l^{N_l^c+i} \\ &> \frac{\sqrt{A}(1-\rho_e)}{v\rho_e} \cdot \left[\sum_{i=-N_h^c+1}^{N_l^t+N_l^c-N_h^c} d^a(i) \cdot \rho_h^{N_h^c+i} + d^a\left(N_l^t + N_l^c - N_h^c\right) \cdot \left(\rho_e - \sum_{i=-N_h^c+1}^{N_l^t+N_l^c-N_h^c} \rho_h^{N_h^c+i} \right) \right] \\ &\geqslant \frac{\sqrt{A}(1-\rho_e)}{v\rho_e} \cdot \sum_{i=-N_h^c+1}^{N_h^t} d^a(i) \cdot \rho_h^{N_h^c+i} \\ &= w^m(\rho_h, N_h^c) \end{aligned}$$

□

Combining Proposition A.2 and A.3 completes the proof that both w^c and w^m monotonically decrease on N^t.

APPENDIX B

Appendices for Chapter 3

B.1 Nomenclature

Table B.1: Notation list of sets, variables, parameters, and functions

Notation	Description
Sets	
$\mathcal{N}$	Set of regions
$\mathcal{T}$	Set of time period
$\mathcal{W}$	Set of OD pairs of ride-sourcing customer demands
$\mathcal{W}^c$	Set of bordered pairs of regions
Variables	
$a_{ij}^c(t)$	Cumulative number of type-(i,j) riders arriving at region i by time t
$m_{ij}(t)$	Cumulative number of type-(i,j) riders being matched at region i by time t
$\pi_{ij}^c m_{ij}(t,s)$	Number of type-(i,j) riders who enter at interval t while matched by time s
$q_{ij}^c(t)$	Number of type-(i,j) riders queueing at region i at time t
$w_{ij}^q(t)$	Time needed to match the number of type-(i,j) riders queueing at time t
$d_{ij}(t)$	Cumulative number of type-(i,j) riders delivered at destination j by time t
$w_{ij}^s(t)$	Service time for type-(i,j) riders who receive matches at time t
$a_i^d(t)$	Total arrival of idle drivers at region i by time t
$c_{ij}(t)$	Cumulative number of idle drivers that cruise from region i to k
$q_i^d(t)$	Number of idle drivers at region i at time t
$b_i(t)$	Cumulative number of drivers abandoning the system at region i by time t
$w_{ij}^m(t)$	Matching time for type-(i,j) riders who arrive at time t
$w_{ij}^p(t)$	Pickup time for type-(i,j) riders matched at time t
$\mathcal{P}_{ij}(t)$	Pricing strategy on type-(i,j) riders at time t
$\boldsymbol{\mathcal{A}}_i(t)$	Matching mechanisms at region i at time t
$\boldsymbol{\mathcal{R}}_i(t)$	Service interventions on idle drivers at region i at time t

Parameters	
$e_i(t)$	Cumulative external arrivals of idle drivers at region i by time t
$\alpha_i(t), \gamma_i(t)$	Coefficients of the pickup time function at region i in period t
$\eta^s_{ij}(t)$	Discounting ratio of ride-pooling for type-i, j riders in period t
$B_{ij}(t)$	Frequency of type-(i, j) users accessing the application in period t
$\boldsymbol{\beta}_{ij}(t)$	Type-(i, j) users' sensitivities on costs in period i
q^0_i	Threshold on the number of idle drivers in matching at region i
$\Lambda^c_{ij}(t)$	Probability of a driver repositioning from region i to j in period t
$\Lambda^b_i(t)$	Probability of a driver abandoning service at region i in period t
Functions	
$D_{ij}(t)$	Demand function of type-(i, j) riders at time t
$M_{ij}(\cdot)$	Aggregate matching function of type-(i, j) riders
$T_{ij}(t)$	Function on type-(i, j) riders' pickup time at time t
$C_{ij}(t)$	Function on drivers' rate of transition from region i to j at time t
$B_i(t)$	Function on drivers' rate of service abandonment at region i time t

APPENDIX C

Appendices for Chapter 6

C.1 Nomenclature

Table C.1: Notation list of sets, variables, parameters and functions

Notation	Description
Sets	
V	Set of nodes
A	Set of links
R	Set of the origin nodes of customer trips; $R \subseteq V$
S	Set of the destination nodes of customer trips; $S \subseteq V$
W	Set of OD pairs of ride-sourcing customer demands
W^b	Set of OD pairs of background regular traffic
W^c	Complete set of OD pairs, including those of RVs and regular vehicles
$M^c(r)$	Set of nodes hailing customers at node $r \in R$ can potentially be matched to
L	Set of nodes with positive accumulations of idle RVs
$M^v(l)$	Set of nodes idle RVs at node $l \in L$ can potentially be matched to
Variables	
C_{rs}	Monetary travel cost between OD pair $(r, s) \in W$
F_{rs}	Fare of a trip from node r to s
w_r^c	Customer's average waiting time at node r, $r \in R$
h_{rs}	Equilibrium or shortest vehicular travel time between node r and s
Q_{rs}	Customer demand between OD pair $(r, s) \in W$
$U_{sr}(U_{sl})$	Idle RVs' utility of cruising from node $s \in S$ to $r \in R(l \in L)$
$w_r^v(w_l^v)$	Idle RVs' average matching time at node $r \in R(l \in L)$
$\bar{F}_r$	Average fare of customer trips originating from node $r \in R$
$\hat{F}_l$	Average earnings of RVs who get matched at node $l \in L$
$\bar{h}_r$	Average service time of customer trips originating from node $r \in R$
$\hat{h}_l$	Average service time of RVs who get matched at node $l \in L$

T^o_{rs}	Occupied RV flow that serves customer demand from node r to s, $(r,s) \in W$
$T^v_{sr}(T^v_{sl})$	Idle RV flow from node $s \in S$ to $r \in R(l \in L)$
T^n_{sr}	Regular traffic flow from node r to s, $(r,s) \in W^b$
$N^v_r(N^v_l)$	Number of idle RVs at node $r \in R(l \in L)$
N^c_r	Number of hailing customers at node $r \in R$
T^m_{lr}	Rate of RVs matched from node $l \in L$ to $r \in M^v(l)$
Parameters	
β^o	Customer's out-of-vehicle value of time ($/h)
β^i	Customer's in-vehicle value of time ($/h)
γ	RVs' value of time $\$/h$
θ	Degree of RV drivers' perception error dispersion
N	Total number of RVs in the network
Functions	
f_{rs}	Function of customer demand versus travel costs for OD pair $(r,s) \in W$
g_{rs}	Function of trip fare versus travel time for OD pair $(r,s) \in W$
m_r	Aggregate matching function for node $r \in R$
$\Phi^v(\Phi^c)$	Node potential function on accumulations of idle RVs(hailing customers)
Δ	Potential difference function for paired nodes with positive matching flows

C.2 Existence and uniqueness of solutions for the inter-node matching equations

This appendix proves that given $\{N_l^v\}$, $\{N_r^c\}$ and $\{h_{lr}\}$ to be fixed and positively valued, there exists a unique matching rate pattern $\{T_{lr}^m\}$ that satisfies the following aggregate matching equation set (C.1),

$$\Phi^v\left(\sum_{k\in M^v(l)} T_{lk}^m, N_l^v\right)\cdot\Phi^c\left(\sum_{k\in M^c(r)} T_{kr}^m, N_r^c\right)=\Delta(T_{lr}^m,h_{lr})\qquad \forall l\in L, r\in M^v(l)\qquad \text{(C.1)}$$

For mathematical convenience, we use $\{\Phi_l^v\}_{l\in L}$, $\{\Phi_r^c\}_{r\in R}$ and $\{\Delta_{lr}\}_{l\in L,r\in M^v(l)}$ to indicate the functions of Φ^v, Φ^c and Δ conditional on $\{N_l^v\}$, $\{N_r^c\}$ and $\{h_{lr}\}$, respectively. Generic properties that pertain to $\{\Phi_l^v(T_l^v),\ \{\Phi_r^c(T_r^c)\}$ (abbrev. $\Phi(T)$) and $\{\Delta_{lr}(T_{lr}^m)\}$ (abbrev. $\Delta(T)$) are summarized as the following conditions C1-C3:

C1. $\Phi(T)$ and $\Delta(T)$ are continuous functions defined on $T\in(0,+\infty)$.

C2. For any $T\in(0,+\infty)$, we have $\Phi(T)>0,\ \Phi'(T)<0$ and $\Delta(T)>0,\ \Delta'(T)>0$.

C3. There exist $p>0$ and $q>0$ such that

$$\Phi(T)=\begin{cases}\Theta(T^{-p}), \text{ as } T\to 0^+\\ \Theta(T^{-q}), \text{ as } T\to+\infty\end{cases}$$

These properties serve as the foundation for the proof below.

In the first stage, we prove the existence of a feasible solution $\{T_{lr}^m\}$ to Eq. (C.1). The proof begins with the Proposition C.1 stated below.

Proposition C.1. *There exist positive constants* $(\omega_o,\ \omega_u,\ \tau_o,\ \tau_u)$ *such that*

$$(\omega_o T^p+\omega_u T^q)^{-1}\leqslant\Phi(T)\leqslant\tau_o T^{-p}+\tau_u T^{-q}$$

Proof. Condition C3 suggests there exist $T_o>0,\omega_o'>0,\tau_o'>0$ and $T_u>0,\omega_u'>0,\tau_u'>0$ such that

$$\begin{aligned}&\omega_o' T^{-p}\leqslant\Phi(T)\leqslant\tau_o' T^{-p} && T\in(0,T_o]\\ &\omega_u' T^{-q}\leqslant\Phi(T)\leqslant\tau_u' T^{-q} && T\in[T_u,+\infty)\end{aligned}$$

Collectively, the above two inequalities yield

$$(\omega_o'^{-1}T^p + \omega_u'^{-1}T^q)^{-1} \leqslant \Phi(T) \leqslant \tau_o' T^{-p} + \tau_u' T^{-q} \qquad T \in (0, T_o] \cup [T_u, +\infty) \tag{C.2}$$

If $T_o \geqslant T_u$, the proposition is proven. Otherwise, as Φ is monotonically decreasing, we have that for any $T \in [T_o, T_u]$,

$$\Phi(T) \geqslant \Phi(T_o) \cdot \frac{\Phi(T_u)}{\Phi(T_o)} \geqslant (\omega_o'^{-1}T^p + \omega_u'^{-1}T^q)^{-1} \cdot \frac{\Phi(T_u)}{\Phi(T_o)} \tag{C.3a}$$

$$\Phi(T) \leqslant \Phi(T_u) \cdot \frac{\Phi(T_o)}{\Phi(T_u)} \leqslant (\tau_o' T^{-p} + \tau_u' T^{-q}) \cdot \frac{\Phi(T_o)}{\Phi(T_u)} \tag{C.3b}$$

where the first inequality results from the monotonicity of Φ and the second one is due to the above relation (C.2). Since $\Phi(T_o)$ and $\Phi(T_u)$ characterize two constants with $\Phi(T_o) \geqslant \Phi(T_u)$, the inequalities (C.3) are extensible for the entire domain of $T \in (0, +\infty)$, which essentially indicates Proposition (C.1). □

Define $Z_{lr}(T)$ as $T^{2p+2q}\Delta_{lr}(T)$. Then, each Z_{lr} characterizes a monotonically increasing function on T. By multiplying both sides of Eq. (C.1) with T^{2p+2q}, we can base on that equation to construct the following mapping Γ that maps $\boldsymbol{T^i}(=\{T_{lr}^i\})$ to $\boldsymbol{T^d}(=\{T_{lr}^d\})$, i.e.

$$\Gamma : \left\{ T_{lr}^d = Z_{lr}^{-1}\left({T_{lr}^i}^{2p+2q} \cdot \Phi_l^v\left(\sum_{k \in M^v(l)} T_{lk}^i \right) \cdot \Phi_r^c\left(\sum_{k \in M^c(r)} T_{kr}^i \right) \right),\ \forall l \in L, r \in M^v(l) \right\} \tag{C.4}$$

