NORTHWESTERN UNIVERSITY

Analysis of ride-hail with pooling: Matching, equilibrium and management

A DISSERTATION

SUBMITTED TO THE GRADUATE SCHOOL
IN PARTIAL FULFILLMENT OF THE REQUIREMENTS

for the degree

DOCTOR OF PHILOSOPHY

Field of Transportation System Analysis & Planning

By

Kenan Zhang

EVANSTON, ILLINOIS

December 2021

Abstract

Analysis of ride-hail with pooling: Matching, equilibrium and management

Kenan Zhang

In the past decade, the e-hail service provided by transportation network companies (TNCs) gained popularity in major cities around the world. By allowing passengers to virtually "hail" vehicles on mobile phones, e-hail has revolutionized the matching process in ride-hail, drastically reducing the existing search friction. However, previous studies found the e-hail service may not only suffer from unexpected efficiency losses under certain market conditions, but also bring considerable pressure on already-congested streets in big cities. Partly motivated by these concerns, TNCs introduced pooling services to complement the regular e-hail service. Passengers who opt for pooling are paired in real-time and share a portion of their trips with others. Despite its great potential in theory, only a fraction of TNC passengers choose to pool. The objective of this dissertation is thus to understand (i) whether pooling could address the efficiency and sustainability concerns about e-hail service in current practice, (ii) under which conditions it could thrive in competition against other transportation modes, and (iii) how it should be better designed, operated and regulated in different market scenarios. To this end, this dissertation develops both analytical and numerical tools to investigate pooling in an urban transportation system with both ride-hail and other travel modes.

This dissertation is divided into four parts. Part 1 proposes a physical model that describes the interaction between passengers and vehicles in the matching process of a pooling ride. The

model supports the analysis of the main trade-off of pooling with respect to matching and serves as a building block of the following studies. Part 2 studies pooling in an aggregate ride-hail market, where passengers make mode choice between pooling and other travel modes, drivers make decisions on whether to join the market, and the service platforms determine the optimal pricing strategy. The analysis is first conducted for a market with a single platform and then extended to tackle the inter-platform competition. Part 3 is devoted to modeling a spatial ride-hail market with pooling. A stylized two-node model is first developed to investigate the congestion effect of ride-hail operations and evaluate different congestion mitigation schemes targeted at them. Then, a game-theoretic approach is proposed for modeling the dynamic routing of ride-hail vehicles in a spatial market. The model is first formulated as a non-cooperative game and then extended to accommodate cooperative routing. Finally, an agent-based ride-hail simulation is developed to support the analysis with more operational details. With the simulator, we demonstrate a metamodel-based simulation optimization approach that embeds an analytical model into the simulation and has it provides updating directions of design variables (e.g., the pricing scheme and matching interval) throughout the simulation. Empirical TNC and traffic data are utilized to construct numerical experiments throughout this dissertation in an attempt to demonstrate the theoretical findings and test the sensitivity of system performance under various market conditions.

Acknowledgements

First and foremost, I feel most indebted to my advisor, Professor Yu Marco Nie, for his continuing guidance, support and encouragement. I still remember the first email I sent to him, where I expressed my interest in his research and asked if he is hiring PhD students. At that time, I was rather confused with my future and not sure whether pursuing a PhD was the correct choice. I was not confident about myself either as I knew little about transportation. Hence, I did not expect his response at all, not to mention the encouraging words. During my PhD study, I have also been plagued with self-doubt multiple times. It was always Marco who backed me and ensured me that I was approaching the right direction. I am also grateful for the absolute freedom in research as well as the timely guidance from him. They enable me to have my research fully driven by interest meanwhile preventing me from being stuck in particular problems. Our heated discussions, sometimes even debates, on various topics further broadened my horizons and deepened my understanding of research. It is his unceasing curiosity about science and passion for research that motivates me to continue pursuing my career in academia.

Next, I want to express my sincere thanks to my committee member, Professor Hani Mahmassani, Professor Yafeng Yin and Professor Anton Braverman. Professor Mahmassani always brought up new perspectives to my research, Professor Yin shared many of his opinions and experience from his research on the same topic, and Professor Braverman guided me to a broader research field outside the transportation community. All three committee members gave a number of constructive comments and suggestions on accomplishing this dissertation. Besides, I want to thank Professor Samitha Samaranayake, Professor Joseph Chow and Professor Nikolas Geroliminis, as well as other

professors I met at conferences and in other encounters, who showed their interest in my work and provide valuable feedback.

I would like to thank the Mobility Research Group for their generous help during my internship at Ford. I really appreciate Mr. James Fishelson and Mr. Yifan Chen for their great efforts in pushing this project. It was my great pleasure working with Archak Mittal, Richard Twumasi-Boakye and Shadi Djvadian. The simulator I developed in collaboration with them becomes an important part of this dissertation. My internship at JD.com Silicon Valley Lab was another memorable adventure thanks to my manager Woody Wu, my mentor Ethan Zhang and other colleagues. It allowed me to stand at the intersection between academia and industry and work on real-world problems.

I would also like to acknowledge my cohort and fellow students at Northwestern, including Dr. Hongyu Chen, Dr. Sida Luo, Dr. Lin Zhong, Dr. Xueyan Wei, Dr. Hang Shu, Dr. Jingyuan Bao, Dr. Michael Hyland, Dr. Jiaxin Chen, Hongyu Zheng, Hongyuan Yang, Jiayang Li and Siwei Hu, for their kind help and support. I am equally grateful to have friends, including Dr. Wei Ma, Dr. Zhengtian Xu, Dr. Jintao Ke, Dr. Ruimin Ke and Dr. Zhaocai Liu, who pursue the same career path as me. Sharing my laughter and tears with them made me feel I was never alone on the way.

Finally, I feel most owed to my beloved family and Xiao. I would not have reached this point without your unconditional support and love. Although we have long been staying far away from each other, I continuously felt accompanied and deeply connected with you. Thank you for understanding every decision I made and always being on my side. It was my greatest fortune to have you in my life.

All the research presented in this dissertation is supported by the NSF Award CMMI 1922665. Besides, I want to thank Northwestern Transportation Center and the McCormick College of Engineering for the generous fellowships that ensure the accomplishment of this dissertation.

To my family and Xiao

Table of Contents

List of Figures

List of Tables

CHAPTER 1

Introduction

1.1. Background and motivation

Pooling is the action of grouping together resources to maximize benefits or minimize risk. In personal mobility, the idea of pooling creates transport modes like carpooling, carsharing and ridesharing [Chan and Shaheen, 2012, Furuhata et al., 2013]. The past decade has witnessed the rapid rise of ridesharing services provided by *transportation network companies* (TNCs) such as Uber, Lyft, Didi Chuxing and Grab. Growing with these TNCs is the debate about the nature and social impacts of their services. Since TNC drivers operate for profit rather than share their own trips with riders, such service is better described as *ridesourcing* or *e-hail* instead of ridesharing [Rayle et al., 2014, Nie, 2017] (for this reason, "e-hail" is used hereafter).

TNCs do provide certain kinds of ridesharing services (e.g., UberPool and LyftShared) that match, in real-time, passengers who can share a portion of their trips. As an incentive, pooling passengers pay a discounted fare while drivers receive a pickup fee for each additional pickup. Typically, pooling tends to increase trip distance and duration due to detours. Hence, the primary trade-off for passengers is between a lower trip fare and a longer travel time. As for TNC platforms, pooling helps increase the service capacity without expanding their vehicle fleets. Serving pooling trips could also generate profit for the platforms because drivers are paid primarily according to their *occupied time and distance*, independent of the number of passengers in-vehicle.[1] The cost for the platform, however, is the extra pickup fee and the vehicle time consumed in detours.

Prior to the emergence of TNCs, the ride-hail market has long been dominated by taxis. Hence, early studies mostly focus on taxis [e.g., Douglas, 1972, Beesley and Glaister, 1983, Arnott, 1996,

[1] See e.g., policies of Uber (`https://www.uber.com/us/en/drive/services/shared-rides/`) and Lyft (`https://help.lyft.com/hc/en-us/articles/115012926987-Shared-ride-driver-pay`).

Yang and Wong, 1998, Lagos, 2000]. They uncovered two salient features of ride-hail service: *matching friction* and *economies of scale*. Matching friction arises from the spatial distribution of demand and supply. It refers to the fact that the passengers cannot meet a vacant vehicle immediately after they enter the market. The friction also gives rise to a lion shared of the wait time, which is the primary measure of ride-hail's level of service (LOS). To achieve a desired LOS, a fraction of supply has to be "wasted" in the form of vacant taxis cruising on street in search for passengers. The economies of scale are related to the matching efficiency. In a taxi market, when waiting passengers and vacant vehicles both double in their numbers, the pickup number is likely to more than double. In other words, the trip production of taxis enjoys *increasing returns to scale* [e.g., Arnott, 1996, Yang et al., 2010].

Although e-hail has dramatically reduced the matching friction thanks to its advanced matching technologies [Azevedo and Weyl, 2016, Cramer and Krueger, 2016], there is growing evidence that it might hurt economies of scale [Nie, 2017, Castillo et al., 2018, Zhang et al., 2019a]. For instance, Nie [2017] and Shapiro [2018] both find e-hail loses competitive advantage over taxis in high-density areas (where the densities of passengers and vehicles are both high). Castillo et al. [2018] identify a catastrophic state named *Wild Goose Chase* (WGC), where the passenger wait time increases while the trip production decreases. The culprit for causing WGC is the pickup process, which could be exceedingly long during a demand surge. Since taxis do not have such a pickup phase, they are free of WGC. These issues have pushed TNCs to develop various incentive mechanisms to balance the demand and supply during peak hours and in high-density areas, such as surge pricing [e.g., Castillo et al., 2018, Guda and Subramanian, 2019, Besbes et al., 2021, Garg and Nazerzadeh, 2021]. Yet, the potential of pooling in mitigating the demand-supply imbalance—the fact that the demand for vehicles would simply be reduced to half if all trips are shared by two passengers—has not been fully exploited.

On the other side, empirical evidence of the negative impact of TNC operations on traffic congestion, particularly that in city centers, has been accumulating in recent years [e.g., Schaller,

2017a, Erhardt et al., 2019, Diao et al., 2021]. This has prompted city managers to aggressively pursue regulations on TNC operations. For instance, a popular policy is to charge a congestion fee on each TNC ride entering a dedicated region, while the charge for pooling trips is usually discounted.[2] However, it remains unclear how these policies would affect TNC operations and traffic conditions.

Several studies have demonstrated the great potential of pooling [e.g., Cici et al., 2014, Alonso-Mora et al., 2017, Tachet et al., 2017]. For instance, it is estimated more than 90% of taxi trips in New York City could be shared with a reasonable detour time [Santi et al., 2014]. However, we found most regions in Chicago have a pooling ratio well below 20% (see Figure 4.3 in Section 4.6). One naturally wonders what has contributed to the large gap between theory and reality, and what we could do to close it. Answering these questions is no easy task due to the complicated interactions among stakeholders (passengers, drivers and platforms) and their trade-offs. Although a great amount of research effort has been devoted to modeling and analyzing taxi and/or e-hail markets lately [e.g., Buchholz, 2019, Frechette et al., 2019, Castillo et al., 2018, Yan et al., 2019, Zha et al., 2016, 2018b, Xu et al., 2020], few have centered their investigations on pooling, as attempted in this dissertation.

On a grander scale, pooling is worth special attention because it aligns with the goal of building a sustainable transport system. A properly operated and regulated pooling service could improve the utility of both passengers and drivers, as well as reduce total vehicle miles traveled and carbon footprint of travel. Moreover, pooling is instrumental to the conception of mobility-as-a-service (MaaS), which strives to integrate multiple transport modes to provide users with a seamless travel experience. Even if two passengers have different origins and destinations, they might be able to share a portion of their trips before switching to another mode. Hence, pooling is likely to be implemented by MaaS operators in various forms to make the best use of their service capacities.

[2] See New York State's Congestion Surcharge (`https://www1.nyc.gov/site/tlc/about/congestion-surcharge.page`) and Chicago TNC congestion pricing (`https://www.chicago.gov/city/en/depts/bacp/supp_info/city_of_chicago_congestion_pricing.html`)

A better understanding of pooling in the context of a ride-hail market thus provides the foundation for analyzing other shared mobility service in the era of MaaS.

In summary, we aim to address three research questions in this dissertation:

- Whether and how does pooling address the efficiency issue of regular e-hail service?
- Why is the pooling ratio observed in practice far below the theoretical limit?
- How should policies be designed to encourage pooling and to mitigate the congestion effect of ride-hail services?

The answer for the first question is rooted in the matching process that is unique for pooling, while a prerequisite for the latter two is explicitly modeling the decision making process of all stakeholders in the market, including passengers, drivers and service operators. In addressing these questions, this dissertation provides a general framework for the analysis of ride-hail services, particularly pooling, in an urban transportation system.

1.2. Research framework

This dissertation studies pooling in a ride-hail market with other transport modes. This research problem is tackled from three aspects, namely, matching, equilibrium and management. The framework is outlined in Figure 1.1.

To differentiate pooling from other ride-hail modes, we first develop a physical model that describes the matching process of a pooling trip. Here, the main focus is not to develop a particular matching algorithm but to generalize the interaction between matching inputs (i.e., the number of waiting passengers and vacant vehicles) and outputs (e.g., expected passenger wait time). To this end, we extend the general matching model developed in Chen et al. [2018] and consider a ride-hail market with both regular e-hail trips (with single passenger and named *solo* trips hereafter) and pooling trips. The main research question here is how to characterize the impact of pooling on the matching process. On the one hand, pooling helps lessens passenger competition for vacant vehicles

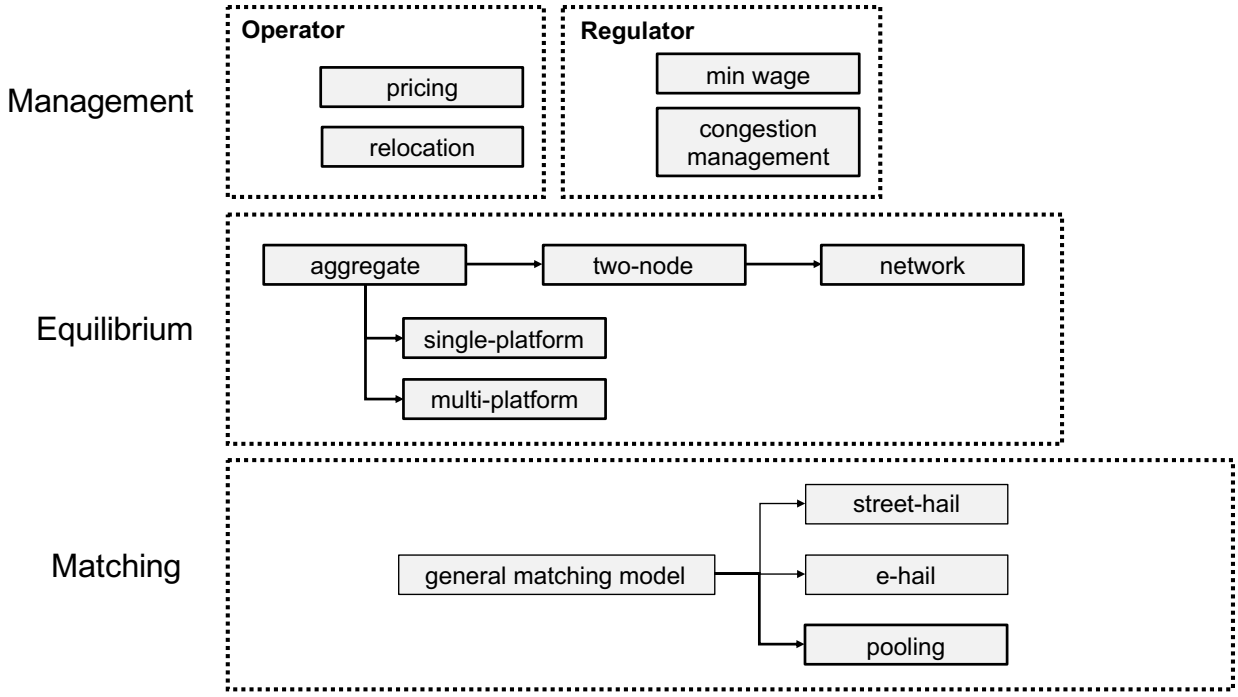


Figure 1.1. Research framework.

and thus the passenger wait time would reduce with the pooling demand. Further, pooling passengers enjoy higher competing power in the passenger-vehicle matching as their peers (who share trips with them) simultaneous search vehicles at a different location. On the other hand, pooling leads to additional detours in the pickup process, which extends the total passenger wait time. With the proposed matching model, we explicitly characterize the underlying physical mechanism of such a trade-off.

As discussed above, one main objective of this dissertation is to understand passengers' mode choice when pooling is introduced. Specifically, passengers choose between a full-priced solo ride and a discounted pooling ride with likely longer travel time, as well as alternative travel modes. We examine this trade-off using a market equilibrium model, where the matching model developed in the first step is integrated to express the interaction between passenger demand and vehicle supply. The equilibrium model is first established for an aggregate market with a single TNC platform, then extended to study the inter-platform competition and the congestion effect of ride-hail vehicles. As the analysis moves from an aggregate market to a spatial market, a new challenge

emerges, that is, modeling the movements of ride-hail vehicles across local markets. Assume idle drivers freely choose search locations to maximize their own payoff. Their collective behaviors, in turn, affect drivers' payoff through the likelihood of picking up passengers in each region (referred to as *meeting probability* hereafter). In other words, each driver needs to make search decisions based on all others' decisions. Accordingly, we formulate behaviors of ride-hail drivers as a dynamic routing game and derive the vehicle distribution at the equilibrium state. The matching model, again, plays an important role in this model as it differentiates the meeting probability function between particular ride-hail modes.

Using the market equilibrium models, we study a number of management problems. For the service operator, i.e., the TNC platform, the main question is whether or not to serve pooling and, if so, how to jointly price solo and pooling trips. To tackle this, we formulate the optimal pricing problem with different market structures and conditions, in which the equilibrium model serves as the constraints. Besides, the vehicle relocation is also briefly discussed in the spatial models. As for the regulator, it is still unclear how the regulations on TNCs would affect their operational strategies about the pooling service. In this dissertation, we examine the impact of minimum wage and various congestion mitigation schemes. This is done by integrating additional constraints into the platform's optimal pricing problem. Specifically, when the congestion effect of ride-hail vehicles is included, the analysis covers not only the stakeholders in the ride-hail market both also other travelers in the transportation system.

1.3. Contribution

From the modeling perspective, this dissertation is one of the few studies that explicitly model the matching process of pooling. The model developed in this work characterizes the matching process that is unique for a pooling ride. Assume each pooling ride is shared by two passengers. Then, the passenger wait time for a pooling ride could be divided into two parts, as illustrated in Figure 1.2. The first one is determined by the distance between the vehicle to the closer passenger ($\tilde{w}_{p1}$), while

the other depends on the distance between the passengers ($\tilde{w}_{p2}$). Accordingly, the wait time is a function of both the vehicle density and the pooling demand density. Besides, pooling enhances passengers' competing power for vacant vehicles. As shown in the second panel of Figure 1.2, the passenger on the left is not able to reach the vehicle given the search distance (equivalent to certain wait time). However, they are matched in a pooling ride because the vehicle falls in the peer passenger's search area. Such an advantage increases with the distance between pooling passengers. Yet, a larger distance also means a longer pickup detour ($\tilde{w}_{p2}$). Such a trade-off is thoroughly analyzed in this dissertation.

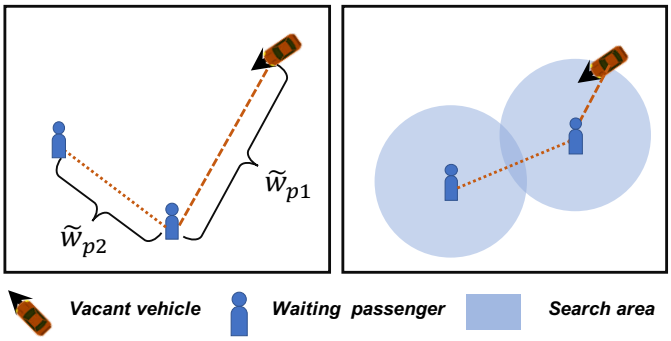


Figure 1.2. Matching in a pooling ride.

The market equilibrium models established based on the matching model also provide useful tools to study a wide array of problems regarding service operations and regulations. Several variants of equilibrium models are developed to support the analysis of platform competition and congestion externality. In all cases, we show the equilibrium can be reduced to a fixed point system and it always exists under mild conditions.

This dissertation also proposes a novel approach to modeling the search behaviors of ride-hail vehicles in a dynamic spatial ride-hail market. As introduced above, we formulate drivers' movements as a routing game, establish the equilibrium and propose iterative algorithm to approximate the equilibrium solution. We further extend the model to consider the cooperative routing, where drivers share the objective to maximize total reward of the system. In a nutshell, the modeling

framework is flexible to account for different ride-hail services, while, as a key component, the meeting probability needs to be specified for a given ride-hail mode. Therefore, the proposed model is capable of comparing different ride-hail services under the same market conditions.

Besides its contribution to the modeling, this dissertation also develops several numerical tools to facilitate the analysis of ride-hail services. One is a general solution algorithm for optimization problems with nonlinear and nonconvex equilibrium constraints. In this dissertation, it is used to solve the optimal pricing problem with the market equilibrium constraints, as well as regulatory constraints when applicable. The key idea is to differentiate the equilibrium constraints at the current solution using automatic differentiation [Baydin et al., 2017] and then use the equilibrium partials to construct gradient of the objective function.

Another numerical tool is an agent-based simulation of ride-hail service. Different from existing simulation frameworks, it is compatible with different ride-hail modes and passenger-vehicle matching algorithms. Besides, it supports mode choice of passengers as well as market entry decisions and strategic searching of drivers. It further provides an interface that allows some of the control variables (e.g., pricing) to be adjusted throughout the simulation. The simulator has been used to calibrate the analytical models proposed in this dissertation. With the control interface, we test a metamodel-based approach for simulation-based optimization [Barton and Meckesheimer, 2006], which embeds an analytical model into the simulation and asks it to guide the updating directions of control variables.

To reflect the issues in real practice, we construct all numerical experiments based on empirical TNC and traffic data. We also conduct extensive sensitivity analyses to test how the results vary against input variables and parameters. The findings derived from this dissertation offer a number of managerial insights for the operations and regulations of ride-hail service with pooling. Some of them have never been presented in the literature, such as the short-term and long-term impacts of minimum wage in a single-platform market, the market failure in the inter-platform competition due

to multi-homing drivers (drivers who join multiple platforms at the same time), and the efficiency loss of selfish routing of ride-hail drivers.

Besides the knowledge gained about pooling in a ride-hail market, this dissertation also sheds lights on the MaaS enterprise. The main problem tackled in this dissertation is how a mobility service provider makes the best use of its capacity. This is also the key question to be addressed in the era of MaaS, where passengers are not offered a single ride, but rather a mobility plan consisting of multiple transport modes. Hence, the MaaS operators need to better understand the underlying mechanisms of supply-demand interactions in order to optimize pricing, trip planning and capacity management. Models proposed in this work thus serve as an essential building block for these new problems.

1.4. Dissertation outline

The remainder of this dissertation is organized as follows. Chapter 2 gives a review of related studies on ride-hail service and pooling. The main focus there is matching, pricing, vehicle routing, platform competition and regulation in the context of market equilibrium as most of the analyses in this dissertation are based on a stationary state of the market.