Then, the existence of a solution to Eq. (C.1) is equivalent to proving that the mapping (C.4) has a fixed point. As per Brouwer's fixed point theorem (De la Fuente, 2000), this requires the existence of a compact convex set Ω, such that $\boldsymbol{T^d} = \Gamma(\boldsymbol{T^i}) \in \Omega$ for any $\boldsymbol{T^i} \in \Omega$. We thus prove the second proposition as follows,

Proposition C.2. *A pair of positive bounds B_o and $B_u(\geqslant B_o)$ can always be found, with $\Omega = \{\boldsymbol{T} | T_{lr} \in [B_o, B_u], \forall l \in L, r \in M^v(l)\}$ satisfying the above condition required by the existence of fixed points.*

Proof. According to Proposition (C.1), there will be a set of constants $(\omega_{k,o}^t, \omega_{k,u}^t, \tau_{k,o}^t, \tau_{k,u}^t)$ for each Φ_k^t such that

$$(\omega_{k,o}^t T^p + \omega_{k,u}^t T^q)^{-1} \leqslant \Phi_k^t(T) \leqslant \tau_{k,o}^t T^{-p} + \tau_{k,u}^t T^{-q}$$

where $(t, k) \in S = \{k \in L \text{ if } t = v \text{ and } k \in R \text{ if } t = c\}$. By defining $\omega_o = \max\{\omega_{k,o}^t\}, \omega_u =$

$\max\{\omega_{k,u}^t\}$ and $\tau_o = \max\{\tau_{k,o}^t\}, \tau_u = \max\{\tau_{k,u}^t\}$ over all $(t,k) \in S$, we thus have

$$(\omega_o T^p + \omega_u T^q)^{-1} \leqslant \Phi_k^t(T) \leqslant \tau_o T^{-p} + \tau_u T^{-q} \qquad \forall (t,k) \in S$$

Next, we define two univariate functions $\Delta_o(T)$ and $\Delta_u(T)$ based on $\{\Delta_{lr}(T)\}$, i.e.,

$$\Delta_o(T) = \min\{\Delta_{lr}(T),\ \forall l \in L, r \in M^v(l)\}$$
$$\Delta_u(T) = \max\{\Delta_{lr}(T),\ \forall l \in L, r \in M^v(l)\}$$

Also, define $Z_i(T) = T^{2p+2q}\Delta_i(T)$ for either $i \in \{o,u\}$. As all functions in $\{\Delta_{lr}(T)\}$ increase monotonically on T, the resultant Δ_o, Δ_u and Z_o, Z_u are all monotonically increasing functions. Suppose T_{lr} will be valued from a given set $[B_o, B_u], \forall l \in L, r \in M^v(l)$. Then,

$$\begin{aligned}
\Gamma_{lr}(\boldsymbol{T}) &\leqslant Z_o^{-1}\left(T_{lr}^{2p+2q} \cdot \Phi_l^v\left(\sum_{k\in M^v(l)} T_{lk}\right) \cdot \Phi_r^c\left(\sum_{k\in M^c(r)} T_{kr}\right)\right) \\
&\leqslant Z_o^{-1}\left(T_{lr}^{2p+2q} \cdot \Phi_l^v\left(T_{lr}\right) \cdot \Phi_r^c\left(T_{lr}\right)\right) \\
&\leqslant Z_o^{-1}\left(T_{lr}^{2p+2q} \cdot \left(\tau_o \cdot T_{lr}^{-p} + \tau_u \cdot T_{lr}^{-q}\right)^2\right) \\
&\leqslant Z_o^{-1}\left((\tau_o B_u^q + \tau_u B_u^p)^2\right)
\end{aligned}$$

Let $\xi(B)$ denote $(\tau_o B^q + \tau_u B^p)^2$. The upper bound B_u can thus be defined as solving $Z_o(B_u) = \xi(B_u)$. Note that by solving the equation, we can always receive a feasible B_u because

$$Z_o(B_u) = \Theta\left(\Delta_o(B_u) \cdot B_u^{2p+2q}\right) \underset{(<)}{>} \Theta(B_u^{2p+2q}) \underset{(<)}{>} \Theta\left((\tau_o B_u^q + \tau_u B_u^p)^2\right) \qquad \text{as } B_u \to \underset{(0^+)}{+\infty}$$

given $p > 0, q > 0$ and $\Delta_o(\cdot)$ increases monotonically ranging from 0^+ to $+\infty$.

On the other hand, by defining constants $n_v = \max_{l\in L}\{|M^v(l)|\}$ and $n_c = \max_{r\in R}\{|M^c(r)|\}$,

$$\begin{aligned}
\Gamma_{lr}(\boldsymbol{T}) &\geqslant Z_u^{-1}\left(T_{lr}^{2p+2q} \cdot \Phi_l^v\left(\sum_{k\in M^v(l)} T_{lk}\right) \cdot \Phi_r^c\left(\sum_{k\in M^c(r)} T_{kr}\right)\right) \\
&\geqslant Z_u^{-1}\left(T_{lr}^{2p+2q} \cdot \Phi_l^v\left(T_{lr} + (n_v - 1)B_u\right) \cdot \Phi_r^c\left(T_{lr} + (n_c - 1)B_u\right)\right) \\
&\textit{Definition: Let } B_u^v = (n_v - 1)B_u \textit{ and } B_u^c = (n_c - 1)B_u \\
&\geqslant Z_u^{-1}\left(\frac{T_{lr}^{p+q}}{\omega_o \cdot (T_{lr} + B_u^v)^p + \omega_u \cdot (T_{lr} + B_u^v)^q} \cdot \frac{T_{lr}^{p+q}}{\omega_o \cdot (T_{lr} + B_u^c)^p + \omega_u \cdot (T_{lr} + B_u^c)^q}\right) \\
&\geqslant Z_u^{-1}\left(\frac{B_o^{p+q}}{\omega_o \cdot (B_o + B_u^v)^p + \omega_u \cdot (B_o + B_u^v)^q} \cdot \frac{B_o^{p+q}}{\omega_o \cdot (B_o + B_u^c)^p + \omega_u \cdot (B_o + B_u^c)^q}\right)
\end{aligned}$$

Define the above term on B_o within the parenthesis behind Z_u^{-1} as $\zeta(B_o)$. Then, the lower bound B_o can be defined as solving $Z_u(B_o) = \zeta(B_o)$. Again, such a B_o always exists because

$$Z_u(B_o) = \Theta\left(\Delta_u(B_o) \cdot B_o^{2p+2q}\right) \underset{(<)}{>} \Theta\left(B_o^{2p+2q}\right) \underset{(\leqslant)}{>} \Theta\left(\zeta(B_o)\right) \qquad \text{as } B_u \underset{(0^+)}{\to} +\infty$$

Further, since $Z_u(B_u) \geqslant Z_o(B_u) = \xi(B_u) \geqslant \zeta(B_u)$, there must exist a B_o in $(0, B_u]$.

It is then straightforward that with any pair of B_o and B_u defined above and $\boldsymbol{T} = \{T_{lr} | T_{lr} \in [B_o, B_u], \forall l \in L, r \in M^v(l)\}$, all the mappings $\{\Gamma_{lr}(\boldsymbol{T})\}$ will compliantly fall in $[B_o, B_u]$. □

With Proposition (C.2) held, the existence of a feasible solution $\{T_{lr}^m\}$ to the equation set (C.1) is always guaranteed. In the second stage, we prove that the ensured solution $\{T_{lr}^m\}$ is unique.

Proposition C.3. *The matching flow pattern $\{T_{lr}^m\}$ that solves the following Eq. (C.6) is unique.*

$$\Phi_l^v\left(T_l^v\right) \cdot \Phi_r^c\left(T_r^c\right) = \Delta_{lr}(T_{lr}^m) \qquad \forall l \in L, r \in M^v(l) \tag{C.6a}$$

$$T_l^v = \sum_{k \in M^v(l)} T_{lk}^m \qquad l \in L \tag{C.6b}$$

$$T_r^c = \sum_{k \in M^c(r)} T_{kr}^m \qquad r \in R \tag{C.6c}$$

Proof. We prove the proposition by contradiction.

Assume there are two distinct flow patterns $\{T_{lr}^{m,1}\}$ and $\{T_{lr}^{m,2}\}$ that solve Eq. (C.6). Firstly, if $T_l^{v,1} = T_l^{v,2}$ and $T_r^{c,1} = T_r^{c,2}$ for all $l \in L$ and $r \in R$, then Eq. (C.6a) suggests the equivalence of $\{T_{lr}^1\}$ and $\{T_{lr}^2\}$. Suppose $T_{l_1}^{v,1} > T_{l_1}^{v,2}$ for $l_1 \in L$, and then $T_{l_1 r_1}^{m,1} > T_{l_1 r_1}^{m,2}$ for $r_1 \in M^v(l_1)$. These two inequalities yield

$$\Phi_{r_1}^c(T_{r_1}^{c,1}) = \frac{\Delta_{l_1 r_1}(T_{l_1 r_1}^{m,1})}{\Phi_{l_1}^v(T_{l_1}^{v,1})} > \frac{\Delta_{l_1 r_1}(T_{l_1 r_1}^{m,2})}{\Phi_{l_1}^v(T_{l_1}^{v,2})} = \Phi_{r_1}^c(T_{r_1}^{c,2})$$

which further gives rise to $T_{r_1}^{c,1} < T_{r_1}^{c,2}$. Thus, there exists an $l_2 \in M^c(r_1)$ such that $T_{l_2 r_1}^{m,1} < T_{l_2 r_1}^{m,2}$. Through a simple conduction, we obtain the following connection between $\Phi_{l_2}^v$ and $\Phi_{l_1}^v$,

$$\frac{\Phi_{l_2}^v(T_{l_2}^{v,1})}{\Phi_{l_2}^v(T_{l_2}^{v,2})} = \frac{\Delta_{l_2 r_1}(T_{l_2 r_1}^{m,1}) / \Delta_{l_1 r_1}(T_{l_1 r_1}^{m,1})}{\Delta_{l_2 r_1}(T_{l_2 r_1}^{m,2}) / \Delta_{l_1 r_1}(T_{l_1 r_1}^{m,2})} \cdot \frac{\Phi_{l_1}^v(T_{l_1}^{v,1})}{\Phi_{l_1}^v(T_{l_1}^{v,2})} < \frac{\Phi_{l_1}^v(T_{l_1}^{v,1})}{\Phi_{l_1}^v(T_{l_1}^{v,2})} < 1 \tag{C.7}$$

The inequality (C.7) implies two folds of relationships: 1. $l_2 \neq l_1$; 2. $T_{l_2}^{v,1} > T_{l_2}^{v,2}$, which together evidences the existence of an $r_2 \in R$ with $T_{l_2 r_2}^{m,1} > T_{l_2 r_2}^{m,2}$ and

$$\frac{\Phi_{r_2}^c(T_{r_2}^{c,1})}{\Phi_{r_2}^c(T_{r_2}^{c,2})} = \frac{\Delta_{l_2 r_2}(T_{l_2 r_2}^{m,1}) / \Delta_{l_2 r_1}(T_{l_2 r_1}^{m,1})}{\Delta_{l_2 r_2}(T_{l_2 r_2}^{m,2}) / \Delta_{l_2 r_1}(T_{l_2 r_1}^{m,2})} \cdot \frac{\Phi_{r_1}^c(T_{r_1}^{c,1})}{\Phi_{r_1}^c(T_{r_1}^{c,2})} > \frac{\Phi_{r_1}^c(T_{r_1}^{c,1})}{\Phi_{r_1}^c(T_{r_1}^{c,2})} > 1$$

Therefore, by repeating the above process iteratively, we can retrieve two ceaselessly nonrepetitive sequences $\{l_3, l_4, l_5, \cdots\} \subseteq L$ and $\{r_3, r_4, r_5, \cdots\} \subseteq R$. This yields an obvious contradiction because both L and R are finite sets. □

Integrating Proposition (C.2) and (C.3) proves the existence and uniqueness of a solution for Eq. (C.1). We present this conclusion formally as the following Theorem (C.1):

Theorem C.1. *By fixing $\{N_l^v\}$, $\{N_r^c\}$ and $\{h_{lr}\}$ with positive values, the aggregate matching equation set (C.1) yields a unique flow pattern $\{T_{lr}^m\}_{l\in L, r\in M^v(l)}$.*