Chapter 3 presents the matching model of pooling, which is extended from the physical model developed in Chen et al. [2018] and Zhang et al. [2019a]. The model characterizes the matching process in a market with both regular e-hail and pooling services, from which we derive the expected passenger wait time of both solo and pooling trips. A set of numerical experiments is then conducted to investigate of the sensitivity of passenger wait times towards the density of waiting passengers and vacant vehicles, as well as the pooling ratio.

Chapters 4 through 6 are dedicated to the analysis of an aggregate ride-hail market with pooling. Specifically, the market equilibrium is established in Chapter 4. We show that, regardless of the number of platforms in the market, the equilibrium is formulated as a system of equations and can be reduced to a fixed point system. Accordingly, the existence of equilibrium is proved by

evoking the fixed point theorem. We then construct a set of numerical experiments to investigate the system performance under various market conditions in both monopoly and duopoly scenarios.

Chapter 5 formulates the optimal pricing problem with the objective to maximize either the platform profit or social welfare. We discuss the optimal pricing strategies of a profit-maximizing platform and demonstrate the analytical results through numerical examples. Two regulations, namely, minimum wage and congestion tax, are also thoroughly studied in the numerical experiments. Chapter 6 continues to investigate the inter-platform competition. We define the equilibrium of the resulting pricing game, with and without regulatory constraints, and discuss its existence. Specifically, two supply modes are introduced and discussed both analytically and numerically. The first one is *single-homing*, which assumes each driver only joins one platform, and the other one is *multi-homing*, which assumes drivers join all platforms if they decide to enter the market. We further investigate the case of asymmetric platforms under the two supply modes in numerical experiments.

Starting from Chapter 7, our discussion moves to a spatial ride-hail market. A stylized two-node model is first developed to investigate the congestion effect of ride-hail vehicles. To characterize the congestion effect, we propose a simple traffic model where the traffic speed is endogenously determined the operation of ride-hail service. Similar to the aggregate mode, we first establish the market equilibrium and then formulate the optimal pricing problem with and without regulatory constraints. A case study is then constructed on the City of Chicago to test three congestion mitigate policies: (i) a trip-based fee charged on each solo trip starting or ending in CBD, (ii) a cordon-based fee charged on each vehicle entering CBD with one or no passenger, and (iii) a cruising cap that requires a minimum occupancy rate for all ride-hail vehicles in CBD.

Chapter 8 tackles the problem of ride-hail vehicle routing in a dynamic spatial market. The behavior of each driver is characterized by a Markov decision process (MDP), and, together, the collective routing is formulated as an MDP congestion game. We define the equilibrium state, show its equivalence to a fixed point and prove the existence via the fixed point theorem. The model

is further extended to consider cooperative routing, where drivers collaborate with each other to maximize a system objective. At the core of is the proposed model is the meeting probability, which gives the likelihood of picking up a passenger after searching in a zone for a time period. We derive its functional form for two ride-hail modes and calibrate them using simulation data. The model is then tested on both hypothetical and real networks to demonstrate the solution algorithm, compare the system outputs of different modes and evaluate the efficiency loss due to selfish routing.

Chapter 9 introduces an agent-based ride-hail simulation framework. The simulator is developed based on MATSim (Multi-Agent Transport Simulation) [Horni et al., 2016] while supporting multiple features that are not available in existing open-sourced simulation tools. Specifically, we develop an interface to adjust control variables throughout the simulation process. This feature is motivated by the idea to embed a design problem into a simulation framework. As the first step, we develop a pricing module that updates the pricing strategy according to the aggregate model developed in Chapter 5.

All numerical experiments conducted in this dissertation are based on the City of Chicago. Chapter 10 provides the description of data used in these experiments and outlines the estimation procedure of key parameters.

Finally, Chapter 11 summarizes this dissertation and provides some directions for future research.

CHAPTER 2

Literature Review

2.1. Market equilibrium

The ride-hail market has long been dominated by taxis. Hence, early research efforts have been devoted to modeling the taxi market. Two seminal works are accomplished by Douglas [1972] and Arnott [1996], which, respectively, characterize the interactions between passenger demand and vehicle supply in street-hail and radio-dispatch taxi market. Both studies, as well as many following works [e.g., De Vany, 1975, Beesley and Glaister, 1983, Cairns and Liston-Heyes, 1996], explicitly capture *matching friction*, the first salient feature of ride-hail service, in the relationship between passenger wait time and the vacant vehicle density. Since drivers are distributed over the space, the passenger cannot meet a vacant vehicle immediately after she enters the market. Therefore, to achieve a desired level of service (LOS), a portion of the vehicle supply must be "wasted" and transforms into the vacant vehicle density. Compared to street-hail, the radio-dispatch service also consumes its supply in the pickup process because a "matched" vehicle is no longer available for other passengers, even though it is still vacant [Zhang et al., 2019a]. This seemly minor difference becomes surprisingly consequential in the case of e-hail, a more advanced radio-dispatch system and the main focus of this thesis. Under certain market conditions, it results in an inefficient state called *Wild Goose Chase*, where a large portion of vehicles are stuck on the way to picking up passengers [Castillo et al., 2018]. Xu et al. [2020] further show that this inefficient state always exists in an e-hail market with finite match radius. Besides the discussion on matching, the aggregate equilibrium model of e-hail market has also been used to examine the pricing strategies on both sides of the market [e.g., Zha et al., 2016, Wang et al., 2016, Yang et al., 2020b, Dong et al., 2021].

As e-hail platforms, such as Uber and Lyft, launch their own on-demand pooling services, a few studies have been devoted to model the aggregate market equilibrium of this new type of service and compare it with regular e-hail mode. Ke et al. [2020a] models an aggregate e-hail market with a single platform serving either solo trips (with one passenger) or pooling trips (shared by two passengers). They formulate the optimal pricing problem under monopoly, social optimum and second-best (i.e., maximizing social welfare while maintaining a nonnegative platform profit) in both cases. With a simplified matching model and presumed detour function, they show a unit decrease in trip fare in a poling market attracts more passengers than that in a solo market and, under certain conditions, the optimal trip fare of pooling trips under monopoly, social optimum and second-bast are always lower than that of solo trips, respectively. In a subsequent study, Ke et al. [2020b] include a linear traffic model to investigate the congestion effect in pooling and solo market. It is found pooling could benefit both ride-hail passengers and background travelers and such a win-win situation depends on the design of matching window in pooling. While a longer matching time extends passengers' wait time, it also leads to pooling pairs with shorter detours and thus help mitigate the congestion externality. Similarly, Vignon et al. [2021] analyze the congestion effect of e-hail service through an aggregate market equilibrium model, though they consider the platform serves both solo and pooling trips and passengers make mode choice according to the total travel cost, including the trip fare and wait time. To simplify the modeling pooling service, they assume passengers in each pooling ride are picked up (dropped off) at the same location, thus there is no pickup (en-route) detour involved. Jacob and Roet-Green [2021] derive the extra travel of pooling trips using a queueing model and consider the optimal pricing problem in three scenarios: the platform offers only solo rides, only pooling rides, and both. In the last scenario, it is assume a group of high-type passengers only choose solo rides while the remaining low-type passengers only choose pooling rides. However, there is no modeling on the supply side and thus the passenger wait time determine by the vehicle supply is fixed regardless of the demand.

While the aggregate models aim to uncover the underlying rules of ride-hail services, the network models are developed to describe the spatiotemporal movements of vehicles and resulting network equilibrium. Research on this topic can be classified in two different ways. The first one is to divide them into centralized models [e.g., Pavone et al., 2012, Braverman et al., 2019] and decentralized models [e.g., Yang and Wong, 1998, Lagos, 2000], according to whether or not drivers behave at their own discretion. This will be further discussed in Section 2.4. Here, I will follow the second rule that splits the network models into static and dynamic. Yang and Wong [1998] develops the first static network model to analyze the spatial taxi market. The vehicle movements are described by vehicular flow across local markets and, at the equilibrium, the total vehicle time conserves and the inflow equals the outflow for each zone. In a subsequent study, an aggregate matching function, which takes the form of Cobb-Douglas production function [Cobb and Douglas, 1928], is introduced to determined the driver search time, along with the passenger wait time [Yang et al., 2010]. The same modeling framework has been applied to investigate a number of research problems of ride-hail service with a focus on both taxis and the e-hail service offered by TNCs [e.g., Yang and Yang, 2011, He and Shen, 2015, He et al., 2018]. In Xu et al. [2019], the assumption that passengers can only be matched with vehicles in the same zone is relaxed. Hence, in addition to occupied and vacant vehicle flows, there also exists pickup vehicle flow across zones. Another line of studies represents the ride-hail market as a network of stations. Hence, at a stationary state, the inflow of vehicles at each station must equal the outflow of occupied and relocating vehicles. This modeling framework is commonly used in the context of autonomous mobility-on-demand (AMoD) system, where vehicles are fully controlled by the platform [e.g., Pavone et al., 2012, Zhang and Pavone, 2016, Braverman et al., 2019].

In both directions of static models, the equilibrium constraint is based on vehicle flow conservation. In contrast, the dynamic models describe the market equilibrium in a different way. Lagos [2000] models the taxi movements in a spatial market and assume, at the equilibrium state, the return obtained by drivers in each zone must be equal thus the drivers' choice of next search

location is indifferent. In particular, the return includes the expected revenue in the current time period as well as in the future. The resulting dynamic equilibrium model is applied to analyze the taxi market in New York City [Lagos, 2003]. The similar framework has been used to investigate the spatial pricing problem of e-hail service [Bimpikis et al., 2019]. Buchholz [2019] also builds a dynamic spatial equilibrium model of taxi market and calibrates it with empirical taxi data of New York City. Instead of assuming all local market features equal return, the equilibrium is defined on a set of state variables. Specifically, at equilibrium, the passenger arrivals in each zone, vehicle flows across zones, as well as the drivers' search policy should be stationary. With a similar model, Shapiro [2018] compare the service efficiency of Uber and taxis in New York City. It is found Uber does not show great advantage in matching over taxis in dense markets, and thus its appeal in these areas are likely due to lower regulatory burden.

Besides the two major groups of network equilibrium models discussed above, a few recent studies extend the classic traffic assignment model Sheffi [1985] to evaluate the service operations and congestion effect of ride-hail services. Di and Ban [2019] consider three travel modes in a congested traffic network, namely, driving solo, ridesharing and e-hail. Accordingly, four types of traffic flows are characterized: personal vehicles flow, e-hail vehicle flow, ridesharing passenger flow and e-hail passenger flow. Each of them corresponds to a copy of the traffic network, which is then integrated into a super network. Ni et al. [2021] also distribute the passenger and vehicle flows on the network links, while the focus is the game among passengers, drivers and TNCs. At the equilibrium state, no player would change their decision and travel flows of both passengers and vehicles coincide on each link.

2.2. Matching

Centered at the equilibrium of a ride-hail market is the matching between waiting passengers and vacant vehicles. Result of this process determines the expected passenger wait time, a primary measure of level-of-service (LOS). Early studies introduce simple models to elaborate the relationship

between passenger wait time and vehicle supply. For instance, Douglas [1972] assumes the expected wait time for street-hail taxis is inversely proportional to the line density of vacant vehicles. While Arnott [1996] derives that the expected wait time for radio-dispatch taxis is inversely proportional to the spatial density of vacant vehicles. Results derived from this thesis align with these general rules, yet provide more detail physics in the matching process and further consider the influence from the demand side.

Due to the difficulty of modeling and calibrating the physical matching process, some studies simply assume it is frictionless, that is, the number of pickups equals the minimum between waiting passengers and vacant vehicles [e.g., Lagos, 2000, Bimpikis et al., 2019]. Although the frictionless matching assumption greatly simplifies the modeling, it sadly does not hold in practice. A traditional street-cruising taxi may not successfully meet a passenger on street even though they are in the same local market. Although in e-hail passengers can be matched with vehicles almost immediately, they still need to endure a pickup time, which could be extensively long during the demand peak [Castillo et al., 2018]. Using a large-scale TNC data, Xu et al. [2021] show the frictionless assumption may hold in the virtual matching in e-hail, but cannot generalize the entire matching process that also includes the pickup phase.

A common way to address the matching in ride-hail research is to introduce an aggregate matching function where the numbers of waiting passengers and vacant vehicles are the inputs while the pickup rate is the output. The most commonly used matching function is the Cobb-Douglas function, which draws analogy between matching process and production [e.g., Yang et al., 2010, Yang and Yang, 2011, He and Shen, 2015, Wang et al., 2016, Zha et al., 2016]. The urn-ball matching function is also used to model the matching as Bernoulli trials [e.g., Shapiro, 2018, Buchholz, 2019]. Relaxing the assumption that passengers can only be matched with vehicles in the same zone, Xu et al. [2019] develop a multi-output matching function by analogize the inter-zonal matching to electrical circuits. Additionally, some studies directly approximate the matching function through simulations [e.g., Frechette et al., 2019]. Another popular approach is to model

the matching process as a queueing system [e.g., Banerjee et al., 2015, Afeche et al., 2018, Zhang and Pavone, 2016, Banerjee et al., 2017, Braverman et al., 2019, Xu et al., 2020]. However, these models all assume passengers are picked up immediately after being matched with a vacant vehicle. The two exceptions are Besbes et al. [2018] and Feng et al. [2020], both of which extend $M/M/n$ queueing model by incorporating the pickup time that depends on the supply-demand relationship. Instead of characterizing the matching process, some studies consider matching as an operational strategy [e.g., Özkan and Ward, 2020, Hu and Zhou, 2021, Wang et al., 2019]. Most of these studies are also based on the queue model thus implicitly assume passengers are first-come-first-serve.

As detailed taxi operation data become wildly available, researcher start to calibrate the market equilibrium model with empirical data in order to conduct hypothetical analysis on taxi market. Frechette et al. [2019] and Buchholz [2019] both examine the impact of matching friction in the taxi market of New York City, envisioning the entry of TNCs. Frechette et al. [2019] finds the market segmentation between taxis and e-hail services could reduce the market thickness, in turn worsen matching frictions in the market. With a spatial equilibrium model, Buchholz [2019] further evaluates different dynamic pricing schemes and finds the location-based pricing achieves the most improvement of total surplus and matching efficiency. Following their works, Shapiro [2018] investigates the competition between taxi and e-hail using both taxi data and scraped Uber data in New York City, concluding that the advantage of e-hail in high-density areas mostly attributes to its lower price and less regulatory burden, rather the matching technology.

A few recent studies restart to consider the matching process in a physical space. Zha et al. [2018b] propose a geometric matching model to estimate the average matching and pickup time, though it still relies on the equilibrium conditions and a presumed matching function. Yang et al. [2020a] investigate the trade-off of matching interval in e-hail. To this end, they define the *dominant-zone* whose area depends on the length of matching interval and the demand level. By definition, there is only one passenger in each dominant-zone thus she will always be matched to the closest vacant vehicle. The matching model developed in this thesis is also rooted in the distance between

the waiting passenger and the closest vacant vehicle that is available for her. Differently, we consider the *matchable vehicles* for each passenger and derive their density in various ride-hail services and market conditions. Hence, our model is more general and can be used for comparison across different service modes.

To the best of our knowledge, on a few studies have been devoted to modeling the matching process of on-demand pooling. Yan et al. [2019] derive the matching time for a pooling ride using a queueing-based model. It assumes arriving pooling passengers wait in a matching queue for a certain time window. If a pooling passenger successes to find a peer, a vacant vehicle is dispatched to pick up both of them at a meeting point in middle of their trip origins. If not, the passenger would ride alone. The same setting is assumed in Jacob and Roet-Green [2021] and Ke et al. [2020a]. Instead of calibrated from empirical data as in Yan et al. [2019], Jacob and Roet-Green [2021] explicitly derive the probability of finding a match within a time window, while Jacob and Roet-Green [2021] assume it exponentially decreases with the number of passengers accumulated within a matching window. Since all three studies assume passengers are picked up at the same location, the pickup detour is simply ignored. However, in real practice, the matching time is usually at least an order of magnitude smaller than the pickup time. Meanwhile, most pooling services do pick up passengers in sequence. Therefore, it is important to characterize the pickup detour of pooling trips, which is the main focus of matching model proposed in this thesis.

2.3. Pricing

The pricing of e-hail platform has attracted a great amount of research interest in recent years. This problem is often tackled from two directions. One is to develop real-time pricing and/or price-aware dispatching algorithms [e.g., Asghari and Shahabi, 2018, Tong et al., 2018, Xu et al., 2018, Nourinejad and Ramezani, 2019, Chen et al., 2020, 2021]. The other, which is also the main focus of this thesis, is to evaluate the system performances under optimal pricing in the context of equilibrium analysis. One of the most heated research topics in this stream of research is *surge*

pricing, i.e., dynamically adjusting price in response to the demand surge [e.g., Cachon et al., 2017, Hu et al., 2018, Garg and Nazerzadeh, 2021]. Various benefits of surge pricing have been proposed in the literature. Castillo et al. [2018] show that surge pricing could effectively prevent the system from WGC. Banerjee et al. [2015] demonstrate dynamic pricing (a more general form of surge pricing) is more robust than static pricing. Besides, it is found surge pricing could be even useful in less profitable areas or regions with excessive supply because it incentivizes drivers to relocate [Besbes et al., 2021, Guda and Subramanian, 2019]. While the platform and drivers in general benefit from surge pricing, passengers have to bear the cost. Hence, regulations might be necessary to guide such practice, especially when a platform has de facto monopoly [Zha et al., 2018a]. Garg and Nazerzadeh [2021] further look into the structure of surge pricing and show that an additive surge induces more incentive compatible behaviors of drivers compared to multiplicative surge.

In the context of network equilibrium, pricing strategies are also designed to obtain a spatial balance between demand and supply. This is often known as *spatial pricing* [e.g., Zha et al., 2018b, Banerjee et al., 2017, Besbes et al., 2021]. Bimpikis et al. [2019] show that the platform profit and consumer surplus under the profit-maximizing pricing scheme are both maximized if the demand pattern satisfies certain "balance" condition, while the origin-based pricing scheme is more profitable when the demand pattern is unbalanced. Afifah and Guo [2020] model the spatial pricing problem as a Stakelberg game in a congested traffic network. The platform plays as the leader and sets the price to minimize the total supply-demand imbalance in the market, while passengers and drivers are the followers and respond to the pricing scheme with their travel and relocation decisions. Another commonly studied strategy to address the spatial demand-supply balance is vacant vehicle rebalancing [e.g., Pavone et al., 2012, Zhang and Pavone, 2016, Braverman et al., 2019], which will be discussed in more detail in Section 2.4. While pricing encourage drivers to relocate on their own initiative, rebalancing directly controls the vehicle flows. Visioning the

equivalence of spatial pricing and rebalancing, a few studies propose to jointly optimize the pricing and rebalancing strategies [e.g., Banerjee et al., 2017, Wollenstein-Betech et al., 2020]

The pricing of pooling service, however, still lacks in the literature. The studies on the market equilibrium of pooling service reviewed in Section 2.1 all formulate the platform's pricing problem and analyze the properties of optimal pricing strategies [Ke et al., 2020a,b, Vignon et al., 2021, Jacob and Roet-Green, 2021]. A common finding from these studies is the optimal trip fare of pooling rides is often lower than that of solo rides and it could be largely affected by the traffic condition. Besides, Lei et al. [2019] propose path-based dynamic pricing strategy that aims to promote pooling.

2.4. Vehicle routing and rebalancing

The literature on the routing problems of ride-hail vehicles may be broadly classified into individual and collective routing problems. In the first category, the main focus is either to learn a representative driver's routing strategy or to generate the optimal routing policy given a driver's current state. As taxi GPS trajectory data become widely available, a great amount of research effort has been dedicated to examine the search behavior of taxi drivers and to enhance their search efficiency [e.g., Liu et al., 2010, Li et al., 2011, Zhang et al., 2014]. Instead of manually defining search strategies, Zhang et al. [2019b] propose an image representation of taxi search behaviors and applied an unsupervised clustering algorithm to identify unique search strategies. Recognizing the nature of sequential decision making, Urata et al. [2021] model the search behaviors as a Markov decision process (MDP) with a nested discrete choice model and calibrate it with large-scale e-hail data. Similar to taxi drivers, e-hail drivers show a clear preference of relocating to high-demand areas, especially during the off-peak period. The MDP framework is also commonly used in the recommendation of optimal search path for ride-hail drivers [e.g., Hwang et al., 2015, Ge et al., 2010, Yuan et al., 2011, Qu et al., 2014, Shou et al., 2020, Yu et al., 2019b]. One notable benefit of this approach is the ability to anticipate and account for payoffs that may realize *later* due to

current decisions. However, a key assumption of these studies is that the environment would be invariant in application, that is, all other drivers would not change their search strategies. This assumption clearly no longer holds if a large group of drivers use follow the recommended paths.

The models of collective routing, on the other hand, typically simplify the movement of vehicles as continuous vehicular flow in a network, which ensures them to scale properly with the fleet size. These models could be further classified as centralized and decentralized. The centralized models assume vehicles are fully controlled by the operator and thus are often discussed in the context of the futuristic autonomous on-demand mobility (AMoD) systems. Pavone et al. [2012] first consider the fluids of passengers and vehicles across a network of stations and model the rebalancing as serving a group of "virtual" passengers. Accordingly, the centralized routing problem becomes finding an optimal rebalancing vehicular flow pattern. This model is then cast into a closed queuing network, where each station is modeled as a single-server node and the link between any two stations is modeled as an infinite-server node [Zhang and Pavone, 2016]. It turns out the optimal rebalancing problem formulated on the two models reduces to the same linear program. Braverman et al. [2019] prove the queuing network converges to a fluid limit as the numbers of passengers and vehicles both approach infinity, which approximately holds in large markets. In this case, the rebalancing problem can be reduced to a linear program, similar to those in the previous studies, and its optimal solution is shown to bound the performance of any state-dependent routing policy from above. Besides fleet management, the same modeling framework has been applied to examine the integration of AMoD system with public transit [Salazar et al., 2018, Wollenstein-Betech et al., 2021], the interaction between rebalancing and charging [Iglesias et al., 2019], and the joint decision of fleet management and charging stations [Estandia et al., 2021]. It is also used to design spatial pricing policies with the objective to address the demand-supply imbalance in ride-hail service [Waserhole and Jost, 2012, Banerjee et al., 2017]. All these models are built on the assumption that there is a single queue of passengers at each station and the passenger at the head would be picked up immediately upon the arrival of a vehicle. In practice, however, matching in ride-hail takes place at a much

larger geographical scale such that waiting is inevitable for passengers even if vacant vehicles are present in the same area. Few centralized models have considered such search friction in the spatial matching between passengers and vehicles. One of the exceptions is Ramezani and Nourinejad [2018], who propose a network-scale taxi dispatch model and utilize the Cobb-Douglas matching function [Cobb and Douglas, 1928, Yang et al., 2010] to predict the passenger pickup rate in each zone.

Different from the centralized model, decentralized vehicle routing models treat each driver as a strategic agent whose objective is to maximize their own utility. Hence, they often serve as the supply-side modeling in the equilibrium analysis of the spatial ride-hail market. These models can be further divided into static and dynamic models. Yang and Wong [1998] first develop a static network model for the spatial taxi market. The city is represented by a network of zones and the taxi movements are described as vehicular flows between them. The search behaviors of taxi drivers are characterized by a multinomial logit (MNL) model that takes the relocation time and expected search time into consideration. At equilibrium, not only does the vehicle inflow equal the outflow for each zone, but also the total vehicle time conserves. In a subsequent study [Yang et al., 2010], the Cobb-Douglas matching function is introduced to determine the passenger wait time and the driver search time in each zone, as in Ramezani and Nourinejad [2018]. The same modeling framework has been applied in a wide range of inquiries concerning both taxis and e-hail services [e.g., Yang and Yang, 2011, He and Shen, 2015, Wang et al., 2016]. Xu et al. [2019] adopt a similar framework but relax the assumption that passengers can only be matched with vacant vehicles in the same zone. Accordingly, a multi-output matching function is proposed to approximate the pickup rate in the inter-zonal matching.

The above static models assume drivers would stay in their selected search zone until picking up a passenger. Another line of research relaxes this assumption by incorporating the *meeting probability*, i.e., the probability that a driver successfully picks up a passenger after searching for a certain time period. The meeting probability dictates the expected reward of a search and,

accordingly, the value of each zone is computed as the expected reward plus the maximum future value from there. In Lagos [2000], it is further assumed all zones have the same value, which also equals the maximum future value, at the steady-state. This condition is weakened in Bimpikis et al. [2019], where the maximum value is only achieved in zones with positive vacant vehicle inflows. In both studies, the meeting probability is simply set to be the minimum between 1 and the ratio of supply to demand. In other words, they assume matching is frictionless, which again does not hold in practice [Castillo et al., 2018, Xu et al., 2021].

Dynamic decentralized routing models discretize the analysis time horizon into multiple periods and consider drivers making search decisions at the beginning/end of each time period. Buchholz [2019] proposes a dynamic spatial equilibrium model to study the taxi market in New York City. Drivers' search behavior is characterized by MNL model like Yang and Wong [1998] while the main factor comes to be the value of searching in each zone. The equilibrium is then defined by a fixed-point system of a set of variables, including the number of vacant vehicles and the meeting probability in each zone, the number of occupied vehicles between each origin-destination (OD) pair and search policies. With a similar model, Shapiro [2018] compare the service efficiency of Uber and taxi in New York City. In both studies, the meeting probability is approximated by a nonlinear function of passenger arrival rate and vacant vehicle number.

All decentralized models discussed above, either static or dynamic, explicitly capture the impact of collective vehicle routing on individual drivers' search decisions through the search time or the meeting probability. Yet, they do not recognize that drivers are playing a routing game against each other. In particular, each driver needs to optimize their search strategies in anticipation of all other drivers' strategies. Salhab et al. [2017] formulate the dynamic decentralized as a Mean Field Game [Lasry and Lions, 2007] and cast the equilibrium state as a fixed point of vacant vehicle distribution over the spatial market. Motivated by the vehicle routing in a spatial ride-hail market, Calderone and Sastry [2017] define the MDP routing game that extends classic routing game in that each agent's routing decision is characterized by an MDP rather than a deterministic path.

The resulting equilibrium is defined on the relocation vehicular flows between zones. The game is then extended to infinite horizon [Calderone and Shankar, 2017] and applied to study related problems [e.g., Li et al., 2019b,a].

To end this section, we note another approach for dynamic decentralized routing is to formulate it as a multi-agent reinforcement learning (MARL) problem and directly learn optimal search policies from simulations. The MARL framework has been applied to tackle the large-scale vehicle rebalancing problem [Lin et al., 2018, Yang et al., 2020c] and solve the lower-level problem in a bi-level optimization structure to design vehicle relocation incentives Shou and Di [2020]. However, as a model-free approach, MARL requires a large number of simulations to achieve the system equilibrium and obtain the corresponding search policy for a single set of market inputs. It is thus not an ideal tool to evaluate the system performance. Besides, as argued in [Xu et al., 2021], the search policies derived from these "black box" models could hardly be generalized to provide managerial insights.

2.5. Platform competition

The e-hail market is an instance of two-sided market in the transportation industry, where one or several platforms match service providers with service consumers meanwhile charge both sides [Rochet and Tirole, 2006]. Two-sided market is often praised for its *cross-side network effect*, that is, a larger number of user on one side of the market will attract users on the other side [Rysman, 2009, Tucker and Zhang, 2010]. However, a number of studies have identified the existence of *same-side congestion effect* that compromises the system efficiency [e.g., Belleflamme and Toulemonde, 2009, Aloui and Jebsi, 2011, Bernstein et al., 2019]. As will be shown in this thesis, such congestion effect is more substantial in e-hail because (i) passengers and vehicles are matched in real-time, and (ii) matched vehicles are no longer available for other passengers, even if they are close to each other.

Although previous studies have integrated the congestion effect into the analysis of platform competition, it is often modeled as a simple function of the user number on the same side (e.g., linear

in Aloui and Jebsi [2011]). One exception is Bai and Tang [2018], who model a duopoly market with symmetric platforms and single-homing drivers and represent the passenger wait time as a general function of demand and supply. They show that when the so-called *pooling effect* is present—i.e., wait time decreases when supply and demand both increases by one unit—both platforms break even at equilibrium. Only in the absence of the pooling effect would the platforms be profitable. In a sequel to Bai and Tang [2018], Wu et al. [2020] explore a sequential driver-passenger game, in which drivers and passengers make decisions sequentially. Unlike the standard simultaneous game, they show no sequential equilibrium exist such that the competing platforms would both possess a positive market share. Bernstein et al. [2019] compares the duopoly equilibria under single-homing and multi-homing in a general two-sided market, where the same-side network effect is described as a function of utilization rate. It is found that, even though individual drivers may prefer multi-homing, all players are worse off when all drivers are multi-homing. The aforementioned study by Zha et al. [2016] also analyzes a duopoly with single-homing drivers. Their main finding is that inter-platform competition does not necessarily lower the trip fare or improve social welfare. This finding is confirmed by Nikzad [2017], who adopts a different matching model and a general supply mode (i.e., drivers could be either single- or multi-homing). Nikzad [2017] further shows that the duopoly price would be higher than the monopoly price when the size of potential supply is small. Yet, the driver wage is always higher in the duopoly equilibrium.

Besides the network effects, the underlying supply mode also affects how platforms compete for users and service providers in a two-sided market. A central issue has to do with how "tightly" drivers are affiliated with a platform. When the affiliation is exclusive, it leads to the so-called single-homing supply mode; otherwise, drivers are allowed to be affiliated with multiple platforms, resulting in the multi-homing mode. This seemingly minor difference could result in distinctive pricing strategies [e.g., Rochet and Tirole, 2003, Armstrong, 2006, Böhme et al., 2010]. However, multi-homing drivers in the ride-hail service are slightly different from those considered in the existing literature. Once a driver accepts a ride request from one platform, they are no longer

available for all other platforms. Therefore, passengers from different platforms need to compete for the same pool of drivers. Based on my reading of the literature, there is still a lack of analysis on the consequence of mutli-homing in ride-hail market.

2.6. Regulation

Before the launch of Uber in 2009, the ride-hail industry in most cities around the world had been tightly regulated, in terms of both price and entry. The rationale of taxi regulations is largely based on the argument that taxi industry is a decreasing-average-cost industry [Douglas, 1972, Beesley and Glaister, 1983] and full competition is unlikely to maximize social welfare [Douglas, 1972, De Vany, 1975, Cairns and Liston-Heyes, 1996]. Besides controlling price and entry, it is also argued the taxi industry should be subsidized to achieve social optimum [e.g., Arnott, 1996]. While e-hail appears to share a similar cost structure with taxi [Zha et al., 2016], it has a radically different supply model. In particular, e-hail drivers are considered "independent contractors" who neither earn a fixed wage nor commit to a fixed work schedule. Therefore, the platform does not have a full control over its service capacity. Nor could the regulator easily manage the market entry through simple policy (e.g., issue of medallion).

The regulations currently under consideration focus on capping the number of operating drivers and setting a minimum "wage" to protect drivers [Joshi et al., 2019]. Gurvich et al. [2019] analyze the service capacity management problem for a service provider relying on a flexible supply model similar to that of e-hail. The authors argue that imposing a minimum wage will force the provider to restrict the number of agents on the platform during certain time periods. This implies that, in response to a minimum wage regulation, e-hail platforms may cap the number of drivers online when the supply is sufficient, effectively limiting their scheduling flexibility. A similar argument is given by Asadpour et al. [2019]. The authors show that, under certain market conditions, the platform cannot satisfy the required wage floor while maintaining a non-negative profit. As a result, it would respond to the regulation by either exiting the market, or reducing drivers' flexibility. However,

Parrott and Reich [2018] conclude that TNCs "could easily absorb an increase in driver pay with a minimal fare adjustment and little inconvenience to passengers". Specifically, their simulation results indicate the minimum wage policy proposed by the New York City will only lead to a relatively minor increase in passenger wait time. A recent study by Li et al. [2019c] shows that a cap on the number of vehicles benefits the platform but hurts drivers. On the other hand, imposing a minimum wage benefit both drivers and passengers because it pushes the platform to hire more drivers. Parrott and Reich [2018] conclude that TNCs "could easily absorb an increase in driver pay with a minimal fare adjustment and little inconvenience to passengers". Commission cap is another policy that has been discussed in literature. It is closely related minimum wage as it bounds the revenue taken by the platform from the drivers. Zha et al. [2016] show that commission cap is sufficient to guarantee a second-best outcome. This finding is further confirmed in Vignon et al. [2021], which incorporates both pooling and congestion externality. Yu et al. [2019a] analyze the welfare effects of the entry control policy in a market where e-hail and traditional taxi services compete for passengers. They conclude that there exists an optimal capacity cap that can best balance competing objectives of various stake holders. Our reading of literature above suggests no study has examined what partially motivates our study: the impact of regulations on an e-hail platform serving both solo and pooling rides.

Another set of regulations also under heated debates is congestion pricing. Although TNC claim e-hail service could help reduce car ownership, the evidence of its adverse traffic impact in already-congested city centers piles up [e.g., Schaller, 2017b, Erhardt et al., 2019, Diao et al., 2021]. As a result, several cities started to regulate TNC operations. For instance, New York City now charges \$2.75 on each TNC trip that passes through a designated congestion zone (south of 96th Street in Manhattan) with only one passenger.[1] The charge is reduced to \$0.75 if the trip is pooled (i.e., shared by at least two passengers). Similarly, Chicago charges \$3.00 for each solo trip (\$1.25 for each pooling trip) that starts or ends inside its downtown area[2]. In 2019, San Francisco

[1]See `https://www.tax.ny.gov/bus/cs/csidx.htm`

[2]See `https://www.chicago.gov/city/en/depts/bacp/supp_info/city_of_chicago_congestion_pricing.html`

enacted a TNC tax, which imposed a 3.25% surcharge on all solo trips and a 1.5% surcharge on all pooling trips[3]. Beside the trip-based fee, the conventional congestion pricing scheme, such as cordon-based pricing [see e.g. de Palma and Lindsey, 2011], is also under consideration. Using a spatial market equilibrium model, Li et al. [2020] compare three congestion pricing policies: (i) uni-directional cordon fee, (ii) bi-directional cordon fee, and (iii) trip-based fee. They find all three policies can reduce vacant vehicle density in the congested area and increase vehicle occupancy across the entire market. However, uni-directional cordon pricing also improves service quality outside the congested area and is more effective in reducing congestion. Recognizing the congestion impact of TNC vehicles is most severe in the city centers during peak hours, Schaller [2018] suggests a surcharge of $50/hour in Midtown and $20/hour in the other congested areas of Manhattan on ride-hail vehicles, which is expected to result in an 8% drop in the demand for ride-hail and an 30% decrease in VMT. Other than congestion pricing, it has been proposed to directly regulating the occupancy rate of TNC vehicles [see e.g., Schaller, 2017a]. Here, the target is the cruising time that is often seen as a counterproductive dead-weight loss from the vintage point of a city manager. Finally, recent research further discusses the potential of joint policy. For instance, Li et al. [2021] suggest imposing a congestion fee, either by trip or based on operation time, and combining it with a minimum wage policy. Their numerical experiments show that, although both pricing schemes reduce ride-hail demand, the time-based pricing can more effectively improve vehicle occupancy.

Among the few studies discussing a mixed e-hail service with both pooling and solo rides, only Vignon et al. [2021] investigate the impact of regulations. Specifically, it is the commission cap can only be imposed on solo rides to achieve the second-best, with however small congestion toll charged on drivers. Moreover, the toll can be replaced by congestion fees charged on solo and pooling rides at different rate.

[3]See `https://www.governing.com/news/headlines/San-Francisco-Will-Increase-Taxes-for-Ride-Hailing-Trips.html`

2.7. Simulation-based model for mobility service design

Compared to the analytical models reviewed above, the simulation-based models began to gain traction in ride-hail research more recently. [Nourinejad and Roorda, 2016] develop an agent-based simulation model to compare centralized and decentralized matching in dynamic ridesharing. [Djavadian and Chow, 2017] simulate the day-to-day adjustment process for both the demand side and the supply side of a mobility-as-a-service market. The results suggest the process is able to converge to a stochastic user equilibrium (SUE) Smith [1984], Horowitz [1984]. [Ruch et al., 2018] propose AMoDeus, a simulation model of autonomous mobility-on-demand (AMoD) systems that provides a wide range of computational tools. [Hyland and Mahmassani, 2018] propose several matching/dispatching algorithms for AMoD systems and evaluate them using agent-based simulation. Built on SimMobility Adnan et al. [2016], [Nahmias-Biran et al., 2019] attempt to integrate different types of on-demand mobility services within a single simulation framework. [Shen et al., 2018] develop an agent-based simulation model of a hybrid system that integrates shared fist/last-mile AMoD service with traditional transit. [Beojone and Geroliminis, 2021] analyze the impact of cruising TNC vehicles on traffic congestion through agent-based simulation, in which the travel time of a trip is rendered from a macroscopic fundamental diagram (MFD) Geroliminis and Daganzo [2008].

Simulation-based models are able to represent the behaviors of passengers, drivers and the operator, as well as their interactions in detail. However, they are not well equipped to solve service design problems such as pricing and fleet size. The main reason is the gradient, i.e., response of the simulated system towards a variation in the design variable, is often too complex to compute. A commonly adopted remedy to this difficulty is to introduce a bi-level structure: the upper level deals with the design question and the lower level simulates the system. One example is Pinto et al. [2020], who tackle the design of an integrated transit–AMoD system. However, this framework often limits the interaction between the design model and the simulator—only the design variables and

selected system performance metrics are exchanged. As a result, the design model still could not get full information about the gradient.

A related but more general research topic is simulation-based optimization (SBO), which integrates microscopic simulations into a stochastic optimization problem. The solution approaches for SBO can be broadly classified into two groups. When the feasible set of decision variables is a finite set, the problem could be solved through random search and metaheuristics, while if it is continuous, then gradient-based and metamodel-based methods are used [Barton and Meckesheimer, 2006]. Specifically, the metamodel approach can be further classified into physical and functional ones. While the former requests a physical representation of the underlying problem, the latter is generic and does not require any information about the objective function or the underlying problem. Therefore, the functional metamodels are more often selected for SBO problems. So as those in transportation research [e.g., Chen et al., 2016, He et al., 2017, Cheng et al., 2019]. Nevertheless, due to the lack of information about the specific problem, this approach often fail to provide . To deal with this issue, a number of studies propose to include an auxiliary physical component into the metamodel, which provides a global approximation of the objective function [e.g., Osorio and Bierlaire, 2013, Chong and Osorio, 2018].

Another group of studies tackles the service design problem through reinforcement learning, where the optimal policies are directly learned from simulation without solving any optimization problem or explicitly modeling the system dynamics Holler et al. [2019], Shou and Di [2020], Chen et al. [2021]. While this approach could obtain a desirable performance in practice, it fails to generate insights into the ride-hail operation due to a lack of understanding of the market. Thus, it is not suitable for the strategic planning of ride-hail service.

Part 1

Matching with pooling

CHAPTER 3

Matching model of pooling

Throughout this thesis, we define a *pooling ride* as a trip shared by two passengers. It starts at the moment an idle driver is dispatched to pick up the passengers, one at a time, and ends when both passengers arrive at their destinations. As its counterpart, a *solo ride* refers to a regular e-hail trip with only one passenger. For both solo and pooling rides, the driver cannot be assigned to another ride unless he finishes the current one.

To simplify notations, we name the passengers who choose solo rides as *solo passengers*, and those choosing pooling as *pooling passengers*. If the platform does not offer pooling rides, all passengers must choose solo rides. We name passengers in such a case as *e-hail passengers* to distinguish them from solo passengers who have the option of pooling. Accordingly, we use subscripts s, p and e to denote variables associated with solo, pooling and e-hail rides, respectively.

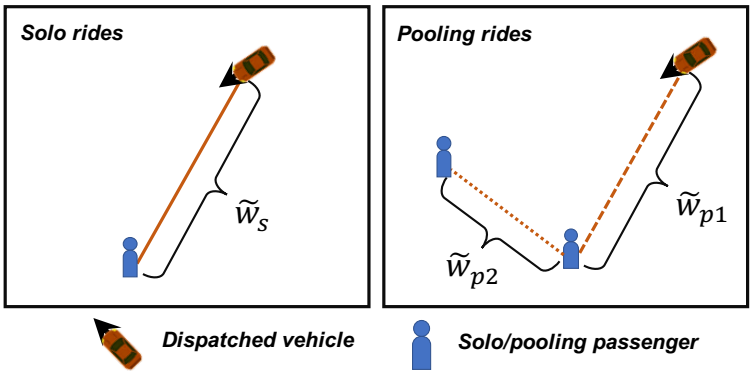


Figure 3.1. Pickup process of solo and pooling rides.

As shown in Figure 3.1, the wait time for a solo ride, denoted as $\tilde{w}_s$, depends on the distance between the passenger and the vehicle assigned to her. For a pooling ride, the wait time consists of two parts. The first part, denoted as $\tilde{w}_{p1}$, is the *pickup time* for the passenger closer to the matched vehicle. The second part, denoted as $\tilde{w}_{p2}$, is the *detour time* consumed in picking up the other

passenger, determined by the distance between the two passengers sharing the trip, as illustrated in Figure 3.1. Notations used in this chapter are listed in Table 3.2.

Before presenting the matching model, we wish to emphasize that the objective here is to represent the relationship between macroscopic variables that can be meaningfully measured and predicted in a highly idealized macroscopic market, such as average wait time, and the density of passengers/vacant vehicles. As a compromise, most details of the matching process, including matching time (i.e., the duration from the moment when the platform receives a request to the moment when it assigns a vehicle to serve the ride), matching criteria (e.g., distance, shareability), and dynamic pooling strategies, are left out. Because these details vary strongly across time, space and platform, it is difficult to represent them explicitly and satisfactorily by the equilibrium state of our model. In addition, their effects on wait time and detour time, which are the most important attributes of the ride-hail service, appear to be of secondary importance. For example, the matching time is usually at least an order of magnitude smaller than the pickup time [Zha et al., 2018b], whether vehicles are assigned to requests immediately [Castillo et al., 2018] or in a batching process [Yang et al., 2020a]. The only exception is when there is a severe supply-demand imbalance and passengers must wait to be matched in a virtual queue. However, such an extreme case need not concern us because (i) the focus here is on a relatively long-term market equilibrium; and (ii) transit would serve as a fallback option. Although we do not model the detailed matching process, we shall implicitly capture its aggregate effect on wait time and detour time by introducing exogenous parameters that can be calibrated from empirical data.

3.1. A general matching model

In this thesis, we adopt the matching model developed in Chen et al. [2018] and extend it to accommodate the matching process in pooling. Consider a passenger entering the ride-hail market. Their wait time depends on the number of vacant vehicles as well as the number of other waiting passengers nearby. Here, a vehicle is considered vacant if there is no passenger in it. Therefore, a

fraction of vacant vehicles may have been dispatched and are en-route to pick up their passengers. We name vehicles in this stage as *matched vehicles*. Accordingly, the idle vehicles are referred to *unmatched vehicles*. It is expected there are very few matched vehicles in a street-hail taxi market, whereas a substantial portion of vehicles could be matched in an e-hail market. Similarly, we split waiting passengers into two groups: (i) the matched, who are waiting to be picked up, and (ii) the unmatched, who are waiting to be matched. Since the matching time is in general much shorter than the pickup time, the majority of the waiting passengers should be matched. To simplify the analysis, the following assumptions are introduced [Chen et al., 2018].

Assumption 1. *Vacant vehicles and waiting passengers, both matched and unmatched, are uniformly distributed. In addition,*

(1) *all vehicles are cruising at the same speed v, and*

(2) *passengers keep waiting at the same location before pickup.*

With Assumption 1, the spatial distribution of vacant vehicles and waiting passenger are represented by their respective densities Λ and Π. Accordingly, the densities of unmatched vehicles (passengers) is given by $\Lambda_0 = b_\Lambda \Lambda$ ($\Pi_0 = b_\Pi \Pi$). Chen et al. [2018] prove that the number of unmatched vehicles, denoted by $\tilde{N}_v(d)$, within a distance d from the passenger forms an Inhomogeneous Poisson Process with intensity $\xi_v(d) = 2\pi d\Lambda_0$. The proof is provided in Section 3.6.2. The key insight of the model is that not all unmatched vehicles are available for the passenger. Instead, the fraction of *matchable vehicles*, denoted by $\gamma(d)$, is determine by matching mechanism. Assume the matchable vehicles are randomly distributed among the unmatched vehicles, then the subprocess, denoted by $\tilde{N}_{mv}(d)$, is also an Inhomogeneous Poisson Process with intensity $\xi_{mv}(d) = 2\pi d\gamma(d)\Lambda_0$.

Suppose the passenger is picked up by the closest matchable vehicle at distance $\tilde{D}$, the probability that at least one matchable vehicle is within d from the passenger is given by

$$F_{\tilde{D}}(d) = Pr(\tilde{D} \leq d) \tag{3.1}$$

$$= 1 - Pr\left(\tilde{N}_{mv}(d) = 0\right) = 1 - \exp\left(-\int_0^d 2\pi u\gamma(u)\Lambda_0 \mathrm{d}u\right).$$

3.1.1. Street-hail taxi

For street-hail taxis, the equality $\Lambda_{0,\mathrm{tx}} = \Lambda_{\mathrm{tx}}$ generally holds, where the subscript "tx" denotes variables associated with taxi service. Chen et al. [2018] further derive the fraction of matchable vehicles as $\gamma_{\mathrm{tx}}(d) = (\sigma d_0)/(2\pi d)$, where σ evaluate the attractiveness of the passenger's waiting location towards nearby taxis and d_0 is the *hail radius*, that is, the maximum distance from which the passenger could see and hail the taxi. As a result, $N_{mv,\mathrm{tx}}$ reduces to a Poisson Process and the cumulative distribution function (CDF) of $\tilde{D}_{\mathrm{tx}}$ becomes

$$F_{\tilde{D}_{\mathrm{tx}}}(d) = 1 - \exp\left(-\int_0^d \sigma d_0 \Lambda_{\mathrm{tx}} \mathrm{d}u\right) = 1 - \exp\left(-\sigma d_0 \Lambda_{\mathrm{tx}} d\right), \tag{3.2}$$

which yields the expected wait time

$$w_{\mathrm{tx}} = \frac{\delta}{v} D_{\mathrm{tx}} = \frac{\delta}{v\sigma d_0 \Lambda_{\mathrm{tx}}}, \tag{3.3}$$

where δ is the detour ratio of road network [e.g., Fairthorne, 1964, Boscoe et al., 2012, Yang et al., 2018].

3.1.2. Regular e-hail service

E-hail enables the real-time matching between passengers and drivers far from each other. Conceptually, it is equivalent to push d_0 in Eq. (3.2) to infinity. However, the ability of reaching every unmatched vehicle also creates the competition among unmatched waiting passengers. As a result, the fraction of matchable vehicles in the context of e-hail is no longer restricted by the hail distance, but the competition from fellow passengers.