C.3 Existence of equilibrium for the inter-node matching system

This appendix proves that for any practical parametric setting, there always exists an equilibrium solution to the inter-node matching system (6.12). Before starting the proof, we first bring back the mapping Γ defined in Appendix C.2 and redefine it as a mapping from $\{\boldsymbol{T^m}, \boldsymbol{N^v}, \boldsymbol{N^c}, \boldsymbol{h}\}$ to $\{\hat{\boldsymbol{T}}^{\boldsymbol{m}}\}$, i.e.,

$$\Gamma : \left\{ \hat{T}^m_{lr} = Z^{-1}_{lr} \left(T^{m\,2p+2q}_{lr} \cdot \Phi^v_l \left(\sum_{k \in M^v(l)} T^m_{lk}, N^v_l \right) \cdot \Phi^c_r \left(\sum_{k \in M^c(r)} T^m_{kr}, N^c_r \right), h_{lr} \right) \right\}_{l \in L, r \in M^v(l)}$$

where the variables $\boldsymbol{N^v}$, $\boldsymbol{N^c}$ and $\boldsymbol{h}$ previously specified as fixed parameters are now endogenously incorporated. We then leverage Γ to create a mapping $\mathcal{F} : \{\boldsymbol{T^m}, \boldsymbol{N^v}, \boldsymbol{N^c}, \boldsymbol{h}\} \to \{\hat{\boldsymbol{T}}^{\boldsymbol{m}}, \hat{\boldsymbol{N}}^{\boldsymbol{v}}, \hat{\boldsymbol{N}}^{\boldsymbol{c}}, \hat{\boldsymbol{h}}\}$ by reconstructing the solution procedures in Section 6.4.4:

$\mathcal{F}$: With $\{\boldsymbol{T^m}, \boldsymbol{N^c}, \boldsymbol{N^v}, \boldsymbol{h}\}$,

1. Obtain $\hat{\boldsymbol{T}}^{\boldsymbol{m}}$ by applying the mapping Γ above.

2. Deduce $\tilde{\boldsymbol{w}}^{\boldsymbol{c}}$ by solving the following equation on each source node $r \in R$:

$$\sum_{s:(r,s)\in W} f_{rs} \left(\beta^o \tilde{w}^c_r + \beta^m \check{h}_r + (\beta^i + \beta^f) h_{rs} \right) = \sum_{l \in M^c(r)} \hat{T}^m_{lr} \qquad \forall r \in R \tag{B1}$$

   Then, retrieve $\{\hat{\boldsymbol{T}}^{\boldsymbol{o}}, \hat{\boldsymbol{w}}^{\boldsymbol{c}}, \hat{\boldsymbol{N}}^{\boldsymbol{c}}\}$ by letting $\hat{T}^o_{rs} = f_{rs} \left(\beta^o \tilde{w}^c_r + \beta^m \check{h}_r + (\beta^i + \beta^f) h_{rs} \right)$ for all $(r, s) \in W$ and $\hat{w}^c_r = \max\{0, \tilde{w}^c_r\}$, $\hat{N}^c_r = \hat{w}^c_r \cdot \sum_{s:(r,s)\in W} \hat{T}^o_{rs}$ for all $r \in R$.

3. Obtain $\{\hat{\boldsymbol{T}}^{\boldsymbol{v}}, \tilde{\boldsymbol{w}}^{\boldsymbol{v}}\}$ by solving IRVM, then update $\hat{w}^v_l = \max\{0, \tilde{w}^v_l\}$ and $\hat{N}^v_l = \min\{N, \hat{w}^v_l \cdot \sum_{s \in S} \hat{T}^v_{sl}\}$ for all $l \in L$.

4. Obtain $\{\hat{\boldsymbol{h}}\}$ by solving PE.

To accommodate the zero-flow movements that may appear in practice, we expand the domain of $\mathcal{F}$ to cover zero flows by prescribing $0 \cdot \log 0 = 0$, $\Phi^v_l(\boldsymbol{T^m}, 0) = \Phi^c_r(\boldsymbol{T^m}, 0) = 0$, and $f_{rs}(-\infty) = 0$ for all $l \in L$, $r \in R$ and $(r, s) \in W$. Again, we apply Brouwer's fixed point theorem to show that the self-mapping $\mathcal{F}$ holds for at least one fixed point. Below first proves the continuity of mapping $\mathcal{F}$,

Proposition C.4. *The mapping $\mathcal{F}$ is continuous on $\{\boldsymbol{T^m}, \boldsymbol{N^v}, \boldsymbol{N^c}, \boldsymbol{h}\} \in \Omega$, given that*

$$\Omega = \left\{ \boldsymbol{T^m}, \boldsymbol{N^v}, \boldsymbol{N^c}, \boldsymbol{h} \;\middle|\; \begin{array}{rcll} T^m_{lr} & \in & [0, \bar{T}^m], & \forall (l,r) \in \{(l,r) | l \in L, r \in M^v(l)\} \\ N^c_r & \in & [0, \bar{N}^c], & \forall r \in R \\ N^v_l & \in & [0, \bar{N}^v], & \forall l \in L \\ h_{ij} & \in & [h^o, h^u], & \forall (i,j) \in A \end{array} \right\}$$

where $\bar{T}^m, \bar{N}^v, \bar{N}^c, h^o$ and h^u are all constants.

Proof. The proof of continuity follows the same sequence as the mapping steps above.

For Step 1, the mapping Γ characterizes a continuous function. Thus, it is clear that the resultant $\hat{\boldsymbol{T}}^{\boldsymbol{m}}$ is continuous on $\{\boldsymbol{T^m}, \boldsymbol{N^v}, \boldsymbol{N^c}, \boldsymbol{h}\}$.

Step 2 obtains $\tilde{\boldsymbol{w}}^{\boldsymbol{c}}$ from solving equations on $\{\hat{\boldsymbol{T}}^{\boldsymbol{m}}, \boldsymbol{T^m}, \boldsymbol{h}\}$. As the superposition of strictly decreasing functions preserves the strict monotonicity, the left-hand side (LHS) of Eq. (B1) is invertible. Thus, we can easily transform $\tilde{\boldsymbol{w}}^{\boldsymbol{c}}$ into continuous functions on $\{\hat{\boldsymbol{T}}^{\boldsymbol{m}}, \boldsymbol{T^m}, \boldsymbol{h}\}$. Since the rest mappings given $\tilde{\boldsymbol{w}}^{\boldsymbol{c}}$ are explicit continuous functions, we thus have $\hat{\boldsymbol{w}}^{\boldsymbol{c}}$, $\hat{\boldsymbol{T}}^{\boldsymbol{o}}$ as well as $\hat{\boldsymbol{N}}^{\boldsymbol{c}}$ be continuous on $\{\boldsymbol{T^m}, \boldsymbol{N^v}, \boldsymbol{N^c}, \boldsymbol{h}\}$. Specifically, it is worth pointing out that when the right-hand side (RHS) of Eq. (B1) is equal to zero, i.e. $\sum_{l \in M^c(r)} \hat{T}^m_{lr} = 0$, $\tilde{w}^c_r$ will value $+\infty$ as we define $f_{rs}(-\infty) = 0$. In that case, $\hat{w}^c_r$ will also be equal to $+\infty$, and then $\hat{N}^c_r$ takes the value $\lim_{w^c_r \to +\infty} w^c_r \cdot f_{rs}(w^c_r | \boldsymbol{T^m}, \boldsymbol{h})$ which remains finite.

Step 3 solves the following mathematical program IRVM by treating $\{\hat{\boldsymbol{T}}^{\boldsymbol{m}}, \hat{\boldsymbol{T}}^{\boldsymbol{o}}, \boldsymbol{h}\}$ as exogenous variables:

$$\begin{aligned} \text{(IRVM)} \quad & \min_{\boldsymbol{T^v}} \sum_{s \in S} \sum_{l \in L} \left[\pi_{sl} T^v_{sl} + \frac{1}{\theta} T^v_{sl} (\ln T^v_{sl} - 1) \right] \\ \text{s.t.} \quad & \sum_{s \in S} T^v_{sl} = \sum_{r \in M^v(l)} \hat{T}^m_{lr} && \forall l \in L && \text{(B2a)} \\ & \sum_{l \in L} T^v_{sl} = \sum_{r:(r,s) \in W} \hat{T}^o_{rs} && \forall s \in S && \text{(B2b)} \end{aligned}$$

where $\pi_{sl} = -\hat{F}_l + \gamma(\hat{h}_l + h_{sl})$. Defining $x \ln x|_{x=0} = 0$, we extend the domain of the objective function to cover zero flow states, i.e., $T^v_{sl} = 0$. Clearly, the objective function is continuous in $(\boldsymbol{T^v}, \boldsymbol{\pi})$ and strictly convex on $\boldsymbol{T^v}$ for each given $\boldsymbol{\pi}$. The two constraints (B2a,b) along with the flow non-negativity condition additionally define a compact convex feasible region for $\boldsymbol{T^v}$, which given its non-emptiness changes continuously in response to $\{\hat{\boldsymbol{T}}^{\boldsymbol{m}}, \hat{\boldsymbol{T}}^{\boldsymbol{o}}\}$. According to the Berge's maximum theorem, the optimal solution $\hat{\boldsymbol{T}}^{\boldsymbol{v}}$ to IRVM is a single-valued continuous function of inputs $\{\hat{\boldsymbol{T}}^{\boldsymbol{m}}, \hat{\boldsymbol{T}}^{\boldsymbol{o}}, \boldsymbol{\pi}\}$. Since the composition of continuous mapping preserve the continuity, $\hat{\boldsymbol{T}}^{\boldsymbol{v}}$ is continuous in $\{\hat{\boldsymbol{T}}^{\boldsymbol{m}}, \hat{\boldsymbol{T}}^{\boldsymbol{o}}, \boldsymbol{h}\}$.

The KKT conditions of IRVM can be stated as follows:

(B2a-b)

$$\pi_{sl} + \frac{1}{\theta} \ln T^v_{sl} + \beta_l + \tau_s = 0 \qquad \forall l \in L, s \in S \tag{B3}$$

where $\boldsymbol{\beta}$ and $\boldsymbol{\tau}$ are Lagrangian multipliers associated with (B2a) and (B2b). From (B3), we have

$$\pi_{sl} + \frac{1}{\theta} \ln T^v_{sl} + \beta_l = \pi_{sk} + \frac{1}{\theta} \ln T^v_{sk} + \beta_k$$

which by substituting $\boldsymbol{\pi}$ is equivalent to

$$\frac{T^v_{sl}}{T^v_{sk}} = \frac{\exp\left\{\theta\left[\hat{F}_l - \gamma\left(\hat{h}_l + h_{sl} + \frac{\beta_l}{\gamma}\right)\right]\right\}}{\exp\left\{\theta\left[\hat{F}_k - \gamma\left(\hat{h}_k + h_{sk} + \frac{\beta_k}{\gamma}\right)\right]\right\}} \qquad \forall l, k \in L, s \in S$$

and further yields

$$\frac{T^v_{sl}}{\sum_{r:(r,s)\in W} T^o_{rs}} = \frac{\exp\left\{\theta\left[\hat{F}_l - \gamma\left(\hat{h}_l + h_{sl} + \frac{\beta_l}{\gamma}\right)\right]\right\}}{\sum_{k\in L}\exp\left\{\theta\left[\hat{F}_k - \gamma\left(\hat{h}_k + h_{sk} + \frac{\beta_k}{\gamma}\right)\right]\right\}} \qquad \forall l \in L, s \in S$$

From the above formula, we can interpret the term β_l/γ as the RVs' waiting time at node l, i.e., $w^v_l = \beta_l/\gamma$ for all $l \in L$. Note that we can always add a constant η to $\boldsymbol{\beta}/\gamma$, such that the above equation still holds. Without loss of generality, we specify $\hat{\boldsymbol{\beta}}$ to be single-valued functions on $(\boldsymbol{\pi}, \hat{\boldsymbol{T}}^{\boldsymbol{v}})$, i.e.,

$$\hat{\beta}_l = -\pi_{sl} - \frac{1}{\theta} \ln \hat{T}^v_{sl} \qquad \forall l \in L, s \in S$$