If the unmatched vehicles are evenly distributed among unmatched passengers, the fraction of matchable vehicles is approximated by $\gamma_e(d) = 1/\Pi_0$. Accordingly, the intensity of $N_{mv,\mathrm{tx}}$ becomes $\xi_{mv,e}(d) = 2\pi d\Lambda_0/\Pi_0$, where ratio between unmatched vehicle and passenger densities can be

further simplified by

$$\frac{\Lambda_0}{\Pi_0} = \frac{b_\Lambda \Lambda}{b_\Pi \Pi}. \tag{3.4}$$

Since b_Λ and b_Π reflect on the supply and demand at different stages of the matching process. To simplify the analysis, the following assumption is also introduced [Zhang et al., 2019a],

Assumption 2. *Through its matching algorithm, the platform can achieve a stable ratio between b_Λ and b_Π, defined as $k := b_\Lambda / b_\Pi$. k is a parameter that measures the matching efficiency. A larger k indicates a higher efficiency.*

Accordingly, CDF of $\tilde{D}_e$ is derived as

$$F_{\tilde{D}_e}(d) = 1 - \exp\left(-\int_0^d \frac{2\pi u k \Lambda}{\Pi} \mathrm{d}u\right) = 1 - \exp\left(-\frac{\pi k \Lambda}{\Pi} d^2\right), \tag{3.5}$$

and the expected wait time for e-hail is

$$w_e = \frac{\delta}{2v}\sqrt{\frac{\Pi}{k\Lambda}}. \tag{3.6}$$

The difference in matching mechanism between taxi and e-hail is easily understood when comparing the intensities of their matchable vehicle processes. While $\xi_{mv,\mathrm{tx}}(d) = \sigma d \Lambda_{\mathrm{tx}}$ is independent of d, $\xi_{mv,e}(d) = 2\pi d k \Lambda_e / \Pi_e$ grows with d. In other words, the matching rate with respect to the wait time is constant for taxi passengers while it increase linearly for e-hail passenger. The benefit from this increasing matching rate, however, is checked by the competition from other waiting passengers. Factor Π_e in $\xi_{mv,e}(d)$ indicates that the unlimited matching radius essentially induces a congestion on the demand side. As demonstrated in Zhang et al. [2019a], this congestion effect is surprisingly consequential such that it even alters the overall economic property of e-hail service.

In real practice, however, the fraction of matchable vehicles depend on specific matching strategy employed by the platform. Sometimes, it is even beneficial to hold certain number of unmatched passenger to improve the overall matching efficiency [e.g., Özkan and Ward, 2020, Hu and Zhou,

2021]. Hence, we let $\gamma_e = 1/\Pi_0^{\varepsilon}$ to capture such effect. Accordingly, the matching efficiency and the intensity become $k = b_\Lambda/(b_\Pi)^{\varepsilon}$ and $2\pi dk\Lambda/(\Pi)^{\varepsilon}$, respectively. For simplicity, we do not consider this adjustment in the equilibrium analysis.

It is worth noting that Eqs. (3.3) and (3.6) align with the wait time function derived in Douglas [1972] and Arnott [1996], yet they capture more characteristics in the matching process (e.g., hail distance, local hotspot effect and passenger competition). Hence, the model can be calibrated using empirical data to evaluate and compare the matching efficiencies across local markets.

3.2. Matching in e-hail service with pooling

To accommodate the pooling service with the model presented above, we first add more assumptions about the distribution of passenger densities.

Assumption 3. *Passengers waiting for solo rides and pooling rides are uniformly distributed with densities* Π_s *and* Π_p*. Among those waiting for pooling rides, the unmatched passengers are uniformly distributed with a density* $\Pi_{0,p} = b_{\Pi_p}\Pi_p$*. Besides, the origin and destination of passengers are uniformly distributed in the market.*

Similar to b_Λ and b_Π, b_{Π_p} is related to the efficiency of pairing pooling passengers. To simplify the notation, we use b instead hereafter.

Recall the wait time of solo passengers is $\tilde{w}_s$ and that of the pooling passengers consists of two parts, i.e., $\tilde{w}_{p1}$ and $\tilde{w}_{p2}$. We first discuss the detour time $\tilde{w}_{p2}$ and solo wait time w_s, as they are easily derived by drawing an analogy from the matching in regular e-hail service. Then, we proceed to discuss $\tilde{w}_{p1}$, which has a more complicated analytical form.

3.2.1. Expected pickup detour in pooling

The matching between pooling passengers is analogous to that between the passenger and vehicles. Define *matchable passengers* as *other* unmatched pooling passengers. Then, follow the exactly the same reasoning of Proposition 1, it is shown that the number of matchable passengers within

distance l from the passenger, denoted by $\tilde{N}_{mp}(l)$, forms an Inhomogeneous Poisson Process with intensity $2\pi l\Pi_{0,p}$. Again, assume the passenger be matched with the

Denote $\tilde{N}_{mp}(l)$ as the number of *matchable passengers*, that is, *other* unmatched pooling passengers, within a distance l from a passenger. With Assumptions 1 and 3, and following exactly the same reasoning used in proving Proposition 1, we can show $\tilde{N}_{mp}(l)$ is an Inhomogeneous Poisson process with intensity function $2\pi l\Pi_{0,p} = 2\pi lb\Pi_p$. Again, assume the passenger be matched with the closest matchable passenger at distance $\tilde{L}$. Then, the expected values of $\tilde{L}$ and $\tilde{w}_{p2}$ are derived by, respectively,

$$L = \frac{1}{2\sqrt{b\Pi_p}}; \quad w_{p2} = \frac{\delta L}{v} = \frac{\delta}{2v\sqrt{b\Pi_p}}. \tag{3.7}$$

3.2.2. Expected wait time for solo rides

Similar to the case of regular e-hail service, the solo wait time is determined by the distance to the closest matchable vehicle. However, the fraction of matchable vehicles has changed due to the existence of pooling rides. It is expected that the competition among passengers is less severe because pooling passengers requires fewer vehicles. In the extreme case, when everyone is pooling, the total number of vehicles required will be cut in half. Therefore, we define *effective density of waiting passenger* as

$$\Pi_{\text{eff}} \approx \Pi_s + \Pi_p/2. \tag{3.8}$$

Substituting the waiting passenger density in Eq. (3.6) with above definition gives the expected value of solo wait time

$$w_s = \frac{\delta}{2v}\sqrt{\frac{\Pi_{\text{eff}}}{k\Lambda}} = \frac{\delta}{2v}\sqrt{\frac{\Pi_s + \Pi_p/2}{k\Lambda}}. \tag{3.9}$$

3.2.3. Expected pickup time in pooling

Different from solo passengers, a pair of pooling passengers compete for unmatched vehicles as a team. To better illustrate the discrepancy, we define *search area* as the area enclosed by a circle of radius d and centered at the passenger's location. As shown in Figure 3.2, the pooling passengers have a greater access to unmatched vehicles given the same search area as the solo passenger. As a result, the solo passenger in the left panel fails to find a match, whereas she would succeed if choosing to pool as there is a vehicle falls in their peer's search area. In other words, pooling passengers enjoy *collective competing power* in the market, which implies $w_{p1} \leq w_s$ in general. Nevertheless, this advantage diminishes when the pooling passengers get closer to each other and their search areas begin to overlap. Evidently, if the two passengers happen to wait at the same location, their competing advantage will be wiped out.

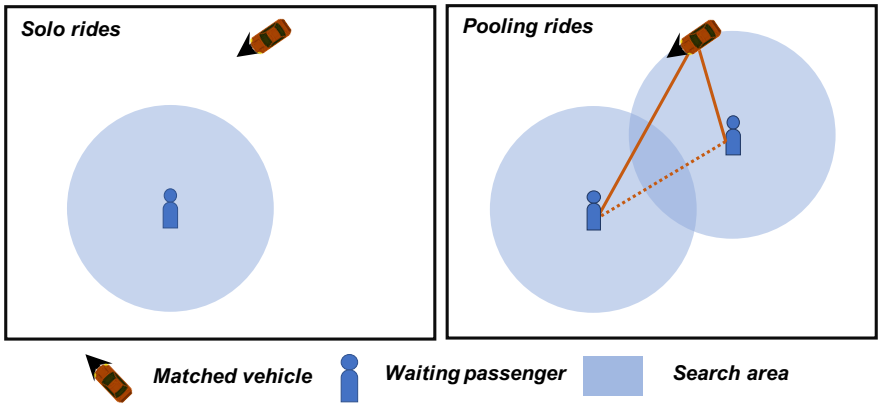


Figure 3.2. Access to unmatched vehicles through search area for solo passengers (Left) vs. pooling passengers (Right).

Taking the above observation into consideration, we derive the following closed-form formula for w_{p1}, which includes some mile approximation and will be explained in detail in Section 3.6.3.

$$w_{p1} \simeq \frac{\delta}{2v}\sqrt{\frac{\Pi_s + \Pi_p/2}{k\Lambda}\frac{\kappa + 4b\Pi_p}{2\kappa + 4b\Pi_p}}, \tag{3.10}$$

where κ is a parameter introduced to approximate the overlapping portions of the pooling passengers' search areas.

3.3. Discussions

3.3.1. En-route pooling

A key assumption about pooling is that passengers are always pooled together before their trips start. In reality, e-hail platforms often allow more flexibility. For example, they may let a pooling passenger leave for her destination without a partner, and attempt to find one en-route. In what follows, we consider the case when each pooling passenger is matched to the closest matchable vehicle, which may be fully or partially vacant. If the passenger rides alone, she could be detoured to pick up another pooling passenger en-route.

Let $\tilde{D}_0$ and $\tilde{D}_h$ be the distance between the passenger and the closest matchable vehicles that are, respectively, fully and partially vacant. Then, the expected pickup time and detour to pick up a second passenger may be specified as

$$w_{p1} = \frac{\delta}{v}\left(E[\tilde{D}_0]Pr(\tilde{D}_0 < \tilde{D}_h) + E[\tilde{D}_h]Pr(\tilde{D}_0 \geq \tilde{D}_h)\right); \tag{3.11}$$

$$w_{p2} = \frac{\delta}{v}E[\tilde{D}_h]\left(1 - Pr^{\beta}(\tilde{D}_0 < \tilde{D}_h)\right), \tag{3.12}$$

where β is the average number of potential pooling passengers along the route.

To interpret Eq. (3.11), note that the passenger may be picked up by either a fully or partially vacant vehicle, and thus her expected pickup time is the average weighted by the probability of each scenario. As for Eq. (3.12), the detour trip occurs only if at least one potential pooling passenger cannot find a closer and fully vacant vehicle. Since $\tilde{D}_0$, $\tilde{D}_h$ and β above depend on the density of partially and fully vacant vehicles and the density of pooling passengers, it seems rather difficult to specify the probabilities in Eqs. (3.11) and (3.12) in a way that can be empirically calibrated and validated.

Additionally, it is unclear whether explicitly modeling en-route pooling would make a meaningful difference in an idealized macroscopic market. The proposed pooling model centers on two basic trade-offs: pooling helps increase the competing power of pooling passengers, and it becomes more attractive when its market share increases (because higher demand reduces the detour distance). In en-route pooling, these basic trade-offs not only exist, but also are expected to play the same dominating role. Because the origin and destination of all trips are uniformly distributed, fully and partially vacant vehicles would also be evenly distributed relative to passengers waiting for pooling rides. Accordingly, pooling passengers still enjoy a greater access to supply because they can hail both fully and partially vacant vehicles. Also, a higher density of pooling passenger still leads to a shorter detour, even if it occurs en-route. Therefore, the total trip time of a pooling ride (inclusive of detour and wait time) should not vary much in en-route pooling. The main difference is where detours occur, which need not concern us.

For the above reasons, en-route pooling is not explicitly modeled in this thesis.

3.3.2. Assumptions on k and b

Assumptions 2 and 3 imply that the platform can and will dynamically adjust the matching and dispatching algorithms to achieve a desired efficiency. We make these assumptions for two reasons. First, the e-hail matching is such a complex process that itself is being actively researched. Previous studies have shown that decision variables like matching interval, matching radius and maximum allowed detour in pooling are all critical to matching performance [e.g., Yang et al., 2020a, Xu et al., 2020, Ke et al., 2020a]. In this study, we choose not to explicitly model these details. Instead, we use k and b to represent the overall efficiency obtained by the platform's matching policy, and calibrate them from empirical observations. This enables us to focus on the main effect of the demand-supply relationship on the matching process. Secondly, by setting k and b as exogenous, the platform's pricing strategies—the focus of this study—are isolated from its matching strategies.

This simplification allows the former to drive the passenger demand and vehicle supply, while the latter's effect is incorporated through the parameters k and b.

We note that Assumptions 2 and 3 may be violated in some cases. If the platform employs a fixed matching strategy (e.g., a constant matching interval), it may not be able to maintain k and b at a stable level when the market conditions vary. While an increasing number of studies consider dynamic matching policies [e.g., Özkan and Ward, 2020, Qin et al., 2021], using a constant matching interval/radius is common in practice [e.g., Yan et al., 2019]. Such a potential violation of the assumptions may introduce estimation errors in passenger wait time and market equilibrium. Assumptions 2 and 3 may also be violated when the market enters a hyper-congested state known as Wild Goose Chase (WGC) [Castillo et al., 2018], which is often accompanied by exceedingly long matching time. WGC is a state when the system throughput Q decreases with passenger wait time w. From Eq. (3.6), we have

$$w_e = \frac{\delta}{2v}\sqrt{\frac{\Pi}{k\Lambda}} \quad \Rightarrow \quad w_e^2 = \frac{\delta}{4v^2}\frac{Qw_e}{kV} \quad \Rightarrow \quad V = \frac{\delta}{4v^2}\frac{Q}{kw_e}.$$

Plugging into the flow conservation $N = V + Q\tau$ yields

$$N = \frac{\delta}{4v^2}\frac{Q}{kw_e} + Q\tau \quad \Rightarrow \quad Q = \frac{N}{\tau + \frac{\delta}{4v^2 k w_e}}.$$

When N and other parameters, including k, are fixed, Q is monotonically increasing with w_e and thus WGC would never emerge. This violation is due to the assumption that k remains constant over time. In reality, it is expected that the matching process becomes inefficient under extreme demand-supply imbalance. In other words, k is more likely to be a piece-wise function of w_e. That is, when w_e is below certain threshold, k is a constant; and as w_e exceeds the threshold, k would decrease with w_e, i.e., $k'(w_e) < 0$. Accordingly,

$$\frac{\partial Q}{\partial w_e} = \frac{N}{\left(\tau + \frac{\delta}{4v^2 k(w_e) w_e}\right)^2} \frac{\delta}{(4v^2 k(w_e) w_e)^2} \left(k(w_e) + k'(w_e) w_e\right),$$

and the system enters WGC when $w_e > -k(w_e)/k'(w_e) \rightarrow \frac{\partial Q}{\partial w_e} < 0$.

Nevertheless, as mentioned before, WGC is unlikely to arise in our setting because the demand for e-hail services is elastic, in the sense that transit is always a feasible fallback option. Hence, passengers would begin to leave the e-hail market long before WGC materializes. In summary, Assumptions 2 and 3 not only simplify the matching model significantly but also separate the optimization of pricing from matching. However, as these assumptions sometimes deviate from the practice in the industry, they may also become a source of estimation errors. We leave it to a future study to relax these assumptions and to refine the matching model for equilibrium analysis.

3.4. Numerical experiment

Eqs. (3.6) and (3.9) suggest that offering pooling always lowers the wait time of solo passengers. However, the benefit is less clear for pooling passengers. Therefore, in this section, we conduct a sensitivity analysis on the wait time of e-hail, solo and pooling rides.

Figure 3.3 examines how waiting passenger density Π, vacant vehicle density Λ and fraction of waiting passenger for pooling $r_\Pi = \Pi_p/\Pi$ affect the passenger wait time and the fraction of pickup detour. When not varying, the parameters set as the default values reported in Table 3.1. Note that the wait time for both passengers in a pooling ride is $w_{p1} + w_{p2}$, because the first passenger to be picked up also endure the pickup detour on the way to pick up the other passenger.

3.4.1. Parameter setting

As shown in Figure 3.3(a), while the wait times of solo and e-hail passengers *increase* with the waiting passenger density, that of pooling passengers *decreases*, thanks to shorter pickup detours. In other words, pooling becomes more desirable when the demand level is high. Although a larger waiting passenger density intensifies the competition and increases the total wait time, its impact on pooling passengers is better mitigated by the collective competing power and the reduction of the detours. Nevertheless, when the system becomes overly congested (larger than 30 passengers

Table 3.1. Default valuesof parameters.

Parameter		**Unit**	**Value**
Detour ratio of road network	δ		1.3
Cruising speed	v	mph	13.6
Matching efficiency	k	/sqmi	0.16
Pooling efficiency	b		0.05
Approximation parameter	κ		4
Vacant vehicle density	Λ	/mi^2	70
Waiting passenger density	Π	/mi^2	24
Fraction of waiting passenger for pooling	r_Π		0.4

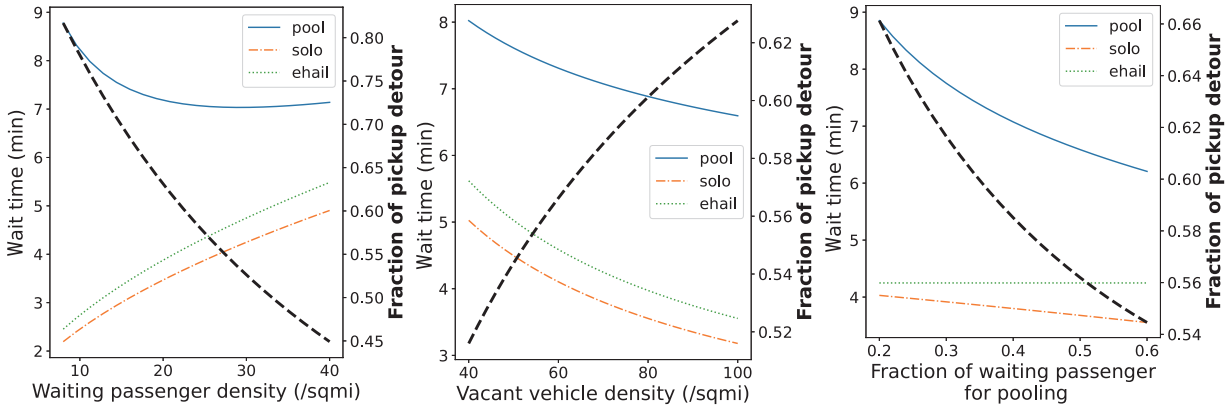


Figure 3.3. Sensitivity of wait time and fraction of pickup detour to (a) waiting passenger density, (b) vacant vehicle density and (c) fraction of waiting passenger for pooling.

per square mile in this example), the benefit of pooling diminishes, slightly pushing up wait time. On the other hand, when the vacant vehicle density increases, all passenger wait times drop, as shown in Figure 3.3(b). Yet, pooling benefits less than the other two due to the existence of pickup detour. According to Figure 3.3(c), while both solo and pooling passengers benefit substantially from the increase in the fraction of waiting passenger for pooling, pooling passengers' gain is greater thanks to the collective competing power of the pair. Figure 3.3 also reveals that the fraction of detour time in the pooling passengers' wait time drops quickly as the waiting passenger density and fraction of pooling passenger among them increase.

To summarize, pooling becomes more appealing as the demand level grows, and the rise of fraction of waiting passenger for pooling further reduces the wait time and thereby attracts more

demand. However, this seemingly positive feedback leaves out an important caveat—pooling tends to prolong a trip due to the pair's different destinations, on top of the detour time incurred in the pickup phase. This *en-route detour time* squeezes the supply because a vehicle serving pooling rides will be occupied longer on average. We will show how this effect is captured in an equilibrium model in Chapters 5 and 6.

3.5. Summary

In this chapter, we present the matching model developed in Chen et al. [2018] and Zhang et al. [2019a], then extend it to accommodate the pooling service. In brief, the existence of pooling benefits both solo and pooling passengers by reducing the competition among passengers. Pooling passengers also enjoy a collective competing power over solo passengers, leading to a shorter pickup time, i.e., the time consumed to pick up the closer passenger. However, pooling passengers endure additional pickup detour that sacrifices the LOS. Both the pickup time and pickup detour depend on the pooling waiting passenger density Π_p, Given other parameters unchanged, the pickup detour decreases with Π_p as it becomes easier to pair pooling passengers. In contrast, the pickup time increases with Π_p because the collective competing power diminishes as the distance between peer passengers decreases. The sensitivity analysis reveals two opposite effects of waiting passenger density on pooling wait time. On the one hand, it helps reduce the pickup detour as passengers could more easily find a peer in close proximity. On the other hand, it intensifies the competition among passengers, which extends the matching and pickup time. Solo and e-hail trips, however, are only subject to the latter effect.

3.6. Appendix

3.6.1. Notations

Table 3.2. List of notations

Variable	Description	Unit
w_e (w_s)	e-hail (solo) passenger wait time	hr
w_{p1}	first part of pooling passenger wait time (matching time plus pickup time of the first passenger)	hr
w_{p2}	second part of pooling passenger wait time (pickup time of the second passenger)	hr
Λ (Λ_0)	vacant (unmatched) vehicle density	/sqmi
Π (Π_0)	waiting (unmatched) passenger density	/sqmi
Π_s (Π_p)	solo (pooling) waiting passenger density	/sqmi
$\tilde{N}_v(d)$ ($\tilde{N}_{mv}(d)$)	number of unmatched (matchable) vehicle within a distance d from a passenger	
$\tilde{N}_p(l)$ ($\tilde{N}_{mp}(l)$)	number of pooling (matchable) passenger within a distance l from a passenger	
v	cruising speed of vacant vehicles	mph
k	coefficient of matching efficiency	/sqmi
b	coefficient of pooling efficiency	
κ	approximation parameter	/sqmi
δ	detour ratio of road network	
$\tilde{D}_e$ ($\tilde{D}_p$)	the distance between the e-hail (pooling) passenger and the closest matchable vehicle (passenger)	mi

3.6.2. Number of vacant vehicles as Poisson Process

Proposition 1. *(Chen et al. [2018] Proposition 1) Under Assumption 1, the counting process $\tilde{N}_v(d)$ is an Inhomogeneous Poisson processes with intensity $2\pi d\Lambda_0$.*

Proof. Due to Assumption 1.1, $\tilde{N}_v(d) = 0$ and the increments of $\tilde{N}_v(d)$ are independent. Consider a ring area defined by d and $d + \Delta d$, and equally cut it into n small pieces with area Δs. Then, the number of vacant vehicle in the ring area follows binomial distribution where each piece contains one vacant vehicle with probability $\Lambda_0 \Delta s$. As n approaches to infinity, such a binomial

distribution can be approximated by a Poisson distribution with rate

$$np = \frac{\pi(d+\Delta d)^2 - \pi d^2}{\Delta s}\Lambda_0 \Delta s = \pi\Lambda_0(2d+\Delta d)\Delta d. \tag{3.13}$$

Hence,

$$Pr(\tilde{N}_v(d+\Delta d) - \tilde{N}_v(d) = 1) = \pi\Lambda_0(2d+\Delta d)\Delta d \exp[-\pi\Lambda_0(2d+\Delta d)\Delta d]$$

$$\Rightarrow \lim_{\Delta d \to 0} \frac{Pr(\tilde{N}_v(d+\Delta d) - \tilde{N}_v(d) = 1)}{\Delta d} = 2\pi d\Lambda_0, \tag{3.14}$$

$$Pr(\tilde{N}_v(d+\Delta d) - \tilde{N}_v(d) > 1) = 1 - \exp[-\pi\Lambda_0(2d+\Delta d)\Delta d] - 2\pi d\Lambda_0$$

$$\Rightarrow \lim_{\Delta d \to 0} \frac{Pr(\tilde{N}_v(d+\Delta d) - \tilde{N}_v(d) > 1)}{\Delta d} \approx \lim_{\Delta d \to 0} \frac{1 - [1 - \pi\Lambda_0(2d+\Delta d)\Delta d]}{\Delta d} - 2\pi d\Lambda_0 = 0. \tag{3.15}$$

Therefore, $\tilde{N}_v(d)$ is an Inhomogeneous Poisson Process with intensity $2\pi d\Lambda_0$. □

3.6.3. Derivation of pickup time in pooling

Define $\tilde{D}_p$ as the minimum distance from *either* passenger to the closest matchable vehicle. To derive the distribution of $\tilde{D}_p$, we introduce *effective search area* as the union of the two search areas. Accordingly, $Pr(\tilde{D}_p \leq d|l)$ gives the probability that at least one matchable vehicle appears inside the effective search area with parameter d conditional on the distance between the pooling pair l.

Let $A(d,l)$ denote the area of the effective search area. As illustrated in Figure 3.4, when $d \leq l/2$, the two passengers' search areas do not overlap. Hence, $A(d,l)$ simply equals $2\pi d^2$. Otherwise, $A(d,l)$ is the total search area less the intersection if $d > l/2$. To summarize,

$$A(d,l) = \begin{cases} 2\pi d^2 & , \ d \leq l/2 \\ 2\pi d^2 - 2\cos^{-1}\left(\frac{l}{2d}\right) d^2 + dl\sqrt{1 - \left(\frac{l}{2d}\right)^2} & , \ d > l/2 \end{cases}. \tag{3.16}$$

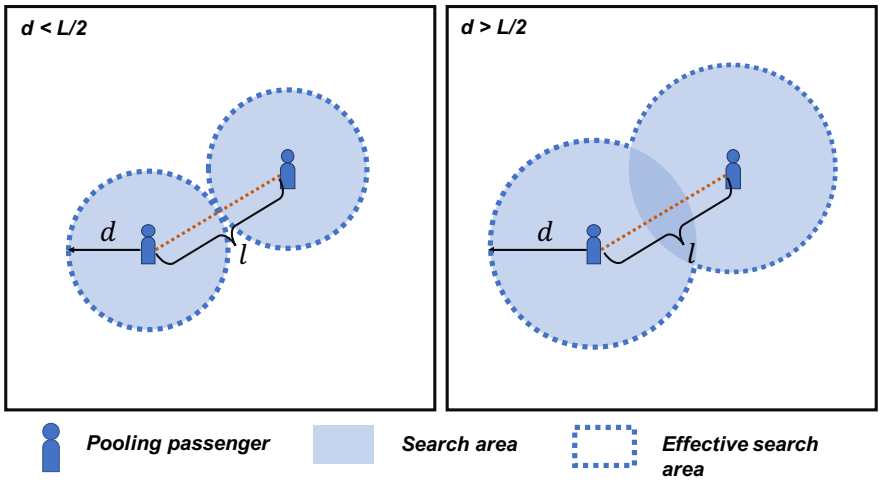


Figure 3.4. Illustration of search area and effective search area.

$Pr(\tilde{D}_p \leq d|l)$ is thus given by

$$Pr(\tilde{D}_p \leq d|l) = 1 - \exp\left(-\frac{k\Lambda}{\Pi_{\text{eff}}} A(d,l)\right). \tag{3.17}$$

As per Eq. (3.17), given d, the collective competing power is controlled by l. It doubles when $l \geq 2d$, though reduces as l, leading to the same wait time as solo passengers when $l = 0$. The non-smoothness of $A(d,l)$, however, bring difficulty to further evaluate the expectation of $\tilde{D}_p$. Therefore, we propose to approximate it with a smooth function as follows:

$$\hat{A}(d,l) = \left(2 - \frac{1}{1+\kappa l^2}\right)\pi d^2. \tag{3.18}$$

It is easy to verify that $\hat{A}(d,l) \to \pi d^2$ as $l = 0$ and $\hat{A}(d,l) \to 2\pi d^2$ as $l \to \infty$. Thus, $\hat{A}(d,l)$ well captures the lower and upper bounds of $A(d,l)$ and the parameter κ may be adjusted to achieve good approximation.

Using $\hat{A}(d,l)$, the conditional expectation $E[\tilde{D}_p|l]$ is derived as

$$E[\tilde{D}_p|l] = \frac{1}{2}\sqrt{\frac{\Pi_{\text{eff}}}{k\Lambda}}\left(2 - \frac{1}{1+\kappa l^2}\right)^{-1/2}. \tag{3.19}$$

Recall that l is a realization of random variable $\tilde{L}$. Thus, the expectation of $\tilde{D}_p$ is given by

$$D_p = E[E[\tilde{D}_p|\tilde{L}]] = \int_0^\infty E[\tilde{D}_p|l]\mathrm{d}F_{\tilde{L}}(l), \tag{3.20}$$

where $F_{\tilde{L}}(l)$ is CDF of $\tilde{L}$.

The above integral cannot be derived analytically due to the functional form of Eq. (3.19). Instead, we introduce the following approximation:

$$D_p = E[E[\tilde{D}_p|\tilde{L}] \approx E[\tilde{D}_p|E[\tilde{L}]] = \frac{1}{2}\sqrt{\frac{\Pi_{\text{eff}}}{k\Lambda}}\left(2 - \frac{1}{1+\kappa L^2}\right)^{-1/2}. \tag{3.21}$$

Plugging Eq. (3.7) into Eq. (3.21) thus yields

$$w_{p1} = \frac{\delta}{v}D_p = \frac{\delta}{2v}\sqrt{\frac{\Pi_{\text{eff}}}{k\Lambda}\frac{\kappa + 4b\Pi_p}{2\kappa + 4b\Pi_p}}. \tag{3.22}$$

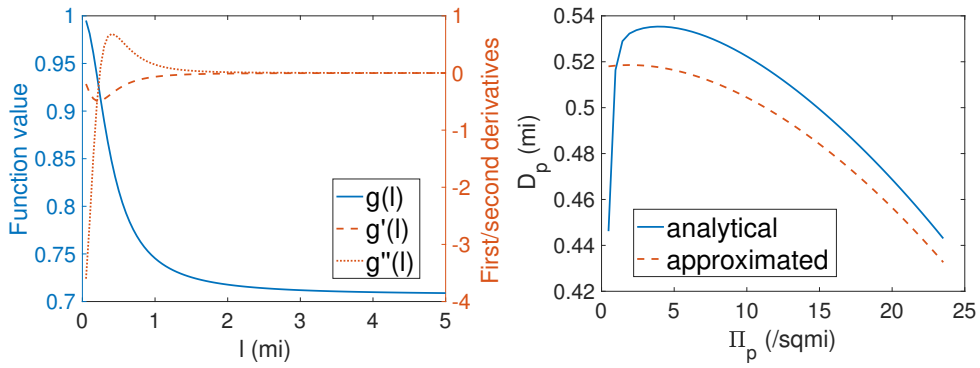
The approximation made in Eq. (3.21) warrants some discussions. Introduce a new function

$$g(l) = \left(2 - \frac{1}{1+\kappa l^2}\right)^{-1/2}. \tag{3.23}$$

Then, the approximation made in Eq. (3.21) is equivalent to $E[g(l)] \approx g(E[l]) = g(L)$. The approximation quality depends on the functional form of $g(l)$. As per Jensen's inequality, if $g(l)$ is a linear function, the approximation is subject to no error. Otherwise, $E[g(l)] \geq g(L)$ holds when $g(l)$ is convex, and $E[g(l)] \geq g(L)$ if $g(l)$ is concave.

Figure 3.5(a) plots the function value, first and second derivatives of $g(l)$ with $\kappa = 4$. It can be seen that $g(l)$ is convex when $l > 0.25$ and quickly converges to $1/\sqrt{2}$. Hence, Eq. (3.21) is likely to underestimate D_p (hence w_{p1}). This finding is confirmed in Figure 3.5(b), which compares the approximated value and the analytical result given by Eq. (3.20) (the analytical result is computed by numerical integration). Specifically, we fix the total waiting passenger density and vacant vehicle density, and vary the waiting passengers density for pooling Π_p. All the parameters are set according to Table 3.1. As shown in Figure 3.5(b), the approximated value slightly underestimates D_p except

when Π_p is close to 0. For most values of Π_p, the error is within 5% and it decreases as Π_p increases. Therefore, Eq. (3.21) does offer a reasonable approximation for D_p.



(a) Functional properties of $g(l)$ (b) Analytical and approximated D_p

Figure 3.5. Analysis of approximation quality

Part 2

Aggregate market with pooling

CHAPTER 4

Aggregate market equilibrium

The main stakeholders in an e-hail market are passengers, drivers and the platform. After the platform sets the trip fare and the compensation rate, passengers and drivers respond with their own decision on mode choice and market entry. In essence, the resulting problem can be described as a Stackelberg game [Von Stackelberg, 2010], in which the platform is the leader while passengers and drivers are followers. The same framework has been widely applied in transportation research [e.g., Zhou et al., 2005, Yang et al., 2007, Zhang and Nie, 2018]. The regulations, however, perform as constraints in the platform's pricing problem and thus influence the market equilibrium indirectly. In this chapter, we will first establish the market equilibrium given the platform's pricing strategy, i.e., the follower's game. Based on the equilibrium, the following two chapters will discuss the platform's pricing problem in both monopoly and duopoly scenarios, i.e., the leader's game, with and without regulatory constraints.

4.1. Main assumptions

The market equilibrium is dictated by the interaction between demand and supply. On the demand side, passengers choose among solo, pooling and transit based on their generalized cost. On the supply side, drivers decide whether or not to enter the market according to the average earning rate. The matching model presented in Chapter 3 connects the demand and the supply, by characterizing the passenger wait time and vehicle occupancy that directly affect passengers' mode choice and drivers' entry choice.

Before we present the equilibrium model, let us first state the main assumptions as follows

Assumption 4. *Transit is a viable mode to all passengers, and is supplied at a constant generalized cost. Also, the transit operator always breaks even, i.e., the fare equals the marginal cost.*

Assumption 5. *Registered e-hail drivers enjoy flexible working schedules. Their decision to enter the service solely depends on the average earning rate offered by the platform relative to that provided by other job opportunities that they can freely pursue.*

Assumption 6. *The amount of vehicular traffic contributed by e-hail services is small, and hence the extra congestion effect is not explicitly considered in the utility of all three modes.*

We note that Assumption 6 is introduced to focus our analysis on the trade-off in pricing between solo and pooling. In Chapter7, it will be relaxed to investigate the congestion impact of e-hail vehicles due to the platform's operational strategies and regulations.

4.2. Passenger demand

We characterize passengers' choice from a discrete set of modes $\mathcal{M} = \{s, p, t\}$, where s, p, and t refers to solo, pooling and transit, respectively. Let ν be the value of time, and f_i and $\tau_i, i \in \mathcal{M}$ be, respectively, the trip fare and the average duration of mode i. Typically, $\tau_p > \tau_s$ because pooling tends to prolong a trip due to the pair's different destinations. We define the generalized cost for passengers choosing one of the three modes as follows.

$$\text{Solo ride: } u_s = f_s + \nu(w_s + \tau_s), \tag{4.1a}$$

$$\text{Pooling ride: } u_p = f_p + \nu(w_{p1} + w_{p2} + \tau_p), \tag{4.1b}$$

$$\text{Transit: } u_t = f_t + (\nu + \zeta)\tau_t. \tag{4.1c}$$

According to Assumptions 4 and 6, f_t, τ_s and τ_t are all treated as constants. Notations used in this chapter are summarized in Table 4.1.

In theory, τ_p should be endogenous, as a decreasing function of the pooling passenger density. Two recent studies investigate this issue for pooling rides shared by two passengers. By simulating the matching process of on-demand pooling using empirical demand data collected in three large cities, Ke et al. [2021] found that the ratio between the detour distance and the average trip distance is inversely proportional to a function of the batch demand (i.e., the number of requests accumulated in a matching interval). Lobel and Martin [2020] analyze the detour and the value associated with pooling. The latter is measured by the reduced total travel distance. The authors define the detour (value) ratio as the total detour (value) normalized by the total travel distance when both passengers take solo rides. They show that the sum of the two ratios is bound by 0.5, which also implies the detour ratio cannot exceed 0.5. Although these results bring valuable insights, implementing them in our framework is impractical. On the one hand, the model proposed in Ke et al. [2021] cannot be properly calibrated with the data we have. On the other hand, Lobel and Martin [2020] only set an upper bound on the detour ratio. Hence, in this study, we simply assume τ_p to be a constant, which is calibrated from the empirical data (see Section 10.2). In Section 4.6, we test our model's sensitivity to τ_p. For the reader who wonders how much a difference an endogenized detour time could make, in Section 4.8.3, we implement a version of our equilibrium model using the detour model of Ke et al. [2021], and perform a sensitivity analysis against key coefficients.

Note that ζ in Eq. (4.1c) represents additional disutility of transit relative to ride-hail services (associated with privacy, comfort, crowdedness, etc.). Following Schwieterman [2019], we set $\zeta = 0.25\nu$ in this study.

Suppose the total demand is D_0 and the share of each mode is a continuous and differentiable function $q : \mathbb{R}^3_+ \to (0, 1)$ of the general costs of all three modes, i.e.,

$$Q_i = D_0 q(u_i, \mathbf{u}_{-i}),\ i \in \mathcal{M}, -i := \mathcal{M} \setminus \{i\}, \tag{4.2}$$

where $\mathbf{u}_{-i}$ refers to modes except for i. Without loss of generality, we assume $\partial Q_i/\partial u_i < 0$ and $\partial Q_i/\partial \mathbf{u}_{-i} > 0$. Accordingly, the pooling ratio is given as $r_Q = Q_p/(Q_s + Q_p)$.

4.3. Vehicle supply

Let S_0 be the potential supply, which may be viewed as the total number of drivers registered on the platform. As per Assumption 5, we define $\tilde{e}_0$ as the earning rate of the alternative employment opportunity available to the drivers. Often known as the *reservation rate*, $\tilde{e}_0$ is modeled as a random variable with a cumulative distribution function (CDF) $g(\cdot)$. Drivers enter the e-hail market only if doing so yields an earning rate $e \geq \tilde{e}_0$. This assumption aligns with several studies that empirically observe a positive wage elasticity of supply among ride-hail drivers [e.g., Angrist et al., 2017, Chen and Sheldon, 2017, Sun et al., 2019] Accordingly, the fleet size is derived as

$$N = S_0 Pr(\tilde{e}_0 \leq e) = S_0 g(e). \tag{4.3}$$

The earning of an e-hail driver is determined by the compensation rate per unit occupied time, denoted as η. In addition, a driver serving pooling rides may also be paid a fixed fee c_p for each additional pickup. Thus, the average earning rate of a driver is computed as

$$e = \frac{1}{N}\left[\eta\left(Q_s\tau_s + \frac{1}{2}Q_p\tau_p\right) + \frac{1}{2}c_pQ_p\right], \tag{4.4}$$

where $Q_s\tau_s + \frac{1}{2}Q_p\tau_p$ is the total occupied time and $\frac{1}{2}Q_p$ denotes the number of additional pickups.

4.3.1. Equilibrium with single platform

With the demand and supply specified in previous two subsections and the wait times derived in Chapter 3, the aggregate equilibrium in a unit time period is characterized by the following system of equations:

$$\text{mode choice:} \quad Q_i = D_0 q(u_i, \mathbf{u}_{-i}),\ i \in \{s, p\}, \tag{4.5a}$$

$$\text{Fleet size:} \quad N = S_0 g\left(\frac{1}{N}\left[\eta\left(Q_s\tau_s + \frac{1}{2}Q_p\tau_p\right) + \frac{1}{2}c_pQ_p\right]\right), \tag{4.5b}$$

$$\text{Flow conservation:} \quad N = V + Q_s\tau_s + \frac{1}{2}Q_p\tau_p, \tag{4.5c}$$

$$\text{Solo wait time:} \quad w_s = \frac{\delta}{2v}\sqrt{\frac{Q_s w_s + Q_p(w_{p1}+w_{p2})/2}{kV}}, \tag{4.5d}$$

$$\text{Pooling wait time:} \quad w_{p1} = w_s\sqrt{\frac{\kappa + 4bQ_p(w_{p1}+w_{p2})}{2\kappa + 4bQ_p(w_{p1}+w_{p2})}}, \tag{4.5e}$$

$$w_{p2} = \frac{\delta}{2v}\frac{1}{\sqrt{bQ_p(w_{p1}+w_{p2})}}. \tag{4.5f}$$

Eq. (4.5c) states that the total vehicle operation time (represented by the fleet size N) consists of three parts: (i) the vacant vehicle time (V) that includes both idle and pickup time; (ii) the time spent on delivering solo passengers; and (iii) the time spent on delivering pooling passengers. Since the market equilibrium is established over unit time period, V substitutes Λ in Eqs. (3.9) and (3.10). Eqs. (4.5d)–(4.5f) are obtained from substituting Π_s and Π_p in Eq. (3.7)–(3.10) by $\Pi_s = Q_s w_s$ and $\Pi_p = Q_p(w_{p1} + w_{p2})$ as per Little's formula [Little, 1961].

4.4. Equilibrium with multiple competing platforms

Consider an aggregate ride-hail market where a set of platforms, denoted as $\mathcal{P} = \{A, B, \dots\}$, competing with each other and against transit, by offering solo and/or pooling rides. The demand model is the same as that defined in Section 4.2, except that the set of modes expands. The supply model, however, depends on the particular supply mode of drivers. Here, we consider two scenarios: *single-homing*, where each driver only joins one platform, and *multi-homing*, where each driver joins all platforms.

4.4.1. Single-homing

In the case of single-homing, a driver working for platform j is paid at a compensation rate η^j per unit occupied time. Then, the average wage rate is given by

$$e^j = \frac{\eta^j}{N^j}\left(Q_s^j \tau_s^j + \frac{1}{2}Q_p^j \tau_p^j\right), \tag{4.6}$$

the flow conservation condition Eq. (4.5c) becomes

$$V^j = N^j - Q_s^j \tau_s^j - \frac{1}{2} Q_p^j \tau_p^j, \tag{4.7}$$

and the effective waiting passenger density is

$$\Pi_{\text{eff}}^j = Q_s^j w_s^j + \frac{1}{2} Q_p^j (w_{p1}^j + w_{p2}^j). \tag{4.8}$$

Recall that the effective waiting passenger density Π_{eff} measures the passenger competition for vacant vehicles. In the case of single-homing, solo and pooling passengers from the same platform compete for the vacant vehicles of that platform.

Similar to the case of single platform, the market equilibrium with multiple platforms can be recast as a fixed point system $\mathbf{x} = F(\mathbf{x})$, where $\mathbf{x} = [\mathbf{w}_s, \mathbf{w}_{p1}, \mathbf{w}_{p2}, \mathbf{N}]^T$, and $\mathbf{w}_s := \{w_s^j, j \in \mathcal{P}\}, \mathbf{w}_{p1} := \{w_{p1}^j, j \in \mathcal{P}\}$ and $\mathbf{w}_{p2} := \{w_{p2}^j, j \in \mathcal{P}\}$. In other words, the market equilibrium can be fully characterized by the vectors of passenger wait time (solo and pooling) and the fleet size.

4.4.2. Multi-homing

In practice, the flexibility promised by TNCs allows drivers to register and work on multiple platforms. The obvious rationale for a driver is to follow jobs wherever they arise, rather than putting all eggs (in this case his time) in one basket. In Chicago, for example, about 25% drivers are affiliated with more than one platform in September 2019, and a preliminary analysis suggests these drivers do serve more trips on average than others[1]. The situation considered here is an extreme version of multi-homing, though it should provide insights on the real practice. Accordingly, we have $N = N^j, j \in \mathcal{P}$ and the average wage rate is given by

$$\bar{e} = e^j = \frac{1}{N} \sum_j \eta^j \left(Q_s^j \tau_s^j + \frac{1}{2} Q_p^j \tau_p^j \right). \tag{4.9}$$

[1]Computed from the Chicago TNC data in September 2019 (see Section 10.1)

Besides, there is only one flow conservation equation Eq. (4.5c), i.e.,

$$V = V^j = N - \sum_j \left(Q_s^j \tau_s^j + \frac{1}{2} Q_p^j \tau_p^j \right). \tag{4.10}$$

Furthermore, as the waiting passengers compete for the same pool of vacant vehicles, the effective waiting passenger density is the sum of all platforms, i.e.,

$$\Pi_{\text{eff}} = \Pi_{\text{eff}}^j = \sum_j \left[Q_s^j w_s^j + \frac{1}{2} Q_p^j (w_{p1}^j + w_{p2}^j) \right]. \tag{4.11}$$

Here, solo and paired pooling passengers from each platform join the competition for the same pool of vacant vehicles. It is worth emphasizing that, although multi-homing intensifies competition for vacant vehicles, it has nothing to do with the process of pairing passengers. Pooling passengers from the same platform are always paired among themselves and together they compete for vacant vehicles. This is reflected in the fact that the detour w_{p2}^j depends on Π_p^j, which is different among platforms. In other words, our model does not allow passengers on different platforms to be matched in the same pooling ride.

Therefore, multi-homing intensifies the inter-passenger competition even though it expands the pool of vacant vehicles. With single-homing, a passenger on one platform only competes with those who use the same platform. Yet, she would have to compete with every other waiting passenger in the aggregate market under multi-homing mode. Consider a system with two symmetric single-homing platforms. At a stationary state, the passenger wait time of each platform is dictated by the ratio Π_{eff}/Λ, where Π_{eff} and Λ are respectively the effective waiting passenger density and vacant vehicle density on *each* platform. If the system switches from single-homing to multi-homing, the ratio remains the same, because both Π_{eff} and Λ double. As a result, the passenger wait time would not be affected at all. In other words, the benefit of having a greater supply pool is offset by the more intense competition on the demand side.

Eq. (4.11) also implies that the solo wait times of all platforms are linearly correlated, i.e., $w_s^{j_1}/w_s^{j_2} = \sqrt{k^{j_2}/k^{j_1}},\ \forall j_1, j_2 \in \mathcal{P}$. Therefore, it suffices to represent the solo wait time vector

$\mathbf{w}_s$ using the solo wait time on any reference platform. If we choose $j = A$ as the reference platform, then the solution to the multi-homing equilibrium problem can be represented using a vector $\mathbf{x} = [w_s^A, \mathbf{w}_{p1}, \mathbf{w}_{p2}, N]^T$.

4.5. Existence and stability

In what follows, we present the analysis of existence and stability of the market equilibrium with single platform, while that for the case of two platforms follows the same rationale. Define $\mathbf{x} = (w_s, w_{p1}, w_{p2})$. Then, Q_s, Q_p, N and V can be viewed as functions of $\mathbf{x}$ according to Eqs. (4.5a)–(4.5c). Plugging them into Eqs. (4.5d)–(4.5f) thus reduces the equilibrium to a fixed-point system $\mathbf{x} = F(\mathbf{x})$. With mild assumptions, we prove the solution existence of such a fixed-point system by invoking Brouwer's theorem [Brouwer, 1911], as summarized in the following proposition.

Proposition 2. *Suppose* $\mathbf{x}$ *is bounded from above by* $\bar{\mathbf{x}} = (\overline{w}_s, \overline{w}_{p1}, \overline{w}_{p2})^T$. *Then, there exists an* $\mathbf{x}^* = (w_s^*, w_{p1}^*, w_{p2}^*)^T$ *that satisfies Eq.* (4.5).

Proof. Brouwer's fixed-point theorem [Brouwer, 1911] states that: if a continuous function $f : \Omega \subset R^n \to \Omega$ maps a compact and convex set Ω to itself, then there exists $\mathbf{x}^* \in \Omega$ such that $\mathbf{x}^* = F(\mathbf{x}^*)$.