Then, as per (6.12x),

$$\sum_{(r,s)\in W^c} \left(\hat{T}^v_{rs} + T^m_{rs} + T^o_{rs}\right) \cdot h_{rs} + \sum_{(s,l):s\in S, l\in L} \left(\frac{\hat{\beta}_l}{\gamma} + \hat{\eta}\right) \cdot \hat{T}^v_{sl} = N$$

which gives rise to

$$\hat{\eta} = \frac{N - \sum_{(r,s)\in W^c} \left(\hat{T}^v_{rs} + T^m_{rs} + T^o_{rs}\right) \cdot h_{rs} - \sum_{(s,l):s\in S, l\in L} \hat{T}^v_{sl} \cdot \hat{\beta}_l/\gamma}{\sum_{(s,l):s\in S, l\in L} \hat{T}^v_{sl}} \tag{B4}$$

and accordingly,

$$\hat{w}_l^v = \max\left\{0,\ \frac{\hat{\beta}_l}{\gamma} + \hat{\eta}\right\} \tag{B5}$$

It can be easily inferred from the above relationships that $\hat{\boldsymbol{w}}^{\boldsymbol{v}}$ is also continuous in $\{\hat{\boldsymbol{T}}^{\boldsymbol{m}}, \hat{\boldsymbol{T}}^{\boldsymbol{o}}, \boldsymbol{h}\}$. Therefore, by solving IRVM with given $\{\hat{\boldsymbol{T}}^{\boldsymbol{m}}, \hat{\boldsymbol{T}}^{\boldsymbol{o}}, \boldsymbol{h}\}$, we can retrieve $\hat{\boldsymbol{T}}^{\boldsymbol{v}}, \hat{\boldsymbol{w}}^{\boldsymbol{v}}$ and $\hat{\boldsymbol{N}}^{\boldsymbol{v}}$, all of which change continuously with respect to the inputs. Note that when $\hat{T}_{sl}^v$ values zero, the dual variable $\hat{\beta}_l$ and probably $\hat{\eta}$ will become $+\infty$. However, the corresponding $\hat{N}_l^v$ as we multiply $\hat{\beta}_l$ or $\hat{\eta}$ to $\hat{T}_{sl}^v$ will stay zero, which ensures the feasibility of Step 3 under the zero-flow circumstances.

As for Step 4, previous studies (see e.g., Hall (1978) and Lu and Nie (2010) for the details) showed that under mild conditions, the unique link flow pattern that solves PE as well as the resultant OD travel time are both continuous in response to the travel demands $\{\hat{\boldsymbol{T}}^{\boldsymbol{m}}, \hat{\boldsymbol{T}}^{\boldsymbol{o}}, \hat{\boldsymbol{T}}^{\boldsymbol{v}}\}$.

Combining the continuity of all stages above proves the proposition. □

Proposition C.5 then proves the existence of a compact and convex set Ω such that $\mathcal{F}: \{\boldsymbol{T}^{\boldsymbol{m}}, \boldsymbol{N}^{\boldsymbol{v}}, \boldsymbol{N}^{\boldsymbol{c}}, \boldsymbol{h}\} \to \{\hat{\boldsymbol{T}}^{\boldsymbol{m}}, \hat{\boldsymbol{N}}^{\boldsymbol{v}}, \hat{\boldsymbol{N}}^{\boldsymbol{c}}, \hat{\boldsymbol{h}}\}$ maps from Ω to itself.

Proposition C.5. *There exists a set of bounds $\{\bar{T}^m, \bar{N}^v, \bar{N}^c, h^o, h^u\}$ such that for any point in Ω,*

$$\Omega = \left\{ \boldsymbol{T}^{\boldsymbol{m}}, \boldsymbol{N}^{\boldsymbol{v}}, \boldsymbol{N}^{\boldsymbol{c}}, \boldsymbol{h} \;\middle|\; \begin{array}{rcll} T_{lr}^m & \in & [0, \bar{T}^m], & \forall (l,r) \in \{(l,r) | l \in L, r \in M^v(l)\} \\ N_r^c & \in & [0, \bar{N}^c], & \forall r \in R \\ N_l^v & \in & [0, \bar{N}^v], & \forall l \in L \\ h_{ij} & \in & [h^o, h^u], & \forall (i,j) \in A \end{array} \right\}$$

The mapping $\mathcal{F}$ maps it back to a point in Ω.

Proof. We first expound the existence of $\bar{N}^c$ and $\bar{N}^v$. The boundedness of N^c can be concluded as the customer demand functions $\{f_{rs}\}_{(r,s)\in W}$ are capped on the top and the limit $\lim_{w_r^c \to +\infty} w_r^c \cdot f_{rs}(w_r^c | \boldsymbol{h})$ stays finite for all $r \in R$. Meanwhile, the being of $\bar{N}^v$ is simply ensured by the finite number of RV fleets N.

Then, we prove the existence of $\hat{T}^m$. Since $Z_{lr}^{-1}(\boldsymbol{T}^{\boldsymbol{m}}, N_l^v, N_r^c, h_{lr})$ is monotonically increasing on N_l^v, N_r^c and decreasing on h_{lr}, we have the following inequality hold on Ω,

$$\Gamma_{lr}(\boldsymbol{T}^{\boldsymbol{m}}, N_l^v, N_r^c, h_{lr}) \leqslant \Gamma_{lr}(\boldsymbol{T}^{\boldsymbol{m}}, \bar{N}^v, \bar{N}^c, h^o), \quad \forall l \in L, r \in M^v(l)$$

Then, by following the same procedure as to find the upper bound in Proposition C.2, we can

retrieve an upper bound $\bar{T}^m$ such that

$$\Gamma_{lr}(\boldsymbol{T^m}, \bar{N}^v, \bar{N}^c, h^o) \leqslant \bar{T}^m$$

for any $\boldsymbol{T^m} \leqslant \bar{T}^m \cdot \nVdash$ given $\bar{N}^v, \bar{N}^c$ and h^o. Note that the lower bound h^0 is guaranteed by the positive free-flow travel time on links. Combining the two inequalities thus yields a common upper bound $\bar{T}^m$ for $\hat{\boldsymbol{T}}^{\boldsymbol{m}}$.

Since the domain of $\hat{\boldsymbol{T}}^{\boldsymbol{m}}$ is bounded, both $\hat{\boldsymbol{T}}^{\boldsymbol{o}}$ and $\hat{\boldsymbol{T}}^{\boldsymbol{v}}$ by flow conservations will also stay finite. The boundedness of travel demand then prevents link travel time from being infinitely large, thereby yielding the existence of the upper bound h^u.

All the above boundedness guards a compact and convex set Ω, on which the mapping $\mathcal{F}$ maps from Ω to itself. □

Joining Proposition C.4 and C.5 concludes the existence of a fixed point for $\mathcal{F}$. However, to ensure the executability, the proposed procedure has strategically embedded several step functions, which could potentially deviate the fixed point of $\mathcal{F}$ from the system equilibrium implied. To radically show the existence, we thus demonstrate in Proposition C.6 that those cut-offs introduced by step functions will not be active on any fixed point achieved. Thence, the existence of a fixed point to $\mathcal{F}$ essentially indicate a rational equilibrium of system states.

Proposition C.6. *On any fixed point of the mapping $\mathcal{F}$, the following three step functions will not activate the cut-offs,*

$$\textit{Step 3:}\ \ \hat{w}_r^c = \max\{0, \tilde{w}_r^c\} \qquad \forall r \in R$$

$$\textit{Step 4:}\ \ \hat{w}_l^v = \max\{0, \tilde{w}_l^v\},\ \ \hat{N}_l^v = \min\{N, \hat{w}_l^v \cdot \sum_{s\in S} \hat{T}_{sl}^v\} \qquad \forall l \in L$$

Proof. We first prove by contradictions that $\tilde{\boldsymbol{w}}^{\boldsymbol{c}}, \tilde{\boldsymbol{w}}^{\boldsymbol{v}} \geqslant \mathbf{0}$ for any fixed point of $\mathcal{F}$. Without loss of generality, assume $\tilde{w}_r^c < 0$ for a node $r \in R$. In this case, we must have $\hat{N}_r^c = N_r^c = 0$, and then $\hat{T}_{lr}^m = 0$ for all $l \in M^c(r)$ which further leads to $\tilde{w}_r^c = +\infty$ and $\hat{N}_r^c = \lim_{w_r^c \to +\infty} w_r^c \cdot f_{rs}(w_r^c) > 0$ by the premise. This, however, creates a contradiction and therefore $\tilde{\boldsymbol{w}}^{\boldsymbol{c}} \geqslant \mathbf{0}$.

Next, we assume $\tilde{w}_l^v < 0$ and thus $\hat{N}_l^v = N_l^v = 0$ on a node $l \in L$. Applying the mapping procedures subsequently yield $\hat{T}_{lr}^m = 0$ for all $r \in M^v(l)$ and $\hat{T}_{sl}^v = 0$ for all $s \in S$. This requires $\hat{\beta}_l = +\infty$ and further $\hat{\eta} = \tilde{w}_l^v - \frac{\hat{\beta}_l}{\gamma} = -\infty$, which as per Eq. (B4) holds only if $\hat{\boldsymbol{T}}^{\boldsymbol{v}} = \mathbf{0}$. However, the necessary condition for $\hat{\boldsymbol{T}}^{\boldsymbol{v}} = \mathbf{0}$ is to have $\hat{\boldsymbol{T}}^{\boldsymbol{m}} = \hat{\boldsymbol{T}}^{\boldsymbol{o}} = \mathbf{0}$, thereby giving rise to $\eta = +\infty$ and yielding a contradiction. Thus, $\tilde{\boldsymbol{w}}^{\boldsymbol{v}}$ must also be nonnegative.

With the above foundations, we can readily conclude that on any fixed point $\hat{w}_l^v \cdot \sum_{s\in S} \hat{T}_{sl}^v \leqslant N$ for all $l \in L$. Otherwise, due to the fleet size conservation, there must exist at least one node

among L where the waiting time of RVs $\tilde{\boldsymbol{w}}^v$ becomes negative. Such condition results in the same contradiction described above. □

With Proposition C.6 excluding the influences of the embedded treatments in the mapping procedure, we now conclude the existence of an equilibrium for the proposed frameworks.

Theorem C.2. *There always exists a solution to the network equilibrium system with inter-node matching (6.12).*

As a special case of inter-node matching, the above property can be readily extrapolated to intra-node matching.

Lemma C.1. *A solution to the intra-node matching system (6.7) is always guaranteed.*

C.4 Parametric settings for numerical experiments on the Nguyen-Dupius network

Table C.2: Characteristics of the Nguyen-Dupius network

(A) Parameters for link performance functions

Link	a_{ij}	b_{ij}	Link	a_{ij}	b_{ij}	Link	a_{ij}	b_{ij}
1-5	7	0.0125	6-10	13	0.005	10-11	6	0.0025
1-12	9	0.01	6-5	3	0.0075	10-6	13	0.005
2-8	9	0.0125	6-12	7	0.0025	10-9	10	0.005
2-11	9	0.005	7-8	5	0.0125	11-2	9	0.005
3-11	8	0.01	7-11	9	0.0125	11-3	8	0.01
3-13	11	0.01	7-6	5	0.0125	11-7	9	0.0125
4-5	9	0.01	8-2	9	0.0125	11-10	6	0.0025
4-9	12	0.005	8-7	5	0.0125	12-6	7	0.0025
5-6	3	0.0075	8-12	14	0.01	12-8	14	0.01
5-9	9	0.0075	9-10	10	0.005	12-1	9	0.01
5-1	7	0.0125	9-13	9	0.005	13-3	11	0.01
5-4	9	0.01	9-4	12	0.005	13-9	9	0.005
6-7	5	0.0125	9-5	9	0.0075			

The units of a_{ij} and b_{ij} are minute and minutes per unit flow, respectively.

(B) Regular traffic demand

O\D	1	2	3	4	5
1	0	225	600	0	0
2	150	0	0	450	300
3	450	0	0	300	240
4	0	375	150	0	0
5	0	150	225	0	0

(C) Potential ride-sourcing service demand (Symmetric/Asymmetric)

O\D	1	2	3	4	5
1	0	150/150	400/400	0	0
2	100/100	0	0	300/120	200/80
3	300/300	0	0	200/80	160/64
4	0	250/430	100/220	0	0
5	0	100/220	150/246	0	0

(D) Parameter values

Param	Value	Unit	Param	Value	Unit	Param	Value
β^o	20	\$/h	F_0	2	\$	q_N^v, q_N^c	1
β^i	6	\$/h	β^f	60	\$/h	q_T^v, q_T^c	0.1
β^m	15	\$/h	$\hat{\theta}$	0.01	1/\$	q_h	0.1
γ	10	\$/h	θ	0.5		η	10

APPENDIX D

Appendices for Chapter 7

D.1 Nomenclature

Table D.1: Notation list of variables

Notation	Description
System state variables	
F	Trip fare for an average distance L_t (\$)
N	RV fleet size
K	Number of provided parking spaces
w^c	Average customer waiting time for getting matched (h)
w^t	Average RV waiting time for getting matched (h)
w^m	Average meeting time for customers and RVs after being matched (h)
N^p	Number of parked RVs
Q	Hourly ride-sourcing customer demand
v	Average network travel speed (mph)
Other state variables	
m^{c-t}	Hourly matching rate between unmatched customers and RVs
J^u, J^u_d	Total number and density of unmatched customers
N^u, N^u_d	Total number and density of unmatched RVs
r^m	Average meeting distance of matched RVs
m^{p-t}	Hourly matching rate between cruising RVs and vacant parking spaces
K^v	Number of vacant parking spaces
N^c	Number of cruising RVs
N^b	Number of vehicles in background traffic
Q^b	Background traffic demand
d^t	Average density of running RVs
A^p	Total area for parking spaces
P_m	Probability for an unmatched RV getting matched in each time step

r	Dominant radius of each unmatched RV
w^f	Average parking search time (h)
S	Total social benefits ($)
Π	Total profit of ride-sourcing company/drivers ($)
F^c	Commission fee that the ride-sourcing company charged per order

Table D.2: Parameters and their default values in the numerical examples

Notation	Interpretation	Default Value
τ	Interval of matching steps (h)	3×10^{-3}
A_w	Area of the whole region under RV service (mi^2)	20
L_r	Total length of the road system (mi)	500
μ	Parameter adjusts Euclidean distance to Manhattan distance	$\sqrt{\pi/2}$
Q^0	Potential ride-sourcing customer demand per hour	1.5×10^7
Q_b^0	Potential background traffic demand per hour	1.5×10^7
κ	Demand sensitivity parameter for ride-sourcing services (1/$)	0.25
κ_b	Demand sensitivity parameter for background traffic (1/h)	1
β	Out-of-vehicle value of time ($/h)	40
γ	In-vehicle value of time ($/h)	20
v_f	Free-flow travel speed (mph)	30
k_j	Jam density (veh/$\mathrm{mi}\cdot\mathrm{lane}$)	250
A_u	Total utilizable area for roads and parking facilities (lane·mi)	2000
L_b	Average travel distance in background traffic (mi)	5
L_t	Average distance of trips served by ride-sourcing services (mi)	2.5
c_b	Hourly cost per-vehicle in background traffic ($)	30
c_p	Operational cost for the e-hailing platform per order ($)	2
c_v	Hourly operational cost for a running RV ($)	25
c_s	Hourly savings for a single parked RV ($)	10

D.2 Relationships between RVs and customers' average waiting time in online matching

By rewriting Eq. 7.7, we can present the average waiting time of customers to be matched w^c as a function of RVs' waiting time w^t, i.e.,

$$w^c = w^t \cdot \ln \frac{w^t}{w^t - \tau}$$

The first- and second-order derivatives of w^c on w^t can then be derived as

$$\frac{\mathrm{d}w^c}{\mathrm{d}w^t} = \ln\left(1 + \frac{\tau}{w^t - \tau}\right) - \frac{\tau}{w^t - \tau}$$

$$\frac{\mathrm{d}^2 w^c}{\mathrm{d}{w^t}^2} = \frac{2\tau^2}{w^t (w^t - \tau)^2}$$

Since $\ln(1+x) - x < 0$ when $x > 0$ and w^t is always greater than τ, we have $\mathrm{d}w^c/\mathrm{d}w^t < 0$ and $\mathrm{d}^2 w^c/\mathrm{d}{w^t}^2 > 0$. That means w^c is a monotonically decreasing convex function on w^t, and vice versa.

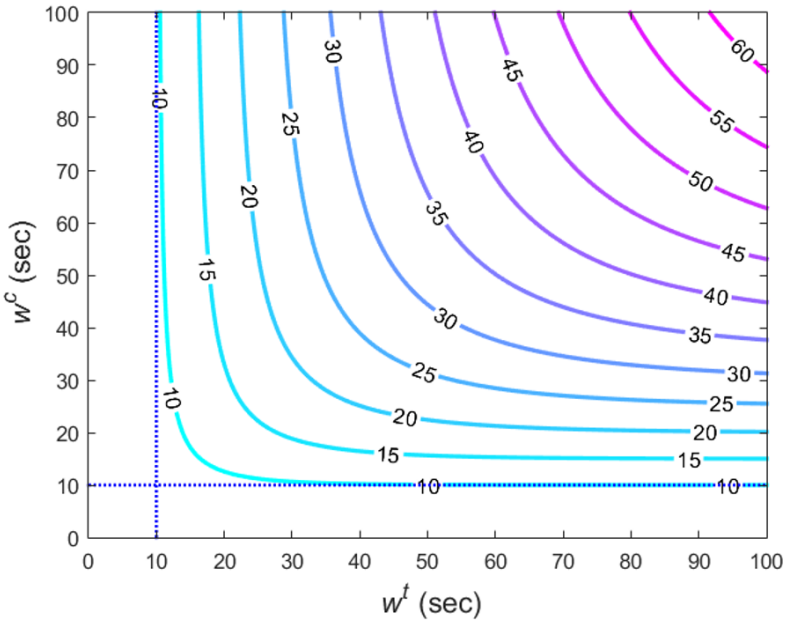


Figure D.1: Relationships between w^c and w^t under different τ [Note: the value of τ is directly numbered on curves (unit: sec)]

D.3 Extreme condition analysis for the deductive meeting distance formula

The expected meeting distance is derived as

$$r^m\left(J_d^u, N_d^u\right) \cong \frac{\mu}{1-\exp\left(-\frac{J_d^u}{N_d^u}\right)} \cdot \left[\frac{1}{2\sqrt{J_d^u}} \cdot \operatorname{erf}\left(\sqrt{\frac{J_d^u}{N_d^u}}\right) - \frac{1}{\sqrt{\pi N_d^u}} \cdot \exp\left(-\frac{J_d^u}{N_d^u}\right)\right] \tag{B1}$$

We analyze in this appendix the asymptotic expression for Eq. (B1) under two extreme conditions, respectively being $J_d^u \gg N_d^u$ and $J_d^u \ll N_d^u$:

When $J_d^u \gg N_d^u$ such that the "competition" among RVs over a customer becomes nil, we have $\sqrt{\frac{J_d^u}{N_d^u}} \approx +\infty$, $\operatorname{erf}\left(\sqrt{\frac{J_d^u}{N_d^u}}\right) \approx 1$ and $\exp\left(-\frac{J_d^u}{N_d^u}\right) \ll \sqrt{\frac{N_d^u}{J_d^u}}$. This gives rise to $r^m = \frac{\mu}{2\sqrt{J_d^u}}$, a result consistent with the meeting distance formula suggested by Daganzo (1978) and Arnott (1996) for the demand responsive transit and radio-dispatched taxi market, respectively. The difference is that in their case, the number of customers is much smaller than the number of vehicles (i.e., $J_d^u \ll N_d^u$) while vehicles' distributions follow the spatial Poisson point process, but here the condition reverses.

In our model, when $J_d^u \ll N_d^u$, the two terms $\operatorname{erf}(\cdot)$ and $\exp(\cdot)$ in Eq. (B1) can be Taylor expanded as the following series:

$$\operatorname{erf}\left(\sqrt{\frac{J_d^u}{N_d^u}}\right) = \frac{2}{\sqrt{\pi}}\left(\sqrt{\frac{J_d^u}{N_d^u}} - \frac{2}{3!}\cdot\left(\sqrt{\frac{J_d^u}{N_d^u}}\right)^3 + \mathrm{O}\left[\left(\sqrt{\frac{J_d^u}{N_d^u}}\right)^5\right]\right) \tag{B2a}$$

$$\exp\left(-\frac{J_d^u}{N_d^u}\right) = 1 - \frac{J_d^u}{N_d^u} + \mathrm{O}\left[\left(\frac{J_d^u}{N_d^u}\right)^2\right] \tag{B2b}$$

Substituting Eqs. (B2) into Eq. (B1) gives

$$\begin{aligned} r^m &= \mu \cdot \frac{\frac{1}{2\sqrt{J_d^u}} \cdot \frac{2}{\sqrt{\pi}}\left(\sqrt{\frac{J_d^u}{N_d^u}} - \frac{2}{3!}\cdot\left(\sqrt{\frac{J_d^u}{N_d^u}}\right)^3 + \mathrm{O}\left[\left(\sqrt{\frac{J_d^u}{N_d^u}}\right)^5\right]\right) - \frac{1}{\sqrt{\pi N_d^u}} \cdot \left(1 - \frac{J_d^u}{N_d^u} + \mathrm{O}\left[\left(\frac{J_d^u}{N_d^u}\right)^2\right]\right)}{\frac{J_d^u}{N_d^u} + \mathrm{O}\left[\left(\frac{J_d^u}{N_d^u}\right)^2\right]} \\ &= \frac{\mu}{\sqrt{\pi N_d^u}} \cdot \frac{\frac{2}{3}\cdot\frac{J_d^u}{N_d^u} + \mathrm{O}\left[\left(\frac{J_d^u}{N_d^u}\right)^2\right]}{\frac{J_d^u}{N_d^u} + \mathrm{O}\left[\left(\frac{J_d^u}{N_d^u}\right)^2\right]} \approx \frac{2\mu}{3\sqrt{\pi N_d^u}} \end{aligned}$$

These formulas suggest that when the size of one side is sufficiently larger than the other, the matching distance mostly relies on the denser side, inversely proportional to the square root

of its density. Since RVs and customers are assumed to follow distinct distributions (the spatial Poisson point process for customers versus uniform for RVs), the coefficients in formulas of the two limiting cases differ from each other.

D.4 Existence of equilibrium for the deductive nonlinear system

The existence of equilibrium under combinations of (N, F, K) is proved using Brouwer's fixed point theorem (De la Fuente, 2000). For analytical convenience, we substitute variable Q with N^u/w^t throughout the proof. The proof proceeds as follows. We firstly construct a compact and convex set Ω on a variable space (w^m, N^u, v). For each element $(w_b^m, N_b^u, v_b) \in \Omega$, we derive three dependent variables (w^t, w^c, N^p) by solving the subsystem below:

$$w^c = w^t \cdot \ln \frac{w^t}{w^t - \tau} \tag{B3a}$$

$$N \cdot \left(w^t + w_b^m + \frac{L_t}{v_b} \right)^{-1} = Q^0 \cdot \exp\left[-\kappa \left(F + \beta \left(w^c + w_b^m \right) + \gamma \frac{L_t}{v_b} \right) \right] \tag{B3b}$$

$$N^p \cdot \left[1 + \frac{L_r}{(K - N^p) \cdot v_b \cdot w^t} \right] = \frac{N \cdot w^t}{w^t + w_b^m + \frac{L_t}{v_b}} \tag{B3c}$$

Eq. (B3a) is directly modified from Eq. (7.11a), while Eqs. (B3b) and (B3c) integrate Eq. (7.11d) respectively with Eqs. (7.11e) and (7.11c) from the original system. We prove that the solution (w^t, w^c, N^p) is unique given each pair of (w_b^m, N_b^u, v_b), and changes continuously with respect to the latter three variables.