We first prove Ω is compact and convex. By definition, $\mathbf{x} = (w_s, w_{p1}, w_{p2})^T \in \Omega \subset \mathbb{R}^3_+$. From the assumption made in the proposition, w_s, w_{p1} and w_{p2} must all be bounded from the above, otherwise the demand for solo and/or pooling rides will disappear all together. We now show that these three variables must also have lower bounds. First, passenger wait times reach lower bounds $\underline{w}_s = \frac{\delta}{2v\sqrt{S_0}}$ and $\underline{w}_p = \frac{\delta}{2v\sqrt{2S_0}}$, when the fleet size approaches its upper bound S_0 and the demand D_0 approaches zero. Second, as Π_p is bounded from above by $D_0(\overline{w}_{p1} + \overline{w}_{p2})$, the lower bound of w_{p2} is given by $\underline{w}_{p2} = \frac{\delta}{2v}[bD_0(\overline{w}_p + \overline{w}_{p2})]^{-1/2}$. Consequently, Ω can be defined as the cubic space $[\underline{w}_s, \overline{w}_s] \times [\underline{w}_{p1}, \overline{w}_{p1}] \times [\underline{w}_{p2}, \overline{w}_{p2}]$, which is compact and convex.

We proceed to show the self-map $F(\cdot)$, i.e., Eqs. (4.5d)–(4.5f), is continuous. Recall that Eqs. (4.5d)–(4.5f) are continuous functions of Q_s, Q_p and V, along with w_s, w_{p1} and w_{p2}. From

(4.5c), we know V is a continuous function of Q_s, Q_p and N. Therefore, to show F is continuous, we only need to prove Q_s, Q_p and N are continuous functions of w_s, w_{p1} and w_{p2}. The continuity of previous two are directly shown from Eqs. (4.5a) and (4.1). The last one is more complicated as it involves the implicit function Eq. (4.5b). This result is formally stated Lemma 1 and its proof is provided in Section 4.8.4.

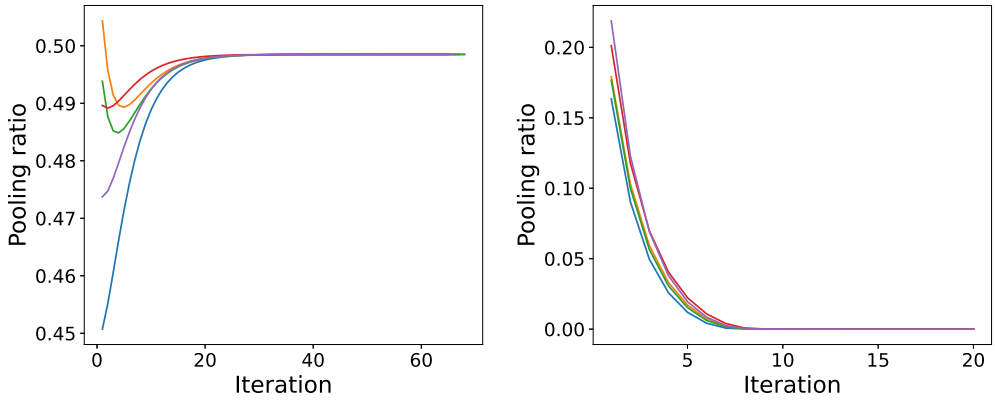
Lemma 1. *The fleet size N defined in Eq.* (4.5b) *can be represented as a continuous function of $\mathbf{x} = (w_s, w_{p1}, w_{p2})^T$.*

Therefore, both conditions stated in Brouwer's fixed-point theorem are satisfied. We hence conclude that the existence of a solution to Eq. (4.5) is guaranteed. □

The assumption made in Proposition 2 effectively sets upper bounds on passenger wait times for both solo and pooling rides. At first glance, this seems at odds with Eqs. (4.5d)–(4.5f), which allow these wait times to grow infinitely. Nevertheless, if passengers have to wait exceedingly long, the demand would be suppressed below a level of practical interest. In other words, Proposition 2 ensures a solution always exists provided that the demand for solo and pooling rides has a lower bound that can, in theory, be arbitrarily close to zero.

An equilibrium solution to Eq. (4.5) can be obtained through an iterative fixed-point algorithm. While implementing such an algorithm is straightforward, we note that, as a highly nonlinear system, Eq. (4.5) may not have a unique equilibrium. In addition, an equilibrium solution may or may not be stable. The stability theory [e.g., Strogatz, 2018] states that a solution $\mathbf{x}^*$ is stable if and only if all eigenvalues of the Jacobian matrix of $F(\cdot)$ at $\mathbf{x}^*$, denoted by $J_{\mathbf{x}^*}$, have absolute values less than 1. Using this result, we screen each equilibrium solution obtained from fixed-point iterations and only keep those that pass the stability test. Since $F(\cdot)$ is explicitly expressed, i.e., Eq. (4.5), we could evaluate $J_{\mathbf{x}^*}$ using automatic differentiation [Baydin et al., 2017].

Figure 4.1 reports the convergence performance of the iterative fixed-point algorithm in two examples, both using the default parameters given in Table 4.2 except for the pricing strategies



(a) $f_s = 17, f_p = 12, \eta = 20$ (b) $f_s = 19, f_p = 17, \eta = 21$

Figure 4.1. Convergence performance of the iterative fixed-point algorithms. (a) All initial solutions lead to the same stable equilibrium; (b) All initial solutions lead to the same unstable equilibrium.

(f_s, f_p and η). The algorithm is terminated when the gap drops below a predefined threshold, set at 10^{-8} in this study. Each convergence curve (gap vs. number of iteration) in the plots represents a different initial solution. These results are representative of our overall experience with the fixed-point algorithm, which is that it generally converges quite fast and the choice of the initial solutions tends to have a negligible impact on convergence. In this particular experiment, the pricing strategy on the left leads to a stable equilibrium and the algorithm is able to locate it regardless of where it starts. However, an unstable equilibrium emerges when the pricing strategy on the right is employed. This happens because the general cost of a pooling ride is close to that of a solo ride when there is no demand for pooling. As a result, a small perturbation in the inputs could make pooling more attractive, forcing the solution to significantly deviate from the pre-perturbation one. Still, the algorithm converges faithfully to this unstable equilibrium, starting from all tested initial points.

In light of the above observations, whenever the market equilibrium is sought, we run the fixed-point algorithm multiple times with randomly generated initial solutions. If all runs converge to the neighborhood of the same fixed point, we take their average as the final equilibrium solution. If there is more than one equilibrium solution, we test their stability and only keep the stable ones.

Curiously, in all numerical experiments conducted in this study, we did not come across a single case where multi equilibrium were found.

4.6. Numerical experiments

In this section, we examine the variations of market equilibrium towards key input parameters under fixed pricing scheme. We first investigate the main equilibrium variables in a monopoly market, then move to analyze the duopoly equilibrium and compare the two supply modes discussed in Section 4.4. The ranges of market inputs, along with other parameters, are set according to the Chicago TNC data (see Chapter 10) and reported in Table 4.2.

4.6.1. Single platform

We characterize passengers' choice based on the random utility theory [Ben-Akiva and Lerman, 1985] and adopt the Multinomial Logit (MNL) model. Therefore, the share of each mode $q(\cdot)$ is estimated as

$$Q_i = D_0 \frac{\exp(-\theta_c u_i)}{\sum_{j\in\mathcal{M}} \exp(-\theta_c u_j)},\ i \in \mathcal{M}, \tag{4.12}$$

where θ_c is a non-negative parameter that reflects the degree of uncertainty in mode choice.

Accordingly, the upper bounds in Proposition 2 can be set based on the minimum demand level considered as meaningful by the modeler. More details are provided in Section 4.8.5.

For simplicity, we assume the drivers' reservation rate $\tilde{e}_0$ follows a uniform distribution on $[0, 2e_0]$, where e_0 is the average reservation rate. Therefore, Eq. (4.3) is simplified as

$$N = S_0 \frac{e}{2e_0} = \frac{S_0}{2e_0 N}\left[\eta\left(Q_s\tau_s + \frac{1}{2}Q_p\tau_p\right) + \frac{1}{2}c_p Q_p\right], \tag{4.13}$$

which yields

$$N = \sqrt{\frac{S_0}{2e_0}\left[\eta\left(Q_s\tau_s + \frac{1}{2}Q_p\tau_p\right) + \frac{1}{2}c_p Q_p\right]}. \tag{4.14}$$

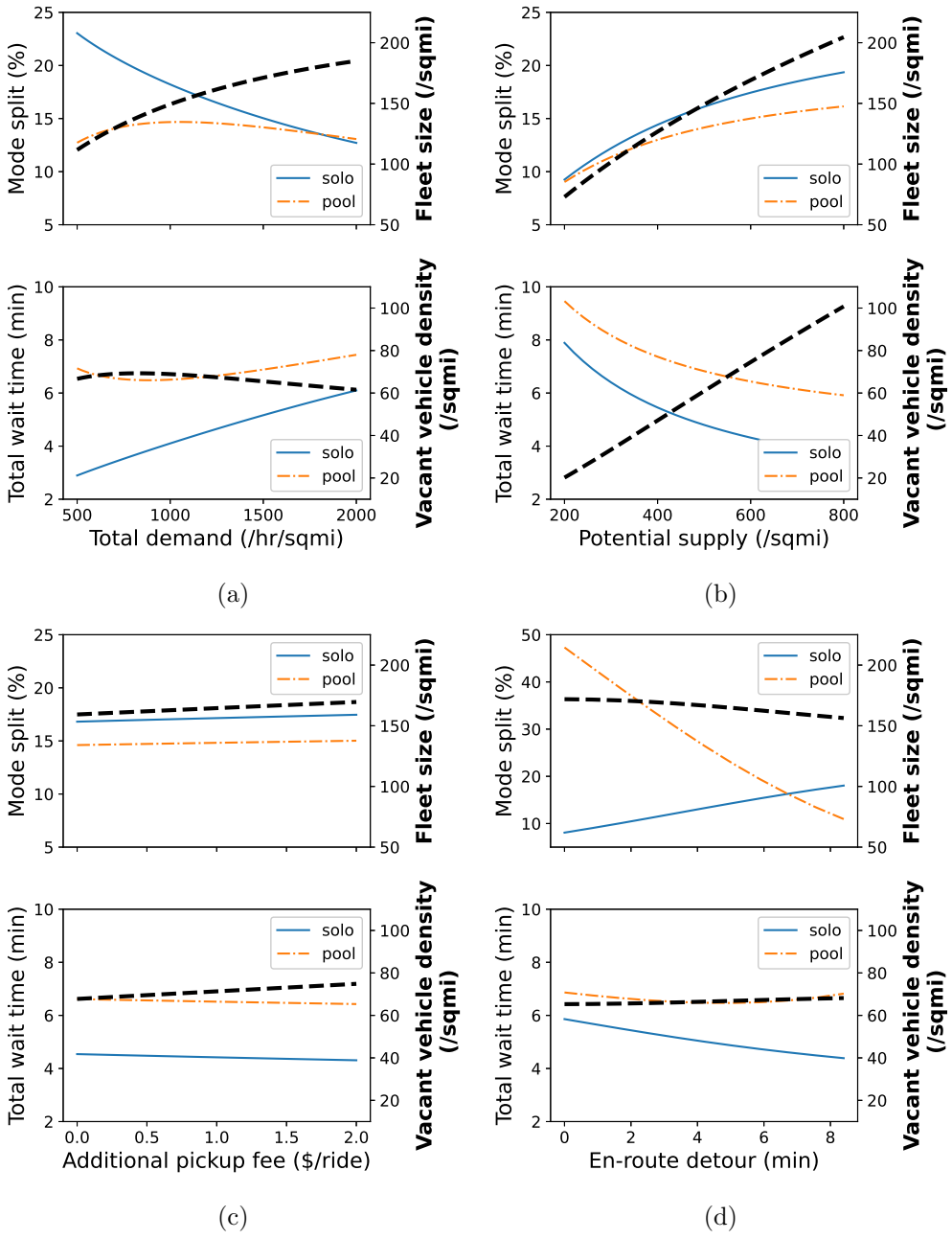


Figure 4.2. Sensitivity of mode share, vehicle supply and passenger wait time to (a) total demand D_0, (b) potential supply S_0, (c) additional pickup fee c_p and (d) en-route detour $\tau_p - \tau_s$.

Figure 4.2 illustrates how mode shares, wait time and vehicle supply vary with the total demand D_0, potential supply S_0, additional pickup fee c_p and en-route detour $\tau_p - \tau_s$. As shown in Figure 4.2(a), the share of pooling rides first increases and then decreases with the total demand.

The share of solo rides drops faster yet remains higher than pooling for most parts. This shift is related to the different impacts of the rising demand on the wait time for the two modes. The rising demand intensifies the competition among waiting passengers, thus steadily increasing the solo wait time. On the other hand, higher demand helps reduce the pickup detour for pooling rides, leading to its initial growth of market share. On the supply side, the growth of total demand induces more drivers to enter the market (top plot). However, it only leads to a mild increase of vacant vehicle density at the very beginning (bottom plot). As the total demand further increases, the level of service for both solo and pooling modes deteriorates.

The growth of S_0 reveals a different pattern; see Figure 4.2(b). The total share of ride-hail increases with the potential supply, while solo rides gain more popularity (top plot). Although the vacant vehicle density increases linearly with the potential supply, passengers enjoy a milder (sub-linear) improvement, owing to competitions on the demand side.

Figure 4.2(c) reveals that the effect of the additional pickup fee for pooling rides (c_p) is almost negligible. When c_p increases from \$0 to \$2, the share of both pooling and solo modes rises less than 0.5% (top plot). Thus, it cannot serve as an effective incentive to encourage drivers to take more pooling rides. Nor could it bring a meaningful improvement to the level of service (the wait time barely changes, see the bottom plot). Given these observations, the additional pickup fee will not be discussed in following analysis, and will be simply set to zero hereafter.

Figure 4.2(d) highlights the importance of the en-route detour time. Specifically, the share of pooling drops sharply as the detour rises from 0. As per Eq. (3.7), the loss of pooling demand results in a longer pickup detour (hence a longer wait time) for pooling passengers, which further discourages pooling. On the other hand, the loss of pooling demand reduces the overall demand level, and as a result, the solo wait time drops despite a shrinking fleet size (top plot). Overall, these findings suggest that *promoting pooling may not be a good strategy for the platform if trip destinations are too scattered to keep the average en-route detour time under control.*

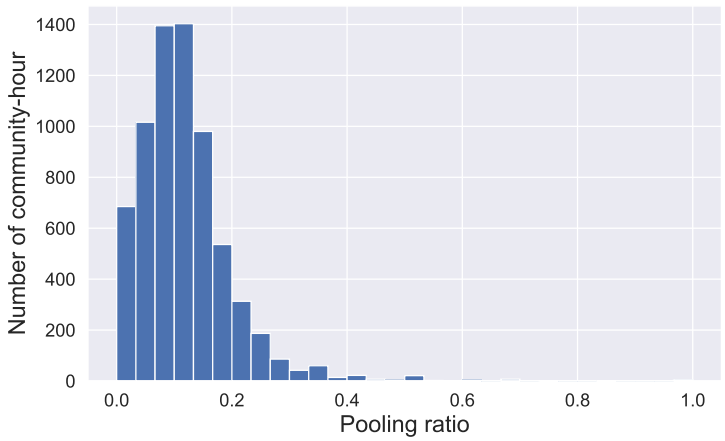


Figure 4.3. Histogram of pooling ratio in the study area and period (based on Chicago TNC data; see Chapter 10.

It is worth noting that results in Figure 4.2 may overestimate the market share of pooling in real practice. In Chicago, from which many of our model's inputs are drawn, the average pooling ratio is less than 15%; see Figure 4.3. A few factors might contribute to this discrepancy. First, our model assumes the trip origins and destinations are uniformly distributed in an aggregate market. The heterogeneous distribution of the real demand is likely to produce strong spatiotemporal imbalance that could undermine the matching efficiency for pooling. The high en-route detour time in Chicago (around 7 min, close to the upper bound in Figure 4.2(d)) may reflect such inefficiency. Second, our model excludes some negative features of pooling (e.g., the loss of privacy and comfort) in favor of simplicity and tractability. Ignoring these factors might underestimate the general cost of pooling.

4.6.2. Two platforms

Different from the single-platform scenario, we estimate a passenger's mode choice using a nested Logit (NL) model [e.g., Williams, 1977], where the generalized cost defined in Eq. (4.1) is used as the observed disutility. Since ride-hail services of different platforms and modes have many features

in common, we introduce a simple nested structure that encompasses them all: the upper level consists of transit and ride-hail while the lower level has four ride-hail modes of both platforms[2].

Accordingly, the mode share of $i \in \{s, p\}$ for platform j is given by

$$Q_i^j = D_0 \frac{\exp(-\theta_c I_c^r)}{\exp(-\theta_c u_t) + \exp(-\theta_c I_c^r)} \frac{\exp(-\theta_c^r u_i^j)}{\exp(-\theta_c^r I_c^r)}, \tag{4.15}$$

Here, θ_c is the uncertainty parameter of choice between transit and ride-hail services. The composite cost of ride-hail services I_c is specified as

$$I_c^r = -\frac{1}{\theta_c^r} \ln \left[\sum_{j \in \mathcal{P}} \sum_{m \in \{s,p\}} \exp(-\theta_c^r u_m^j) \right], \tag{4.16}$$

where θ_c^r is the uncertainty parameter of choice among different modes and platforms (a larger value corresponds to larger uncertainty).

Based on the assumption of NL, we have $\theta_c^r \geq \theta_c$, and the ratio θ_c/θ_c^r measures the magnitude of correlations among the ride-hail modes. $\theta_c/\theta_c^r = 1$ implies that all ride-hail modes are independent alternatives, i.e., the independence of irrelevant alternatives (IIA) assumption holds. On the other hand, if $\theta_c/\theta_c^r \to 0$, all ride-hail alternatives will be viewed together as a single ride-hail mode (with a utility equal to the average of the individual alternatives), and the share of this "composite mode" will be evenly split among the original alternatives.

Since the data available is insufficient to calibrate the dispersion parameters, in most experiments, we simply set $\theta_c^r = \theta_c$ and $\theta_d^r = \theta_d$. Because this setting effectively upholds the IIA assumption, it tends to overestimate the total rid-hail market share in the duopoly case. To gauge this impact, a sensitivity analysis is performed against θ_d^r and θ_c^r in Section 4.8.6.

Similarly, we characterize drivers' decision using an NL model. For single-homing, drivers first decide whether to enter the ride-hail market and then choose which platform to join. Let e_0 be the wage rate of the best alternative opportunity outside the ride-hail market, then the fleet size

[2]Although more sophisticated nested structures could be accommodated within our framework, we note that testing the suitability of these structures is a task beyond the scope of the present study.

of platform j (Eq. (4.5b)) is specified as

$$N^j = S_0 \frac{\exp(\theta_d I_d^r)}{\exp(\theta_d e_0) + \exp(\theta_d I_d^r)} \frac{\exp(\theta_d^r e^j)}{\exp(\theta_d^r I_d^r)}, \tag{4.17}$$

where the composite utility of ride-hail is

$$I_d^r = \frac{1}{\theta_d^r} \ln \left[\sum_{j \in \mathcal{P}} \exp(\theta_d^r e^j) \right]. \tag{4.18}$$

Again, θ_d and θ_d^r are the uncertainty parameters with respect to each level of the NL model.

For multi-homing, since each driver joins all platforms, the choice model reduces to an MNL model with two alternatives and the fleet size of all platforms is given by

$$N = N^j = S_0 \frac{\exp(\theta_d \bar{e})}{\exp(\theta_d e_0) + \exp(\theta_d \bar{e})}. \tag{4.19}$$

Along with the analysis in previous section, we examine how passenger wait times, market share and vacant vehicle density at the market equilibrium vary with there key inputs: the total demand D_0, the potential supply S_0 and the en-route detour $\tau_p - \tau_s$. Each market equilibrium, as well as the duopoly equilibrium D1 or D2, is solved multiple times with randomly selected initial solutions. We include a brief discussion of the existence of multiple equilibria in Section 4.8.7.

Figure 4.4 presents the main results of the sensitivity analysis. Here, we only plot the market share of one platform because the other one is the same due to symmetry. The vacant vehicle density under multi-homing is the total vacant vehicle divided by two, i.e., the expected vehicle supply available for one platform.

As the total demand increases (Figure 4.4(a)), the passenger wait time tends to increase and the e-hail's market share tends to shrink, primarily due to greater competition. The vacant vehicle density first rises with the demand, reaches a long stretch of plateau and then begins a slow decline. These general trends hold for all supply modes: monopoly, single-homing and multi-homing. A noticeable deviation is the wait times for pooling rides: they first decrease mildly as the demand

increases, thanks to the reduced pickup detour w_{p2}. However, when the demand continues to rise, the stronger inter-passenger competition pushes up the pickup time w_{p1} that eventually more than offsets the savings in w_{p2}. This pattern explains why market share of the pooling rides (the middle plot in Figure 4.4) first increases and then decreases as the demand goes up.

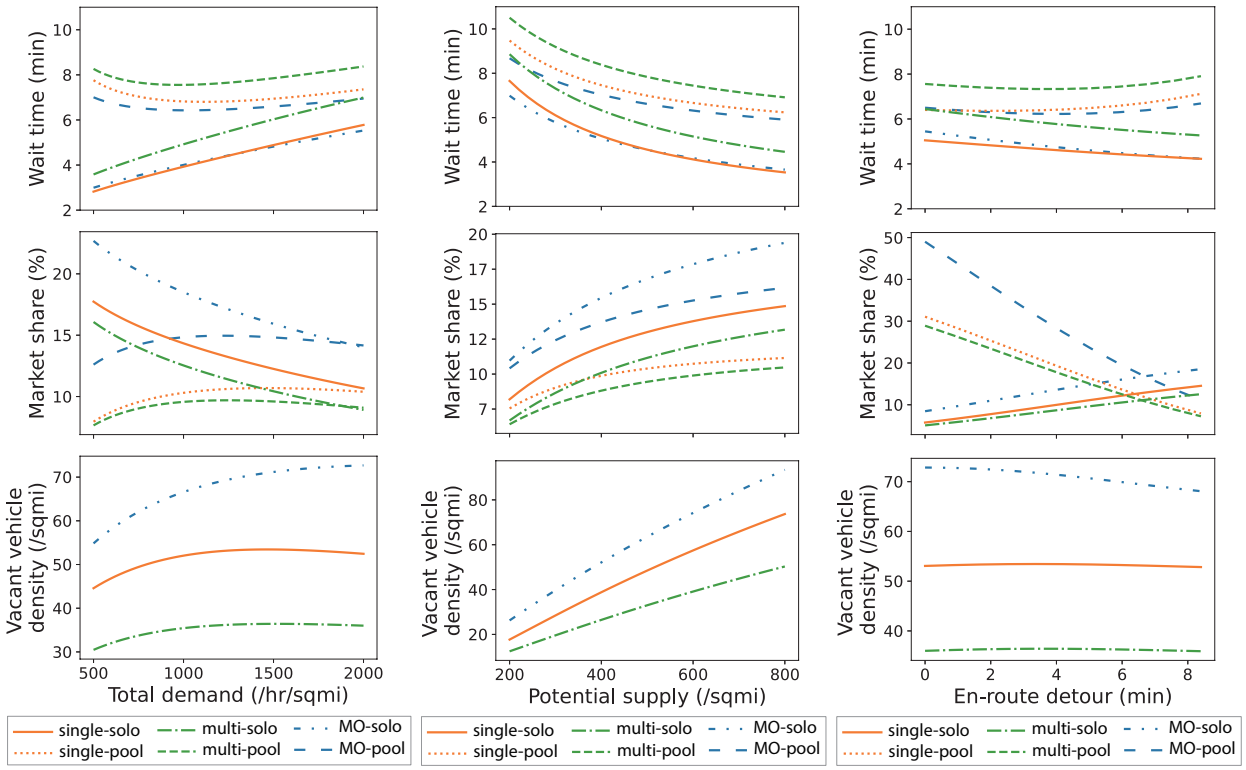


Figure 4.4. Sensitivity of passenger wait time, market share and vacant vehicle density to (a) total demand D_0, (b) potential supply S_0, and (c) en-route detour $\tau_p - \tau_s$.