The subsystem (w^t, w^c, N^p) can be projected back into the (w^m, N^u, v) space via the remaining three relationships of system 7.11, respectively given as:

$$w_a^m = \frac{\mu / v_b}{1 - \exp\left(-\frac{w^c}{w^t}\right)} \cdot \left[\frac{1}{2} \sqrt{\frac{A_w w^t}{N_b^u w^c}} \cdot \operatorname{erf}\left(\sqrt{\frac{w^c}{w^t}} \right) - \frac{1}{\sqrt{\pi}} \sqrt{\frac{A_w}{N_b^u}} \cdot \exp\left(-\frac{w^c}{w^t} \right) \right] \tag{B4a}$$

$$N_a^u = \frac{N \cdot w^t}{w^t + w_b^m + \frac{L_t}{v_b}} \tag{B4b}$$

$$\exp\left(\frac{\kappa_b \cdot L_b}{v_a} \right) \cdot \left[v_f \cdot \left(1 - \frac{N - N^p}{k_j A_u - K} \right) - v_a \right] = \frac{v_f \cdot Q_b^0 \cdot L_b}{k_j A_u - K} \cdot \frac{1}{v_b} \tag{B4c}$$

Eqs. (B3) and (B4) together construct a mapping from the variable space (w^m, N^u, v) to itself. We then show that the realized (w_a^m, N_a^u, v_a) will always be in set Ω. By virtue of Brouwer's fixed point theorem, such a continuous mapping from Ω to its subset guarantees the existence of a solution that solves the whole system.

We start by proving the following propositions for the subsystem (B3) given (w_b^m, v_b).

Proposition D.1. *For each pair of (w_b^m, v_b), there exists a unique pair of (w^c, w^t) that solves Eqs. (B3a) and (B3b).*

Proof. Firstly, according to Appendix D.2, $\mathrm{d}w^c/\mathrm{d}w^t < 0$. By substituting w^c with its equivalence from Eq. (B3a), Eq. (B3b) is then written as an equation purely on w^t:

$$\frac{\exp\left[\kappa\beta\cdot w^c\left(w^t\right)\right]}{w^t+w_{\mathrm{b}}^m+\frac{L_t}{v_b}}=Q^0\cdot N^{-1}\cdot\exp\left[-\kappa\left(F+\beta w_b^m+\gamma\frac{L_t}{v_b}\right)\right] \tag{B5}$$

We denote the left-hand side formula as a function $L_1\left(w^t\right)$. When $w^t\to\tau$, we have $w^c\to+\infty$ and thus $\lim_{w^t\to\tau}L_1\left(w^t\right)=+\infty$; When $w^t\to+\infty$, we have $w^c\to\tau$ and thus $\lim_{w^t\to+\infty}L_1\left(w^t\right)=0$. As $L_1\left(w^t\right)$ is decreasing with w^t and the right-hand side of Eq. (B5) is a finite constant, there exists a unique w^t solving the equation. The unique w^c can then be determined by applying Eq. (B3a). □

Proposition D.2. *The solution (w^c, w^t) to Eqs. (B3a), (B3b) changes continuously with respect to (w_b^m, v_b).*

The proof is straightforward and thus is omitted here.

Proposition D.3. *Given (w_b^m, v_b, w^t), the solution N_p to Eq. (B3c) is unique in $(0, \min\{K, N\})$.*

Proof. Denote the left-hand side of Eq. (B3c) as a function $L_2\left(N^p\right)$. Then, we have $\lim_{N^p\to0}L_2\left(N^p\right)=0$ and $\lim_{N^p\to K}L_2\left(K\right)=+\infty$. Since $L_2\left(N^p\right)$ is strictly increasing with N^p and the right-hand side of Eq. (B3c) is less than N, there exists a unique N^p between 0 and K that solves the equation. When $K>N$, we have the left-hand side $L_2\left(N\right)>N$, while the right-hand side is less than N. Therefore, the solution N^p in this case should be less than N. □

Proposition D.4. *The solution N_p to Eq. (B3c) changes continuously with respect to (w_b^m, v_b).*

Then, we prove the validity of Eq. (B4c) as a mapping from N^p, v_b to v_a.

Proposition D.5. *The solution v_a to Eq. (B4c) is unique given (N^p, v_b), and changes continuously with respect to the latter two parameters.*

Proof. Again, we only prove the existence and uniqueness of a solution v_a. By denoting the left-hand side of Eq. (B4c) as $L_3\left(v_a\right)$, we then have $\lim_{v_a\to0}L_3\left(v_a\right)=+\infty$ and $L_3\left(v_f\right)<0$. Since $L_3\left(v_a\right)$ is strictly decreasing on v_a and the right-hand side of Eq. (B4c) is a positive constant, there exists a unique v_a between 0 and v_f that solves the equation. □

As the other two functions in Eq. (B4) are both explicit and continuous on $(w_b^m, N_b^u, v_b, w^t, w^c, N^p)$, Propositions D.1-D.5 give rise to the following theorem:

Theorem D.1. *When fixing (N, F, K), function sets (B3) and (B4) define a continuous mapping from (w_b^m, N_b^u, v_b) to (w_a^m, N_a^u, v_a).*

Next, we prove that there exists a nonempty, compact and convex set Ω such that $(w_a^m, N_a^u, v_a) \in \Omega$ holds for any $(w_b^m, N_b^u, v_b) \in \Omega$.

Proposition D.6. *There exist* $v_s, v_l \in (0, v_f)$ *and* $v_s < v_l$ *such that* $v_a \in [v_s, v_l]$ *for any* $v_b \in [v_s, v_l]$.

Proof. To preclude the system from falling into complete gridlock, we assume $N + K$ is always be less than $k_j A_u$. Denote $L_4(v_a, N^p, v_b)$ as

$$L_4(v_a, N^p, v_b) = \exp\left(\frac{\kappa_b \cdot L_b}{v_a}\right) \cdot \left[v_f \cdot \left(1 - \frac{N - N^p}{k_j A_u - K}\right) - v_a\right] - \frac{v_f \cdot Q_b^0 \cdot L_b}{k_j A_u - K} \cdot \frac{1}{v_b}$$

Eq. (B4c) thus pertains to $L_4(v_a, N^p, v_b) = 0$. Note that the equation $L_4(v, N^p, v) = 0$ on v has feasible solutions for all $N^p \in [0, N]$, because L_4 is continuous on v, $\lim_{v \to 0} L_4(v, N^p, v) \sim O\left(\exp\left(\frac{\kappa_b \cdot L_b}{v_a}\right)\right) = +\infty$ and $L_4(v_f, N^p, v_f) < 0$. Let v_s and v_l be the largest v that solve $L_4(v, 0, v) = 0$ and solve $L_4(v, N, v) = 0$, respectively. Then, $L_4(v, N, v) < 0$ for $\forall v > v_l$. Since $L_4(v_s, N, v_s) > L_4(v_s, 0, v_s) = 0$, we have v_s always be less than v_l.

Such a pair of v_s and v_l are proven to satisfy the condition that for any $v_b \in [v_s, v_l]$, v_a is also in $[v_s, v_l]$. We denote $v_a = I(N^p, v_b)$ as the mapping from (N^p, v_b) to v_a by solving $L_4(v_a, N^p, v_b) = 0$. Then,

$$\frac{\partial L_4}{\partial v_a} \cdot \frac{\partial I}{\partial v_b} + \frac{\partial L_4}{\partial v_b} = 0 \tag{B6a}$$

$$\frac{\partial L_4}{\partial v_a} \cdot \frac{\partial I}{\partial N^p} + \frac{\partial L_4}{\partial N^p} = 0 \tag{B6b}$$

Since it can be easily proven that $\frac{\partial L_4}{\partial N^p} > 0$, $\frac{\partial L_4}{\partial w_b} > 0$ and $\frac{\partial L_4}{\partial w_a} < 0$ when $v_a = I(N^p, v_b)$, we have $\frac{\partial I}{\partial N^p} > 0$ and $\frac{\partial I}{\partial v_b} > 0$, which further result in the following two inequalities when $v_b \in [v_s, v_l]$:

$$v_a = I(N^p, v_b) \leq I(N, v_l) = v_l$$
$$v_a = I(N^p, v_b) \geq I(0, v_s) = v_s$$

This completes the proof. □

Proposition D.7. w_a^m *realized by Eq. (B4a) should fall in set* $\left[0, \frac{2}{3\sqrt{\pi}} \cdot \frac{\mu}{v_b}\sqrt{\frac{A_w}{N_b^u}}\right]$.

Proof. We reorganize Eq. (B4a) as follows,

$$w_a^m = \frac{\mu}{v_b}\sqrt{\frac{A_w}{N_b^u}} \cdot \frac{\frac{1}{2}\sqrt{\frac{w^t}{w^c}} \cdot \mathrm{erf}\left(\sqrt{\frac{w^c}{w^t}}\right) - \frac{1}{\sqrt{\pi}} \cdot \exp\left(-\frac{w^c}{w^t}\right)}{1 - \exp\left(-\frac{w^c}{w^t}\right)} \tag{B8}$$

Then, the problem becomes to prove that the right-hand side of Eq. (B8) excluding $\frac{\mu}{v_b}\sqrt{\frac{A_w}{N_b^u}}$ is bounded. Denote w^c/w^t as g. It is easy to derive from Eq. (B3a) that $\frac{\mathrm{d}g}{\mathrm{d}w^t} < 0$ and $\frac{\mathrm{d}g}{\mathrm{d}w^c} > 0$. Then, we substitute g into Eq. (B8) and rewrite it as $w_a^m = \frac{\mu}{v_b}\sqrt{\frac{A_w}{N_b^u}} \cdot G(g)$, where $G(g) = \frac{\frac{1}{2}\cdot\mathrm{erf}\left(\sqrt{g}\right) - \frac{1}{\sqrt{\pi}}\sqrt{\mathrm{g}}\cdot\exp(-g)}{\sqrt{g}\cdot[1-\exp(-g)\,]}$.

The derivative of $G(g)$ is derived as

$$G'(g) \leq -\frac{\exp(-g)\ \cdot\left[\sqrt{\mathrm{g}} - \frac{\sqrt{\pi}}{2}\cdot\mathrm{erf}\left(\sqrt{g}\right)\right]}{\sqrt{\pi\cdot g}\cdot[1-\exp(-g)\,]^2}$$

Since $x - \int_0^x e^{-t^2}\mathrm{d}t > 0$ when $x > 0$, we have $\sqrt{\mathrm{g}} - \frac{\sqrt{\pi}}{2}\cdot\mathrm{erf}\left(\sqrt{g}\right) > 0$ which further yields $G'(g) < 0$. As can be easily derived that $\lim_{g\to+\infty} G(g) = 0$, we have $G(g) \geq 0$ and thus $w_a^m \geq 0$. Further, as derived in Appendix D.3 that $\lim_{g\to 0} G(g) = \frac{2}{3\sqrt{\pi}}$, the monotonicity of G then gives rise to $G(g) \leq \frac{2}{3\sqrt{\pi}}$. Therefore, we have $w_a^m \leq \frac{2}{3\sqrt{\pi}}\cdot\frac{\mu}{v_b}\sqrt{\frac{A_w}{N_b^u}}$. □

This implies if N_b^u is bounded from below by a positive number M_N, w_a^m will fall into the set $\left[0, \frac{2}{3\sqrt{\pi}}\cdot\frac{\mu}{v_s}\sqrt{\frac{A_w}{M_N}}\right]$ as long as w_b^m is nonnegative.

Proposition D.8. *There exists a constant $M_N \in (0, N)$ such that $N_a^u \in [M_N, N]$ for any positive N_b^u.*

Proof. We assume that such an M_N does not exist and prove by contradiction. The nonexistence of M_N implies that N_a^u can be infinitely close to 0, which from Eq. (B4b) is ascribed to the unboundedness of $\frac{w_b^m}{w^t}$. Since w^t is greater than τ, the necessary condition for $\frac{w_b^m}{w^t}$ to be infinitely large is to have w_b^m approach infinity. However, when $w_b^m \to +\infty$, Eq. (B5) shows that $w^t \sim O(\exp(\kappa\beta w_b^m)\,)$ which further leads to $\frac{w_b^m}{w^t} \to 0$. This contradicts with the assumption that $\frac{w_b^m}{w^t}$ can be infinitely large. □

Combining Proposition D.6-D.8 leads to the following theorem:

Theorem D.2. *There exist constants $v_s, v_l \in (0, v_f)$ and $M_N \in (0, N)$ such that the mapping by Eq. (B3), (B4) from*

$$\Omega = \left\{ (w^m, N^u, v) \;\middle|\; \begin{array}{l} w^m \in \left[0, \frac{2}{3\sqrt{\pi}}\cdot\frac{\mu}{v_s}\sqrt{\frac{A_w}{M_N}}\right] \\ N^u \in [M_N,\ \ N] \\ v \in [v_s,\ \ v_l] \end{array} \right\}$$

will be projected to a subset in Ω.

It is clear that the set Ω in Theorem D.2 is nonempty, compact and convex on (w^m, N^u, v). Thus, as per Brouwer's fixed point theorem, Theorem D.1 and D.2 together confirm the general existence of equilibrium for the deductive nonlinear system. We state it as Theorem D.3.

Theorem D.3. *When the RV fleet size N does not exceed the total road space, the deductive nonlinear system is always guaranteed with a solution.*

BIBLIOGRAPHY

Afèche, P., Liu, Z., and Maglaras, C. (2018). Ride-hailing networks with strategic drivers: The impact of platform control capabilities on performance. Available at SSRN: https://ssrn.com/abstract=3120544. Accessed on February 4, 2019.

Anderson, D. N. (2014). "not just a taxi"? for-profit ridesharing, driver strategies, and vmt. *Transportation*, 41(5):1099–1117.

Anderson, M. (2019). Mobile technology and home broadband 2019. *Pew Research Center*.

Arnott, R. (1996). Taxi travel should be subsidized. *Journal of Urban Economics*, 40(3):316–333.

Arnott, R. and Inci, E. (2006). An integrated model of downtown parking and traffic congestion. *Journal of Urban Economics*, 60(3):418–442.

Arnott, R. and Inci, E. (2010). The stability of downtown parking and traffic congestion. *Journal of Urban Economics*, 68(3):260–276.

Arnott, R., Inci, E., and Rowse, J. (2015). Downtown curbside parking capacity. *Journal of Urban Economics*, 86:83–97.

Arnott, R. and Rowse, J. (1999). Modeling parking. *Journal of urban economics*, 45(1):97–124.

Arnott, R. and Rowse, J. (2009). Downtown parking in auto city. *Regional Science and Urban Economics*, 39(1):1–14.

Ban, J. X., Dessouky, M., Pang, J.-S., and Fan, R. (2018). A general equilibrium model for transportation systems with e-hailing services and flow congestion. *Manuscript, Department of Industrial and Systems Engineering, University of Southern California*.

Ban, X. J., Dessouky, M., Pang, J.-S., and Fan, R. (2019). A general equilibrium model for transportation systems with e-hailing services and flow congestion. *Transportation Research Part B: Methodological*, 129:273–304.

Banerjee, S., Johari, R., and Riquelme, C. (2016). Dynamic pricing in ridesharing platforms. *ACM SIGecom Exchanges*, 15(1):65–70.

Bimpikis, K., Candogan, O., and Saban, D. (2019). Spatial pricing in ride-sharing networks. *Operations Research*.

Braverman, A., Dai, J. G., Liu, X., and Ying, L. (2019). Empty-car routing in ridesharing systems. *Operations Research*, 67(5):1437–1452.

Buchholz, N. (2018). Spatial equilibrium, search frictions and dynamic efficiency in the taxi industry. Available at https://scholar.princeton.edu/nbuchholz/research. Accessed on February 4, 2019.

Cachon, G. P., Daniels, K. M., and Lobel, R. (2017). The role of surge pricing on a service platform with self-scheduling capacity. *Manufacturing & Service Operations Management*, 19(3):368–384.

Cao, J. and Menendez, M. (2015). System dynamics of urban traffic based on its parking-related-states. *Transportation Research Part B: Methodological*, 81:718–736.

Castiglione, J., Chang, T., Cooper, D., Hobson, J., Logan, W., Young, E., Charlton, B., Wilson, C., Mislove, A., Chen, L., et al. (2017). Tncs today: a profile of san francisco transportation network company activity. Technical report, San Francisco County Transportation Authority.

Castiglione, J., Cooper, D., Sana, B., Tischler, D., Chang, T., Erhardt, G. D., Roy, S., Chen, M., and Mucci, A. (2018). Tncs & congestion. Technical report, San Francisco County Transportation Authority.

Castillo, J. C., Knoepfle, D. T., and Weyl, E. G. (2018). Surge pricing solves the wild goose chase. Available at SSRN: https://ssrn.com/abstract=2890666. Accessed on May 15, 2018.

Chen, M. K. and Sheldon, M. (2016). Dynamic pricing in a labor market: Surge pricing and flexible work on the uber platform. In *Ec*, page 455.

Chiu, S. N., Stoyan, D., Kendall, W. S., and Mecke, J. (2013). *Stochastic geometry and its applications*. John Wiley & Sons.

Cooper, J., Farrell, S., and Simpson, P. (2010). Identifying demand and optimal location for taxi ranks in a liberalized market. Technical Report No. 10-3146, Transportation Research Board 89th Annual Meeting.

Cox, D. (1972). Regression models and life-tables. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 187–220.

Cramer, J. and Krueger, A. B. (2016). Disruptive change in the taxi business: The case of uber. *American Economic Review*, 106(5):177–82.

Daganzo, C. F. (1978). An approximate analytic model of many-to-many demand responsive transportation systems. *Transportation Research*, 12(5):325–333.

Daganzo, C. F. (2007). Urban gridlock: Macroscopic modeling and mitigation approaches. *Transportation Research Part B: Methodological*, 41(1):49–62.

Daganzo, C. F., Gayah, V. V., and Gonzales, E. J. (2012). The potential of parsimonious models for understanding large scale transportation systems and answering big picture questions. *EURO Journal on Transportation and Logistics*, 1(1-2):47–65.

De la Fuente, A. (2000). *Mathematical methods and models for economists*. Cambridge University Press.

Dong, H., Zhang, X., Dong, Y., Chen, C., and Rao, F. (2014). Recommend a profitable cruising route for taxi drivers. In *17th International IEEE Conference on Intelligent Transportation Systems (ITSC)*, pages 2003–2008. IEEE.

Douglas, G. W. (1972). Price regulation and optimal service standards: The taxicab industry. *Journal of Transport Economics and Policy*, pages 116–127.

Erhardt, G. D., Roy, S., Cooper, D., Sana, B., Chen, M., and Castiglione, J. (2019). Do transportation network companies decrease or increase congestion? *Science advances*, 5(5):eaau2670.

Fagnant, D. J. and Kockelman, K. (2015). Preparing a nation for autonomous vehicles: opportunities, barriers and policy recommendations. *Transportation Research Part A: Policy and Practice*, 77:167–181.

Fosgerau, M., Frejinger, E., and Karlstrom, A. (2013). A link based network route choice model with unrestricted choice set. *Transportation Research Part B: Methodological*, 56:70–80.

Gao, Y., Jiang, D., and Xu, Y. (2018). Optimize taxi driving strategies based on reinforcement learning. *International Journal of Geographical Information Science*, 32(8):1677–1696.

Ge, Y., Knittel, C. R., MacKenzie, D., and Zoepf, S. (2016). Racial and gender discrimination in transportation network companies. Technical Report No. w22776, National Bureau of Economic Research.

Geroliminis, N. (2015). Cruising-for-parking in congested cities with an mfd representation. *Economics of Transportation*, 4(3):156–165.

Geroliminis, N. and Daganzo, C. F. (2008). Existence of urban-scale macroscopic fundamental diagrams: Some experimental findings. *Transportation Research Part B: Methodological*, 42(9):759–770.

Greenshields, B., Channing, W., Miller, H., et al. (1935). A study of traffic capacity. In *Highway research board proceedings*, volume 1935. National Research Council (USA), Highway Research Board.

Greer, L., Fraser, J. L., Hicks, D., Mercer, M., Thompson, K., et al. (2018). Intelligent transportation systems benefits, costs, and lessons learned: 2018 update report. Technical report, United States. Dept. of Transportation. ITS Joint Program Office.

Hahn, R. and Metcalfe, R. (2017). The ridesharing revolution: economic survey and synthesis. In *More equal by design: economic design responses to inequality*, volume 4.

Hall, J. V. and Krueger, A. B. (2018). An analysis of the labor market for uber's driver-partners in the united states. *ILR Review*, 71(3):705–732.

Hall, M. A. (1978). Properties of the equilibrium state in transportation networks. *Transportation Science*, 12(3):208–216.

He, F. and Shen, Z.-J. M. (2015). Modeling taxi services with smartphone-based e-hailing applications. *Transportation Research Part C: Emerging Technologies*, 58:93–106.

He, F., Wang, X., Lin, X., and Tang, X. (2018). Pricing and penalty/compensation strategies of a taxi-hailing platform. *Transportation Research Part C: Emerging Technologies*, 86:263–279.

Hughes, R. and MacKenzie, D. (2016). Transportation network company wait times in greater seattle, and relationship to socioeconomic indicators. *Journal of Transport Geography*, 56:36–44.

Iqbal, M. (2019). Uber revenue and usage statistics. Available at Business of Apps: http://www.businessofapps.com/data/uber-statistics/. Accessed on June 20, 2019.

Johari, R. and Kumar, S. (2010). Congestible services and network effects. *EC*, 10:93–94.

Kashyap, B. R. (1966). The double-ended queue with bulk service and limited waiting space. *Operations Research*, 14(5):822–834.

Ke, J., Xiao, F., Yang, H., and Ye, J. (2019). Optimizing online matching for ride-sourcing services with multi-agent deep reinforcement learning. Available at arXiv preprint:1902.06228. Accessed on December 15, 2019.

Koopman, C., Mitchell, M., and Thierer, A. (2014). The sharing economy and consumer protection regulation: The case for policy change. *J. Bus. Entrepreneurship & L.*, 8:529.

Kooti, F., Grbovic, M., Aiello, L. M., Djuric, N., Radosavljevic, V., and Lerman, K. (2017). Analyzing uber's ride-sharing economy. In *Proceedings of the 26th International Conference on World Wide Web Companion*, pages 574–582. International World Wide Web Conferences Steering Committee.

Kornhauser, A., Chang, A., Clark, C., Gao, J., Korac, D., Lebowitz, B., and Swoboda, A. (2013). Uncongested mobility for all: New jersey's area-wide ataxi system. *Princeton University: Princeton, NJ, USA*.

Krueger, R., Rashidi, T. H., and Rose, J. M. (2016). Preferences for shared autonomous vehicles. *Transportation research part C: emerging technologies*, 69:343–355.

Lam, C. T. and Liu, M. (2017). Demand and consumer surplus in the on-demand economy: the case of ride sharing. *Social Science Electronic Publishing*, 17(8):376–388.

Li, Z., Hong, Y., and Zhang, Z. (2016). Do ride-sharing services affect traffic congestion? an empirical study of uber entry. *Soc. Sci. Res. Netw*, 2002:1–29.

Lin, K., Zhao, R., Xu, Z., and Zhou, J. (2018). Efficient large-scale fleet management via multi-agent deep reinforcement learning. In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pages 1774–1783.

Liu, S., Araujo, M., Brunskill, E., Rossetti, R., Barros, J., and Krishnan, R. (2013). Understanding sequential decisions via inverse reinforcement learning. In *2013 IEEE 14th International Conference on Mobile Data Management*, volume 1, pages 177–186. IEEE.

Liu, W. and Geroliminis, N. (2016). Modeling the morning commute for urban networks with cruising-for-parking: An mfd approach. *Transportation Research Part B: Methodological*, 93:470–494.

Lu, S. and Nie, Y. M. (2010). Stability of user-equilibrium route flow solutions for the traffic assignment problem. *Transportation Research Part B: Methodological*, 44(4):609–617.

MacKie-Mason, J. K. and Varian, H. R. (1995). Pricing congestible network resources. *IEEE journal on Selected Areas in Communications*, 13(7):1141–1149.

Mahmassani, H. S., Williams, J. C., and Herman, R. (1984). Investigation of network-level traffic flow relationships: some simulation results. *Transportation Research Record*, 971:121–130.

Manski, C. and Wright, J. (1976). The nature of equilibrium in the market for taxi services. *Transportation Research Record*, 619:11–15.

Martinez, L. and Crist, P. (2015). Urban mobility system upgrade: How shared self-driving cars could change city traffic. In *International Transport Forum, Paris*.

Matsushima, K. and Kobayashi, K. (2006). Endogenous market formation with matching externality: an implication for taxi spot markets. *Structural Change in Transportation and Communications in the Knowledge Economy*, pages 313–336.