Figure 4.4(b) shows that, as expected, when the potential supply increases, the wait time decreases, and both the market share of e-hail and the vacant vehicle density increases. The impact of en-route detour is more nuanced, as shown in Figure 4.4(c). This detour, which measures how much additional time pooling would add to the trip duration, has a dramatic effect on the market share of pooling rides, regardless of the supply mode. For every two minutes of additional detour, the market share roughly drops by 5%. Solo rides gain market share, but not enough to make up for the loss suffered by pooling rides. Interestingly, neither the wait time nor the vacant

vehicle density is much affected by the detour. For solo rides, the wait time slightly decreases, likely because the declining market share eases up competition. Yet, the wait time for the pooling rides eventually turns back up, thanks to the increase in w_{p2} caused by the loss in the market share.

We now examine the differences associated with the supply mode. In all cases, the monopoly has the highest share for both solo and pooling rides, followed by the single-homing and multi-homing duopoly. This is expected because a monopoly platform enjoys a higher market power. However, the total market shares in a duopoly market is higher than those in a monopoly. We caution that this result likely overestimates the appeal of the duopoly (for both single- and multi-homing modes) due to the IIA assumption. For the same reason, the fleet size under single-homing could be overestimated, leading to a shorter wait time compared to that under multi-homing.

The bottom row of Figure 4.4 shows the monopoly enjoys a much higher vacant vehicle density than the other two markets. However, its level of service (LOS) is only marginally better as measured by passenger wait time (the saving is less than 1 minute). This is primarily due to the higher demand rate, which intensifies the competition among waiting passengers and thus drags down the LOS. On the other hand, the wait time under multi-homing is consistently higher than that under single-homing, both for solo and pooling rides. This result could be explained by the much lower vacant vehicle density.

Although the change in market shares against the total demand and supply follows the same pattern under the two duopoly modes, multi-homing shows a higher sensitivity. Specifically, under multi-homing the market share of ride-hail decreases faster as total demand increases, and increases faster as potential supply increases, as shown in the middle subplots of Figures 4.4(a) and (b).

4.7. Summary

In this chapter, we established the market equilibrium of an aggregate ride-hail market with pooling service. The interactions between passengers and drivers are characterized based on the matching model presented in Chapter 3. We showed that, regardless of the number platforms in the market,

this system always has an equilibrium under mile conditions. Besides, the equilibrium state can be solved through fixed point iterations. For the case that multiple platforms competiting in the market, two supply modes are explicitly analyzed. One is single-homing, where each driver joins only one platform, and the other is multi-homing, where drivers join all platforms if they choose to enter the market. It is found this single difference leads to distinct matching mechanisms. Counterintuitively, multi-homing does not improve LOS. Although the passengers get access to a greater pool of vehicles, they also face a more intense competition against passengers from other platforms. This result is also demonstrated through numerical examples. Besides, numerical experiments are conducted to examine the sensitivity of market equilibrium under fixed pricing scheme. The main findings are summarized as follows:

- As expected, pooling is desirable when demand is high but supply is scarce. However, its benefit diminishes quickly as the average en-route detour time (i.e., the difference between the average duration of solo and pooling trips) increases. This also explains the low pooling ratio observed in practice. Therefore, keeping this value under control is the key to the success of pooling.
- The competition reduced the market share of each platform though increases the total demand for ride-hail. Similarly, the fleet size shrinks under platform competition. The declines in demand and supply are comparable in the case of single-homing thus the LOS is hardly affected. In contrast, platforms under multi-homing suffer from a greater loss of vehicle supply and thus observe a more considerable drop in LOS.

The equilibrium model built in this chapter serves an important foundation for the following analysis on the platform pricing and regulations in the context of aggregate ride-hail market, which will be studied in following two chapters.

4.8. Appendix

4.8.1. Notations

Table 4.1. List of notations

Variable	Description	Unit
w_e (w_s)	e-hail (solo) passenger wait time	hr
w_{p1}	first part of pooling passenger wait time (matching time plus pickup time of the first passenger)	hr
w_{p2}	second part of pooling passenger wait time (pickup time of the second passenger)	hr
Λ	vacant vehicle density	/sqmi
Π_s (Π_p)	solo (pooling) waiting passenger density	/sqmi
v	cruising speed of vacant vehicles	mph
k	coefficient of matching efficiency	/sqmi
b	coefficient of pooling efficiency	
κ	approximation parameter	/sqmi
δ	detour ratio of road network	
Q_s (Q_p)	solo (pooling) demand rate	/hr/sqmi
f_s (f_p, f_t)	trip fare of solo rides (pooling rides, transit)	$
τ_s (τ_p, τ_t)	travel time of solo rides (pooling rides, transit)	hr
u_s (u_p, u_t)	generalized cost of solo rides (pooling rides, transit)	$
ν	value of time	$/hr
$\theta_c(\theta_c^r)$	Mode choice uncertainty (among ride-hail options)	/$
$\theta_d(\theta_d^r)$	Market entry uncertainty (among ride-hail platforms)	/$
ζ	disutility factor of transit trips	$/hr
S_0	potential supply	/sqmi
N	fleet size (number of drivers in operation)	/sqmi
V	vacant vehicle time	hr/sqmi
$\tilde{e}_0$ (e_0)	random (average) reservation rate	$/hr
e	driver's earning rate	$/hr
η	compensation rate (payment per unit occupied time)	$/hr

4.8.2. Parameter setting

Table 4.2. Default values and ranges of parameters.

Parameter		**Unit**	**Default value**	**Variation**
Detour ratio of road network	δ		1.3	
Cruising speed	v	mph	13.6	
Matching efficiency	k	/sqmi	0.16	
Pooling efficiency	b		0.05	
Approximation parameter	κ		4	
Average solo trip duration	τ_s	hr	0.28	
Average pooling trip duration	τ_p	hr	0.40	
Average transit trip duration	τ_t	hr	0.53	
Passengers' value of time	ν	\$/hr	27.69	
Relative disutility rate of transit	ζ	\$/hr	6.92	
Mode choice uncertainty in monopoly	θ_c		1	
Mode choice uncertainty in duopoly	θ_c (θ_c^r)		0.5	
Market entry uncertainty in duopoly	θ_d (θ_d^r)		0.25	
Average reservation rate in monopoly	e_0	\$/hr	19.31	
Average reservation rate in duopoly	e_0	\$/hr	15	
Total demand	D_0	/sqmi/hr	1200	500–2000
Potential supply	S_0	/sqmi/hr	550	200–800
Solo trip fare	f_s	\$/ride	14	
Pooling trip fare	f_p	\$/ride	10	
Transit trip fare	f_t	\$/ride	2.69	
Compensation rate	η	\$/hr	20	
Additional pickup fare	c_p	\$/ride	0	0–2

4.8.3. Sensitivity of market equilibrium to en-route detour

As discussed in Section 4.2, the en-route detour $\tau_p - \tau_s$ is expected to decrease with pooling demand. Yet, in all numerical experiments presented in the main text, we have assumed the detour to be constant for simplicity. In this appendix, we adopt the results of Ke et al. [2021] and Lobel and Martin [2020] to test the sensitivity of our findings to the endogeneous en-route detour.

Ke et al. [2021] empirically observe the passenger detour distance Δl follows

$$\frac{\Delta l}{\bar{l}} = \frac{1}{\alpha N + \beta}, \tag{4.20}$$

where $\bar{l}$ is the average trip distance, N is the number of requests accumulated in a matching interval, and α, β are coefficients[3].

Since vehicles travel at a constant speed v, $\Delta l/l = (\tau_p - \tau_s)/\tau_s$. In addition, the batch demand N in Eq. (4.20) can be replaced with $b\Pi_p$, which is the unmatched pooling passenger density. Accordingly, we adjust the average trip duration of pooling rides as follows:

$$\tau_p = \tau_s \left(1 + \min\left(0.5, \frac{1}{\alpha b \Pi_p + \beta}\right)\right). \tag{4.21}$$

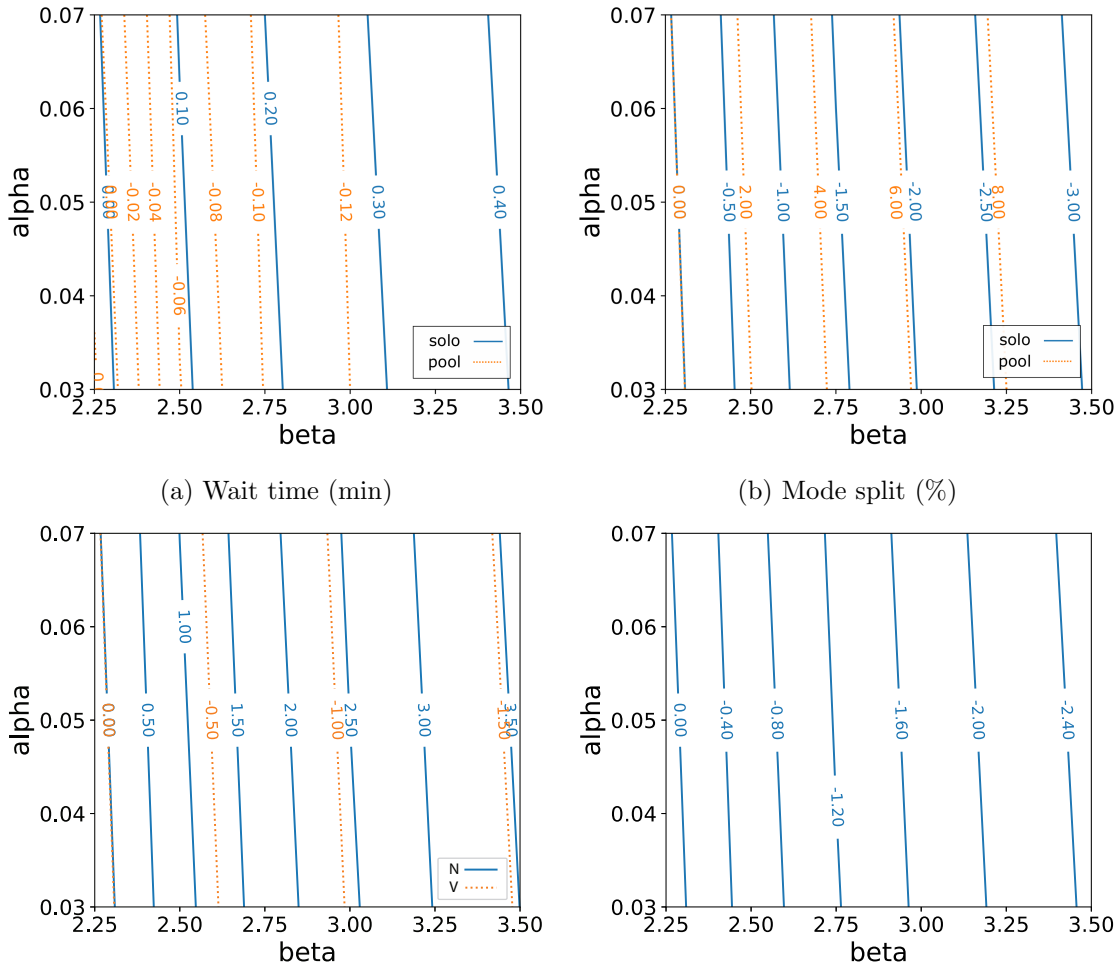
Note that the upper bound 0.5 on $1/(\alpha b\Pi_p + \beta)$ follows from the result proved in Lobel and Martin [2020].

Ke et al. [2021] calibrated the coefficients α, β for New York City and two other cities in China with various matching radius. We take the range of coefficient values associated with New York City and calculate the market equilibrium with default total demand ($D_0 = 1200$/hr/sqmi) and potential supply ($S_0 = 550$/sqmi).

Figure 4.5 shows how the key outputs of the equilibrium model vary with α and β. First, the market equilibrium is clearly insensitive to α, mainly because the first term $\alpha b\Pi_p$ is much smaller compared to β. In other words, the pooling demand has a minor impact on the en-route detour, as long as the pooling demand is not too large. The differences in passenger wait times and average trip duration of pooling rides are small (less 10%), and those in fleet size and vacant vehicle density are almost negligible (less than 1%). The mode split of pooling rides is the most sensitive to β. As shown in Figure 4.5(b), the share of pooling increases in the adjusted model by more than 8%, when β increases from 2.25 to 3.5. As a larger β corresponds to a smaller detour ratio, this result is expected.

We continue to examine the sensitivity of market equilibrium to different supply-demand levels using the adjusted model. Since the results are demonstrated insensitive to α, we fix it as 0.05 and consider two extreme cases of β, i.e., $\beta = 2.25$ and $\beta = 3.5$. The results, along with the market

[3]Here, we follow the same notations in Ke et al. [2021], hence there are a few conflicts with the notations in our main text



(a) Wait time (min)

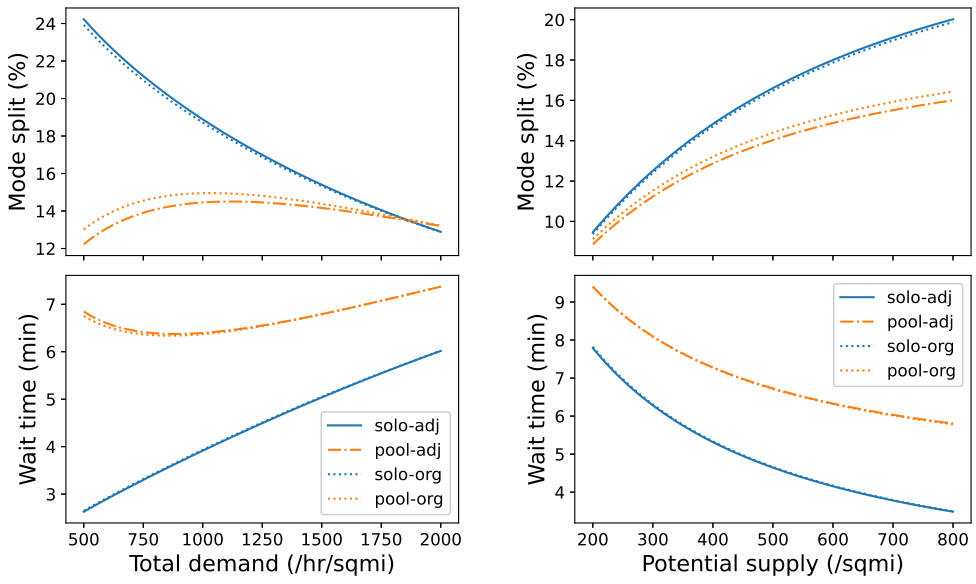
(b) Mode split (%)

(c) Fleet size and vacant vehicle density (/sqmi)

(d) Pooing trip duration (min)

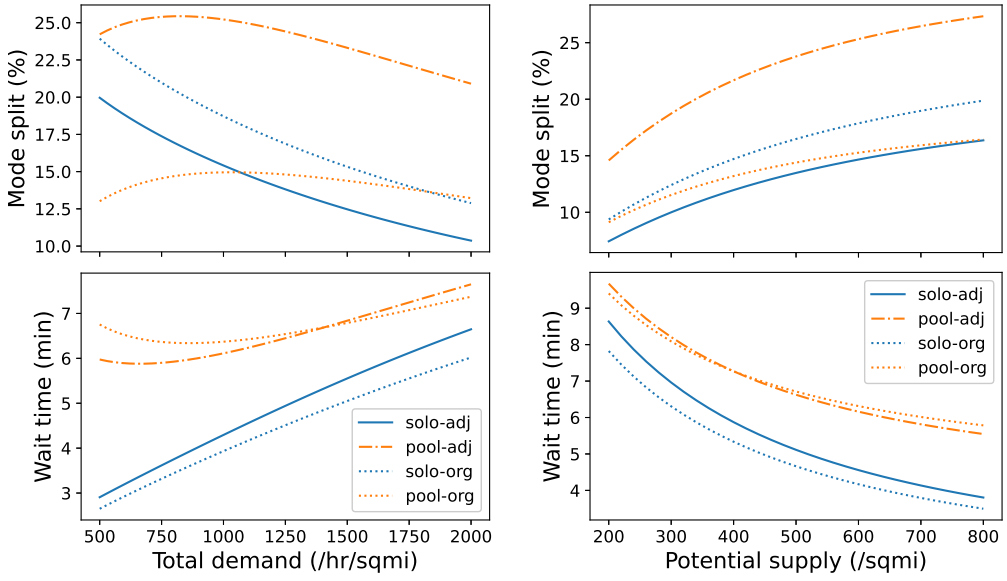
Figure 4.5. Differences in equilibrium outputs between the original model (with constant τ_p) and the adjusted model (with τ_p being specified by Eq.(4.21)).

equilibria with constant en-route detour, are illustrated in Figures 4.6 and 4.7. As expected, when $\beta = 2.25$, the adjusted model barely causes any meaningful changes in the equilibrium solution. The most prominent sensitivity is again shown in the mode split when β is large. Specifically, the difference in pooling share reaches about 10% when the total demand is low or the potential supply is high. The difference in solo share, on the other hand, is quite stable across different supply and demand levels, at around 3-5%. The differences in passenger wait time is even smaller, less than half of an minute in most scenarios.



(a) (b)

Figure 4.6. Sensitivity of market equilibrium to (a) total demand D_0 and (b) potential supply S_0 with $\alpha = 0.05, \beta = 2.25$.



(a) (b)

Figure 4.7. Sensitivity of market equilibrium to (a) total demand D_0 and (b) potential supply S_0 with $\alpha = 0.05, \beta = 3.5$.

4.8.4. Proof of Lemma 1

We apply the implicit function theorem [Krantz and Parks, 2012] to prove the result.

Consider a continuously differentiable function $L : \mathbb{R}^{n+m} \to \mathbb{R}^m$ and a point $(\mathbf{x}_0, \mathbf{y}_0), \mathbf{x}_0 \in \mathbb{R}^n, \mathbf{y}_0 \in \mathbb{R}^m$ such that $L(\mathbf{x}_0, \mathbf{y}_0) = 0$. The theorem states that, if the Jacobian matrix

$$J_{L,\mathbf{y}}(\mathbf{x}_0, \mathbf{y}_0) = \left[\frac{\partial L_i}{\partial y_j}(\mathbf{x}_0, \mathbf{y}_0)\right],\ i = 1, \ldots, m,\ j = 1, \ldots, m \tag{4.22}$$

is invertible, then there is a neighborhood of $\mathbf{x}_0$, denoted as $U \subset \mathbb{R}^n$, and a unique continuously differentiable function $g : U \to \mathbb{R}^m$ such that $\mathbf{y} = g(\mathbf{x}), \forall \mathbf{x} \in U$.

To apply the above result, let us rewrite Eq. (4.5b) as

$$L(\mathbf{x}, N) = S_0 G\left(\frac{1}{N}\left[\eta\left(Q_s(\mathbf{x})\tau_s + \frac{1}{2}Q_p(\mathbf{x})\tau_p\right) + \frac{c_p}{2}Q_p(\mathbf{x})\right]\right) - N = 0. \tag{4.23}$$

Therefore, for any point that satisfies Eq. (4.23), N is a continuous function of $\mathbf{x}$ in a neighborhood of that point provided $\frac{\partial L}{\partial N}$ is invertible, or equivalently,

$$\frac{\partial L}{\partial N} = -S_0 G' \frac{1}{N^2}\left[\eta\left(Q_s(\mathbf{x})\tau_s + \frac{1}{2}Q_p(\mathbf{x})\tau_p\right) + \frac{c_p}{2}Q_p(\mathbf{x})\right] - 1 \neq 0. \tag{4.24}$$

To see why (4.24) must hold, note that G' is the probability density function of the drivers' reservation rate. Thus, G', as well as all other variables, must be nonnegative. Accordingly, $\partial L/\partial N \leq -1$ and hence it cannot be zero. This completes the proof.

4.8.5. Determination of upper bounds forw_s, w_{p1} and w_{p2},

The MNL model implies that the mode share decreases exponentially with its general cost but never reaches zero. However, in reality no one would choose solo or pooling if the corresponding wait time (w_s and $w_{p1} + w_{p2}$) are too long. Let ε be the minimum demand considered to be meaningful for analysis. We next show the upper bound for w_{p2}, denoted as $\overline{w}_{p2}$, can be derived as a function

of ε. Note that

$$Q_p = D_0 \frac{e^{-\theta u_p}}{\sum_i e^{-\theta u_i}} \leq D_0 \frac{e^{-\theta u_p}}{2e^{-\theta \underline{u}} + e^{-\theta u_p}}, \tag{4.25}$$

where $\underline{u} = \min\{u_s, u_t\} = \min\{f_s + \nu(\underline{w}_s + \tau_s), f_t + \nu\tau_t\}$ (recall $\underline{w}_s = \frac{\delta}{2v\sqrt{S_0}}$; see the proof of Proposition 2). Setting

$$D_0 \frac{e^{-\theta u_p}}{2e^{-\theta \underline{u}} + e^{-\theta u_p}} \leq \varepsilon$$

yields

$$\begin{aligned} \left(\frac{D_0}{\varepsilon} - 1\right) e^{-\theta u_p} \leq 2e^{-\theta \underline{u}} \quad &\Rightarrow \quad u_p \geq \underline{u} + \frac{1}{\theta} \log \frac{D_0/\varepsilon - 1}{2} \\ &\Rightarrow \quad w_{p2} \geq \frac{1}{\nu}(\underline{u} - f_p) + \frac{1}{\theta\nu} \log \frac{D_0/\varepsilon - 1}{2} - w_{p1}. \end{aligned} \tag{4.26a}$$

In other words, whenever Eq. (4.26a) is satisfied, the demand for pooling would reduce to no more than ε. Since $w_{p1} \geq 0$, we may set

$$\overline{w}_{p2} \equiv \frac{1}{\nu}(\underline{u} - f_p) + \frac{1}{\theta\nu} \log \frac{D_0/\varepsilon - 1}{2}. \tag{4.27}$$

It is clear that the upper bound established this way is likely to be loose. However, it suffices for our purpose to show a finite upper bound does exist. $\overline{w}_s$ and $\overline{w}_p$ can be obtained in a similar fashion, and the details are omitted for brevity.