Matsushima, K. and Kobayashi, K. (2010). Spatial equilibrium of taxi spot markets and social welfare. In *the 12th World Conference on Transport Research*, Lisbon, Portugal.

Newell, G. F. (1971). *Applications of queueing theory*. Chapman and Hall, London.

Nguyen, S. and Dupuis, C. (1984). An efficient method for computing traffic equilibria in networks with asymmetric transportation costs. *Transportation Science*, 18(2):185–202.

Nie, Y. M. (2017). How can the taxi industry survive the tide of ridesourcing? evidence from shenzhen, china. *Transportation Research Part C: Emerging Technologies*, 79:242–256.

Nourinejad, M. and Ramezani, M. (2020). Ride-sourcing modeling and pricing in non-equilibrium two-sided markets. *Transportation Research Part B: Methodological*, 132:340–357.

Qin, G., Li, T., Yu, B., Wang, Y., Huang, Z., and Sun, J. (2017). Mining factors affecting taxi drivers' incomes using gps trajectories. *Transportation Research Part C: Emerging Technologies*, 79:103–118.

Qu, M., Zhu, H., Liu, J., Liu, G., and Xiong, H. (2014). A cost-effective recommender system for taxi drivers. In *Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 45–54.

Ramezani, M. and Nourinejad, M. (2018). Dynamic modeling and control of taxi services in large-scale urban networks: A macroscopic approach. *Transportation Research Part C: Emerging Technologies*, 94:203–219.

Ranchordás, S. (2015). Does sharing mean caring: Regulating innovation in the sharing economy. *Minn. JL Sci. & Tech.*, 16:413.

Rayle, L., Dai, D., Chan, N., Cervero, R., and Shaheen, S. (2016). Just a better taxi? a survey-based comparison of taxis, transit, and ridesourcing services in san francisco. *Transport Policy*, 45:168–178.

Rong, H., Zhou, X., Yang, C., Shafiq, Z., and Liu, A. (2016). The rich and the poor: A markov decision process approach to optimizing taxi driver revenue efficiency. In *Proceedings of the 25th ACM International on Conference on Information and Knowledge Management*, pages 2329–2334.

Rust, J. (1987). Optimal replacement of gmc bus engines: An empirical model of harold zurcher. *Econometrica: Journal of the Econometric Society*, pages 999–1033.

Sahr, K., White, D., and Kimerling, A. J. (2003). Geodesic discrete global grid systems. *Cartography and Geographic Information Science*, 30(2):121–134.

Salanova, J. M., Estrada, M., Aifadopoulou, G., and Mitsakis, E. (2011). A review of the modeling of taxi services. *Procedia-Social and Behavioral Sciences*, 20:150–161.

Salanova, J. M., Romeu, M. E., and Amat, C. (2014). Aggregated modeling of urban taxi services. *Procedia-Social and Behavioral Sciences*, 160:352–361.

Schaller, B. (2017). Empty seats, full streets: Fixing manhattan's traffic problem. Technical Report 3, Schaller Consulting.

Schaller, B. (2018). The new automobility: Lyft, uber and the future of american cities. Technical report, Schaller Consulting.

Schroeter, J. R. (1983). A model of taxi service under fare structure and fleet size regulation. *The Bell Journal of Economics*, pages 81–96.

Shi, Y. and Lian, Z. (2016). Optimization and strategic behavior in a passenger–taxi service system. *European Journal of Operational Research*, 249(3):1024–1032.

Shou, Z., Di, X., Ye, J., Zhu, H., Zhang, H., and Hampshire, R. (2019). Optimal passenger-seeking policies on e-hailing platforms using markov decision process and imitation learning. Available at arXiv preprint:1905.09906. Accessed on December 15, 2019.

Shou, Z., Di, X., Ye, J., Zhu, H., Zhang, H., and Hampshire, R. (2020). Optimal passenger-seeking policies on e-hailing platforms using markov decision process and imitation learning. *Transportation Research Part C: Emerging Technologies*, 111:91–113.

Sloboden, J., Lewis, J., Alexiadis, V., Chiu, Y., and Nava, E. (2012). Traffic analysis toolbox volume xiv: Guidebook on the utilization of dynamic traffic assignment in modeling. federal highway administration. Technical report, Report FHWA-HOP-13-015, US Department of Transportation, Washington, DC.

Small, K. A. and Chu, X. (2003). Hypercongestion. *Journal of Transport Economics and Policy (JTEP)*, 37(3):319–352.

Small, K. A., Verhoef, E. T., and Lindsey, R. (2007). *The economics of urban transportation*. Routledge.

Smart, R., Rowe, B., Hawken, A., et al. (2015). Faster and cheaper: How ride-sourcing fills a gap in low-income los angeles neighborhoods. Technical report, BOTEC Analysis Corp.

Szeto, W. Y., Wong, R. C. P., Wong, S. C., and Yang, H. (2013). A time-dependent logit-based taxi customer-search model. *International Journal of Urban Sciences*, 17(2):184–198.

Tang, J., Wang, Y., Hao, W., Liu, F., Huang, H., and Wang, Y. (2019). A mixed path size logit-based taxi customer-search model considering spatio-temporal factors in route choice. *IEEE Transactions on Intelligent Transportation Systems*.

Taylor, T. A. (2018). On-demand service platforms. *Manufacturing & Service Operations Management*, 20(4):704–720.

Tchetgen, E. J. T., Walter, S., Vansteelandt, S., Martinussen, T., and Glymour, M. (2015). Instrumental variable estimation in a survival context. *Epidemiology (Cambridge, Mass.)*, 26(3):402.

Train, K. E. (2009). *Discrete choice methods with simulation*. Cambridge university press.

Urata, J., Xu, Z., Ke, J., Wu, G., Yin, Y., Yang, H., and Ye, J. (2020). Learning ride-sourcing drivers' customer-searching behavior: A dynamic discrete choice approach. *Working Paper, University of Michigan*.

Verma, T., Varakantham, P., Kraus, S., and Lau, H. C. (2017). Augmenting decisions of taxi drivers through reinforcement learning for improving revenues. In *Twenty-Seventh International Conference on Automated Planning and Scheduling*.

Wallsten, S. (2015). The competitive effects of the sharing economy: how is uber changing taxis. *Technology Policy Institute*, 22:1–21.

Wang, H. and Yang, H. (2019). Ridesourcing systems: A framework and review. *Transportation Research Part B: Methodological*, 129:122–155.

Wang, X., Yang, H., and Zhu, D. (2018). Driver-rider cost-sharing strategies and equilibria in a ridesharing program. *Transportation Science*, 52(4):868–881.

Watson, P. L. and Westin, R. B. (1975). Transferability of disaggregate mode choice models. *Regional Science and Urban Economics*, 5(2):227–249.

Wong, K., Wong, S., and Yang, H. (2001). Modeling urban taxi services in congested road networks with elastic demand. *Transportation Research Part B: Methodological*, 35(9):819–842.

Wong, K., Wong, S. C., Yang, H., and Wu, J. (2008). Modeling urban taxi services with multiple user classes and vehicle modes. *Transportation Research Part B: Methodological*, 42(10):985–1007.

Wong, R., Szeto, W., and Wong, S. (2014a). Bi-level decisions of vacant taxi drivers traveling towards taxi stands in customer-search: Modeling methodology and policy implications. *Transport Policy*, 33:73–81.

Wong, R., Szeto, W., and Wong, S. (2014b). A cell-based logit-opportunity taxi customer-search model. *Transportation Research Part C: Emerging Technologies*, 48:84–96.

Wong, R., Szeto, W., and Wong, S. (2015a). A two-stage approach to modeling vacant taxi movements. *Transportation Research Part C: Emerging Technologies*, 59:147–163.

Wong, R., Szeto, W., Wong, S., and Yang, H. (2014c). Modelling multi-period customer-searching behaviour of taxi drivers. *Transportmetrica B: Transport Dynamics*, 2(1):40–59.

Wong, R. C., Szeto, W., and Wong, S. (2015b). Sequential logit approach to modeling the customer-search decisions of taxi drivers. *Asian Transport Studies*, 3(4):398–415.

Xu, Z., Chen, Z., and Yin, Y. (2019a). Equilibrium analysis of urban traffic networks with ride-sourcing services. Available at SSRN: https://ssrn.com/abstract=3422294. Accessed on December 15, 2019.

Xu, Z., Li, Z., Guan, Q., Zhang, D., Li, Q., Ke, W., Nan, J., Liu, C., Bian, W., and Ye, J. (2018). Large-scale order dispatch in on-demand ride-hailing platforms: A learning and planning approach. In *KDD '18: The 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, August 19-23, 2018, London, United Kingdom*. ACM, New York, NY, USA, 9 pages.

Xu, Z., Yin, Y., and Ye, J. (2019b). On the supply curve of ride-hailing systems. *Transportation Research Part B: Methodological*.

Xu, Z., Yin, Y., and Ye, J. (2020). On the supply curve of ride-hailing systems. *Transportation Research Part B: Methodological*, 132.

Xu, Z., Yin, Y., and Zha, L. (2017). Optimal parking provision for ride-sourcing services. *Transportation Research Part B: Methodological*, 105:559–578.

Yang, H., Leung, C. W., Wong, S. C., and Bell, M. G. (2010). Equilibria of bilateral taxi–customer searching and meeting on networks. *Transportation Research Part B: Methodological*, 44(8-9):1067–1083.

Yang, H., Qin, X., Ke, J., and Ye, J. (2020a). Optimizing matching time interval and matching radius in on-demand ride-sourcing markets. *Transportation Research Part B: Methodological*, 131:84–105.

Yang, H., Shao, C., Wang, H., and Ye, J. (2020b). Integrated reward scheme and surge pricing in a ridesourcing market. *Transportation Research Part B: Methodological*, 134:126–142.

Yang, H. and Wong, S. (1998). A network model of urban taxi services. *Transportation Research Part B: Methodological*, 32(4):235–246.

Yang, H., Wong, S. C., and Wong, K. I. (2002). Demand-supply equilibrium of taxi services in a network under competition and regulation. *Transportation Research Part B: Methodological*, 36(9):799–819.

Yang, H. and Yang, T. (2011). Equilibrium properties of taxi markets with search frictions. *Transportation Research Part B: Methodological*, 45(4):696–713.

Yang, H., Ye, M., Tang, W. H., and Wong, S. C. (2005). Regulating taxi services in the presence of congestion externality. *Transportation Research Part A: Policy and Practice*, 39(1):17–40.

Yu, J. J., Tang, C. S., Max Shen, Z.-J., and Chen, X. M. (2019a). A balancing act of regulating on-demand ride services. *Management Science*.

Yu, X., Gao, S., Hu, X., and Park, H. (2019b). A markov decision process approach to vacant taxi routing with e-hailing. *Transportation Research Part B: Methodological*, 121:114–134.

Zangui, M., Yin, Y., Lawphongpanich, S., and Chen, S. (2013). Differentiated congestion pricing of urban transportation networks with vehicle-tracking technologies. *Transportation Research Part C: Emerging Technologies*, 36:434–445.

Zha, L., Yin, Y., and Du, Y. (2018a). Surge pricing and labor supply in the ride-sourcing market. *Transportation Research Part B: Methodological*, 117(PB):708–722.

Zha, L., Yin, Y., and Xu, Z. (2018b). Geometric matching and spatial pricing in ride-sourcing markets. *Transportation Research Part C: Emerging Technologies*, 92:58–75.

Zha, L., Yin, Y., and Yang, H. (2016). Economic analysis of ride-sourcing markets. *Transportation Research Part C: Emerging Technologies*, 71:249–266.

Zhang, D., He, T., Lin, S., Munir, S., and Stankovic, J. A. (2014). Online cruising mile reduction in large-scale taxicab networks. *IEEE Transactions on Parallel and Distributed Systems*, 26(11):3122–3135.

Zhang, K., Chen, H., Yao, S., Xu, L., Ge, J., Liu, X., and Nie, M. (2019). An efficiency paradox of uberization. Available at SSRN: https://ssrn.com/abstract=3462912. Accessed on December 15, 2019.

Zheng, Z., Rasouli, S., and Timmermans, H. (2018). Modeling taxi driver anticipatory behavior. *Computers, Environment and Urban Systems*, 69:133–141.