4.8.6. Impact of IIA assumption

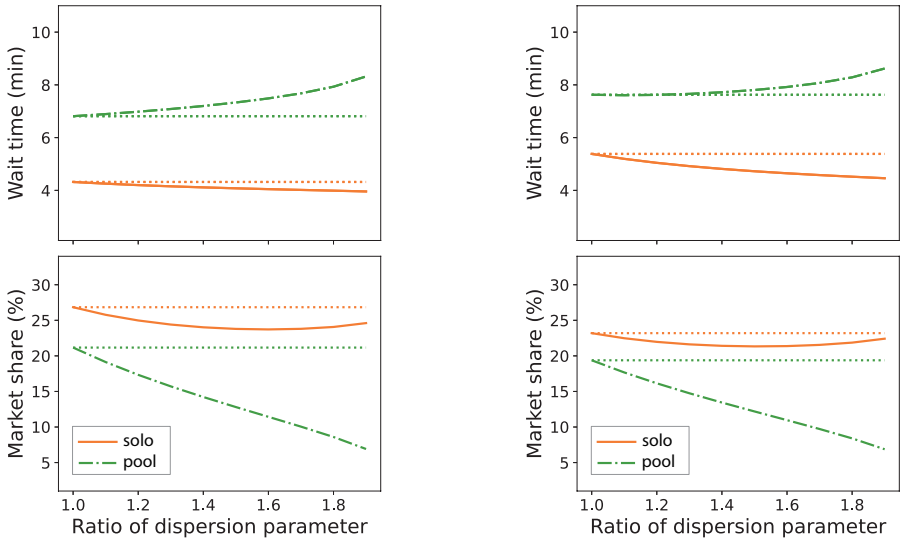
To relax the IIA assumption, we simply allow in our model the ratio of θ_c^r/θ_c to increase from 1 to 2. In the single-homing duopoly, we also similarly vary the ratio θ_d^r/θ_d. For simplicity, we assume the two ratios in this case are the same, referred to as the "ratio of dispersion parameter" in Figure 4.8.

As expected, when θ_c^r/θ_c increases—meaning the four options are increasingly viewed as good substitutes for each other—the total market share of ride-hail dips. In the case of single-homing,

the total share (including both solo and pooling) decreases from about 50% to a little more than 30% as θ_c^r/θ_c doubles. It is the pooling rides, however, that is mostly affected. Indeed, the vast majority of the losses is inflicted on pooling rides, while the market share of solo rides is largely intact. The drop in the market share slightly improves the wait time for solo rides, thanks to less inter-passenger competition. However, LOS degrades for pooling rides, likely because the collapsing market share significantly prolongs the detour time w_{p2}.

Overall, the above finding seems to indicate that the IIA assumption could result in overly optimistic estimation about the potential of pooling rides. However, it is worth noting that the simple setup of our NL model may treat pooling rides unfairly. Intuitively, pooling and solo rides provided by the same platform should be viewed as much more distinctive than pooling rides provided by two different platforms. Yet, they are all lumped into the same nest in the current setting, without recognizing this distinction. Addressing this issue requires a nested structure with two layers, allowing passengers to first choose a mode before choosing a platform (or vice versa). However, without data needed to calibrate such an NL model, we would be guessing the actual specification of the NL model (hence the results) one way or the other. Thus, it seems appropriate to leave a more sophisticated choice modeling to a future study.

The reader should be aware that the findings made with the IIA assumption in the following sections are subject to the biases that comes with it, especially those related to the comparison of the total market share between the monopoly and the duopoly. Most sensitivity results, however, may not be affected by this bias. In particular, Figure 4.8 suggests the impact of IIA is insensitive to the supply mode (note that the patterns of single- and multi-homing results are very similar). In other words, most findings about the relative performance between the two supply modes should stand free of the assumption.



(a) Single-homing (b) Multi-homing

Figure 4.8. Sensitivity of passenger wait time and market share to the ratio of dispersion parameter ($\theta_c^r/\theta_c = \theta_d^r/\theta_d$). Market shares for single- and multi-homing cases are the sum of the two platforms' market shares. Other parameters are set as the default values in Table 4.2.

4.8.7. Existence of multiple equilibria

In all numerical experiments, we solve the duopoly equilibrium D1 or D2 multiple times with randomly selected initial solutions. Yet, in our experiments, we do find multiple equilibria for different initial solutions. To illustrate these equilibria, we solve the unregulated duopoly game with symmetric platforms using default parameter values reported in Table 4.2 and various combination of initial prices. Specifically, we fix the initial values of f_s and η, while varying the values of f_p for the two platforms.

Figure 4.9 plots the region of attraction (ROA) under the single-homing supply mode[4]. In each plot, ROA is colored according to the type of solutions. It is worth emphasizing that, *for each type, there is always a unique equilibrium.* Thus, once the initial solution falls into an ROA, the outcome of the game will be "attracted" to that equilibrium associated with that ROA. In all cases, only four ROAs are identified and they can be characterized based on the market share of the pooling

[4]The results for multi-homing supply are very similar and thus are omitted here for brevity.

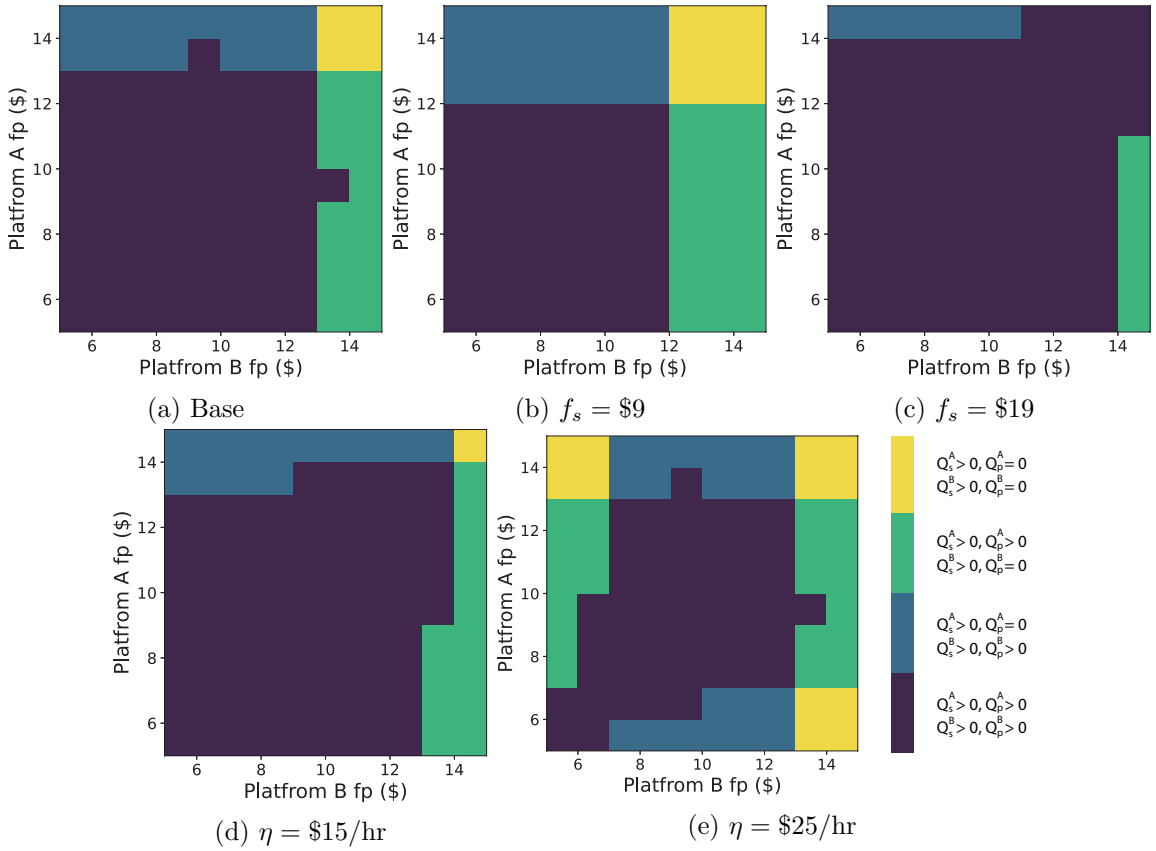


Figure 4.9. Equilibrium solutions of unregulated duopoly game under single-homing supply mode. Each grid represents an initial solution defined by f_p^A and f_p^B, rounded to full dollar.

rides at the corresponding equilibrium: (i) $Q_p^A = Q_p^B = 0$; (ii) $Q_p^A > 0$ and $Q_p^B = 0$; (iii) $Q_p^A = 0$ and $Q_p^B > 0$ and (iv) $Q_p^A > 0$ and $Q_p^B > 0$.

As expected, if a platform initially prices the pooling rides too high (roughly \$13 in the base case; see Figure 4.9(a)), it risks losing all pooling market share as the game evolves. This threshold varies with the solo fare f_s and the compensation rate η. When f_s is lower than the default value (Figure 4.9(b)), the threshold drops to \$12; and when f_s increases to \$19 (Figure 4.9(b)), it rises to \$14. In addition, a high initial compensation rate η attracts a large fleet size and thus the platforms prefer serving more solo rides. Accordingly, in this case (Figure 4.9(e)), it is much more likely for the game to reach an equilibrium where at least one platform completely eliminates pooling rides. Because pooling is one of our primary interests, we will only focus on the equilibrium pattern (iv)

above, i.e. $Q_p^A > 0$ and $Q_p^B > 0$. It is worth emphasizing again that, in all experiments conducted, such a solution is always unique.

CHAPTER 5

Monopoly pricing in aggregate market

Given the equilibrium established in Chapter 4, we consider the market is dominated by a single platform. For the platform, pooling helps increase the service capacity without hiring more drivers. Also, as drivers are paid by their occupied time, the platform could generate a greater profit by serving more pooling rides. However, the detour in pooling trips compromises the level-of-service (LOS) and makes pooling less attractive. Meanwhile, it also consumes extra vehicle time. Therefore, the platform needs to strategically design the pricing scheme to achieve the desired pooling ratio. The regulations imposed on the ride-hail operations also affect the platform's operational strategies, as well as the behaviors of passengers and drivers in the market. This chapter is thus dedicated to investigating the trade-off of platforms in serving pooling trips, with and without regulatory constraints.

5.1. Optimal pricing of a monopolized platform

5.1.1. Profit-maximization pricing without regulation

By choosing a combination of f_s, f_p and η, the platform could control the demand split between solo and pooling rides, thereby the pooling ratio, to maximize its profit. If pooling is not profitable, the platform can simply set its price equal to or higher than solo rides (i.e., $f_s \leq f_p$) to eliminate pooling. On the other hand, the platform may increase the gap between f_s and f_p to encourage pooling.

Without loss of generality, we assume the transit fare f_t is fixed, and the platform aims to maximize its gross profit by choosing a price vector $\mathbf{y} = (f_s, f_p, \eta)^T$. The pricing problem is then

formulated as

$$\text{(M1)} \quad \max_{\mathbf{y}} \; R = f_s Q_s + f_p Q_p - \eta \left(Q_s \tau_s + \frac{1}{2} Q_p \tau_p \right). \tag{5.1}$$

Here, the platform's gross profit R equals the total revenue less the expense directly related to the production of trips. Specifically, the first two terms (i.e., $f_s Q_s$ and $f_p Q_p$) are revenues generated from solo and pooling rides, respectively, while $\eta \left(Q_s \tau_s + \frac{1}{2} Q_p \tau_p \right)$ is the payment to drivers based on their occupied time. In particular, Q_s and Q_p are derived from the market equilibrium induced by certain pricing strategy $\mathbf{y}$.

To simplify the notation, let $Q_s^{(1)}$, $Q_s^{(2)}$ and $Q_s^{(3)}$ denote the partial derivative of Q_s with respect to f_s, f_p, η, respectively. $Q_p^{(1)}$, $Q_p^{(2)}$ and $Q_p^{(3)}$ are defined similarly. Accordingly, the first-order conditions of M1 are reduced to

$$f_s = \eta \tau_s - \frac{Q_s Q_p^{(2)} - Q_p Q_p^{(1)}}{Q_s^{(1)} Q_p^{(2)} - Q_s^{(2)} Q_p^{(1)}}; \tag{5.2a}$$

$$f_p = \frac{1}{2} \eta \tau_p - \frac{Q_p Q_s^{(1)} - Q_s Q_s^{(2)}}{Q_s^{(1)} Q_p^{(2)} - Q_s^{(2)} Q_p^{(1)}}; \tag{5.2b}$$

$$Q_s \left[1 + \left(\tau_s + \frac{1}{2} \frac{Q_p}{Q_s} \tau_p \right) \frac{Q_s^{(1)}}{Q_s^{(3)}} \right] = 0; \tag{5.2c}$$

$$Q_p \left[1 + \left(\frac{Q_s}{Q_p} \tau_s + \frac{1}{2} \tau_p \right) \frac{Q_p^{(2)}}{Q_p^{(3)}} \right] = 0. \tag{5.2d}$$

Eqs. (5.2a) and (5.2b) bear similarity with the Lerner formula [Lerner, 1934], where the first term represents the marginal cost of each solo (pooling) ride (i.e., compensation paid to the driver) and the second term is a mark-up that measures the market power of the platform. It also aligns with the optimal trip fare derived in Zha et al. [2016] and Ke et al. [2020a]. Eqs. (5.2c) and (5.2d) imply that, at the optimal solution, either solo (pooling) demand equals 0 or $Q_s^{(1)}/Q_s^{(3)}$ ($Q_p^{(2)}/Q_p^{(3)}$) is dictated by the market shares and average journey times of the two modes.

For comparison, we derive the first-order conditions under single-mode operation, i.e., when only solo or pooling rides are served. In the case of pure solo rides, the single-mode operation

yields

$$f_s = \eta\tau_s - \frac{Q_s}{Q_s^{(1)}}; \tag{5.3a}$$

$$Q_s\left(1 + \tau_s\frac{Q_s^{(1)}}{Q_s^{(3)}}\right) = 0, \tag{5.3b}$$

and for pure pooling rides,

$$f_p = \frac{1}{2}\eta\tau_p - \frac{Q_p}{Q_p^{(2)}}; \tag{5.4a}$$

$$Q_p\left(1 + \frac{1}{2}\tau_p\frac{Q_p^{(2)}}{Q_p^{(3)}}\right) = 0. \tag{5.4b}$$

Eqs. (5.3) and (5.4) share the same structures as Eq. (5.2)—the optimal trip fare is the marginal cost plus the platform's mark-up, and the marginal changes of demand with respect to trip fare and compensation rate should be a constant when demand is positive.

Eqs. (5.3b) and (5.4b) further imply that, at the optimal solution with positive demand, the marginal change of demand due to an increase in trip fare should be proportional to that due to an increase in compensation rate. Specifically, the rate is $-1/\tau_s$ for solo and $-2/\tau_p$ for pool. In other words, in response to an increased trip fare, the platform must raise the compensation rate to improve LOS so that the service remains attractive to passengers. In the mix-mode scenario, however, the pressure to raise LOS is relieved, i.e., the absolute value of the rate is smaller in Eqs. (5.2c) and (5.2d). Because increasing the trip fare of one mode would make the other more attractive, the platform need not increase the compensation rate as much as in the single mode operation in order to hold on to the market share.

We proceed to compare the platform's market power under single and mixed operation modes. Dividing both the numerator and the denominator of the second term in Eq. (5.2a) by $Q_p^{(2)}$ yields $\frac{Q_s - Q_p Q_p^{(1)}/Q_p^{(2)}}{Q_s^{(1)} - Q_s^{(2)} Q_p^{(1)}/Q_p^{(2)}}$. Here, $Q_p^{(1)}/Q_p^{(2)} < 0$ as f_s and f_p have opposite influences on Q_p, and $Q_s^{(2)} > 0$ because increasing f_p makes solo rides more appealing. Also, $|Q_s^{(1)}| > |Q_s^{(2)}|$ and $|Q_p^{(2)}| > |Q_p^{(1)}|$

because a change in the trip fare of one mode has larger impact on that mode than the other mode. These observations lead to

$$\left|\frac{Q_s - Q_p Q_p^{(1)}/Q_p^{(2)}}{Q_s^{(1)} - Q_s^{(2)} Q_p^{(1)}/Q_p^{(2)}}\right| > \left|\frac{Q_s}{Q_s^{(1)}}\right|,$$

indicating that serving both pooling and solo rides has the potential to boost the platform's market power (hence profit) compared to the case when only solo rides are served. However, the change of total trips served by the platform is unclear, because the impact on pooling rides is unclear.

5.1.2. Profit-maximization pricing under regulations

Mathematically, if the regulation is imposed on certain equilibrium variables (e.g., drivers' average earning rate), the effect can be captured by adding an inequality constraint to the original problem, leading to

$$\begin{aligned} \text{(M2)} \quad & \max_{\mathbf{y}} \; R = f_s Q_s + f_p Q_p - \eta \left(Q_s \tau_s + \frac{1}{2} Q_p \tau_p \right), \\ & s.t. \quad h(\mathbf{y}) \leq 0, . \end{aligned} \tag{5.5}$$

Here, we consider the minimum wage policy (with a wage floor $\underline{e}$), the constraint $h(\mathbf{y})$ is given by

$$\underline{e} - \frac{\eta}{N} \left(Q_s \tau_s + \frac{1}{2} Q_p \tau_p \right) \leq 0 \, . \tag{5.6}$$

Accordingly, the Lagrangian of M2 is equivalent to

$$\mathcal{L} = f_s Q_s + f_p Q_p - \left(1 - \frac{\lambda}{N}\right) \left[\eta \left(Q_s \tau_s + \frac{1}{2} Q_p \tau_p \right) \right], \tag{5.7}$$

and the first-order conditions are reduced to

$$f_s = \left(1 - \frac{\lambda}{N}\right) \eta \tau_s - \frac{Q_s Q_p^{(2)} - Q_p Q_p^{(1)}}{Q_s^{(1)} Q_p^{(2)} - Q_s^{(2)} Q_p^{(1)}} \tag{5.8a}$$

$$+\frac{\lambda}{N^2}\left[\eta\left(Q_s\tau_s+\frac{1}{2}Q_p\tau_p\right)\right]\frac{N^{(1)}Q_p^{(2)}-N^{(2)}Q_p^{(1)}}{Q_s^{(1)}Q_p^{(2)}-Q_s^{(2)}Q_p^{(1)}};$$

$$f_p=\left(1-\frac{\lambda}{N}\right)\left(\frac{1}{2}\eta\tau_p+\frac{1}{2}c_p\right)-\frac{Q_pQ_s^{(1)}-Q_sQ_s^{(2)}}{Q_s^{(1)}Q_p^{(2)}-Q_s^{(2)}Q_p^{(1)}} \tag{5.8b}$$

$$+\frac{\lambda}{N^2}\left[\eta\left(Q_s\tau_s+\frac{1}{2}Q_p\tau_p\right)\right]\frac{N^{(2)}Q_s^{(1)}-N^{(1)}Q_s^{(2)}}{Q_s^{(1)}Q_p^{(2)}-Q_s^{(2)}Q_p^{(1)}};$$

$$Q_s\left[1+\left(\tau_s+\frac{1}{2}\frac{Q_p}{Q_s}\tau_p\right)\frac{Q_s^{(1)}}{Q_s^{(3)}}\right]=0; \tag{5.8c}$$

$$Q_p\left[1+\left(\frac{Q_s}{Q_p}\tau_s+\frac{1}{2}\tau_p\right)\frac{Q_p^{(2)}}{Q_p^{(3)}}\right]=0, \tag{5.8d}$$

where $N^{(1)}$ and $N^{(2)}$ denote the partial derivatives of the total vehicle supply to f_s and f_p, respectively.

On the one hand, Eqs. (5.8c) and (5.8d) suggest that the regulation does not affect the optimal conditions regarding the compensation rate. On the other hand, comparing Eq. (5.8a) with Eq. (5.2a) reveals how minimum wage affects the optimal ride price: the first term is discounted by a factor $1-\lambda/N$, and the newly added third term tends to reduce the platform's market power, as $N^{(1)}$ and $N^{(2)}$ are positive in general. Hence, the platform's profit is likely to suffer under the minimum wage policy. However, the actual change in the ride price is not clear, because it also depends on η.

Another type of regulation directly change the market input and thus is easier to integrate into the model. For instance, a few U.S. cities have issued a congestion tax on TNC trips starting and/or ending in designated areas during peak times[1]. In order to encourage ride-sharing, the proposed tax is lower for a pooling ride than a solo one. Since such a tax is likely to be fully passed to passengers, we assume a constant *extra* tax c_s is charged on each solo ride while pooling rides

[1]See an example of the City of Chicago`https://www.chicago.gov/city/en/depts/bacp/supp_info/city_of_chicago_congestion_pricing.html`

are free of extra charge. Accordingly, the generalized cost of solo trips Eq. (4.1a) becomes

$$u_s = f_s + \nu(w_s + \tau_s) + c_s. \tag{5.9}$$

Since the congestion tax does not introduce a new constraint, it should have less impact on the platform's pricing strategies than minimum wage discussed above. However, a larger shift toward pooling is expected, because the congestion tax is precisely levied against solo rides.

5.1.3. Social optimal pricing

We finally consider a second-best social optimal pricing problem that seeks to maximize social welfare. The policy is "second-best" because, while the goal is to maximize social welfare, the platform is not allowed to run a deficit. The corresponding optimization problem is formulated as follows:

$$\begin{aligned} \text{(M3)} \quad \max_{\mathbf{y}} W = {} & D_0\bar{u} + f_s Q_s + f_p Q_p \\ & - \int_0^N g^{-1}(n/S_0)\mathrm{d}n - c_0 N + c_s Q_s \\ s.t.\ & R \geq 0. \end{aligned} \tag{5.10}$$

The social welfare W consists of five parts: (i) the surplus of passengers, measured by the total expected general cost saving because of switching from transit to ride-hail services, where $\bar{u}$ denotes the average saving of each passenger and will be specified in Section 5.3; (ii) the total platform revenue $f_s Q_s + f_p Q_p$; (iii) the opportunity cost of drivers $\int_0^N g^{-1}(n/S_0)\mathrm{d}n$, where $g^{-1}(\cdot)$ is the inverse function of the CDF of the reservation rate; (iv) the approximate congestion cost caused by the ride-hail fleet, where c_0 is a constant cost caused by the entry of each driver (see Section 10.2 for the estimation of c_0); and (v) the tax revenue due to the congestion tax $c_s Q_s$, if implemented.

It is worthwhile to clarify several issues before we continue the analysis:

(1) Ideally, capturing congestion externality of ride-hail requires an explicit traffic flow model. However, since the main focus of the current analysis is the trade-off of pricing between solo and pooling trips, we simply assume each additional vehicle's contribution to congestion (i.e., the marginal cost) is a constant. In Chapter 7, we will deliver a more thorough analysis of the congestion impact of ride-hail service.

(2) The congestion tax is a transfer payment within the system and hence should neither increase nor decrease the social welfare. Accordingly, the tax revenue (i.e., item (v)) is included in the social welfare to offset the reduction in the passenger surplus of solo rides.

(3) According to Assumption 4, the decisions related to ride-hail (pricing and regulations) do not affect transit operations and the transit system always breaks even. Therefore, neither the supply cost nor the revenue of the transit service is included in the social welfare.

5.2. Solution existence and algorithm

The existence of solution to the program M1 is easily proved by evoking Theorem 1 in Harker and Pang [1988]. Consider the problem

$$\begin{aligned} \min_{x} \quad & f(x,y), \\ s.t. \quad & x \in X,\ y \in Y(x), \end{aligned} \tag{5.11}$$

where f is a continuous function.

The theorem states that there exists an optimal solution to the problem if the following two assumptions are satisfied:

(1) The feasible set is nonempty and closed.

(2) There exists a scalar $\alpha > 0$ and feasible solution (u, v) with $||(u, v)|| \leq \alpha$ such that $f(x, y) \geq f(y, v)$ for all feasible solution (x, y) with $||(x, y)|| \geq \alpha$.

The existence of market equilibrium established by Proposition 2 ensures the first assumption. Conceptually, the second assumption requires the candidate solutions to be contained in a compact set. This is naturally satisfied in the optimal pricing problem because a substantially large trip fare will push the demand to zero. Hence, M1 must have an optimal solution. As for the constrained programs M2 and M3, the solution exists whenever the feasible set is not empty.

Although M1 is in general non-convex, the objective function is continuously differentiable. Hence, we implement the gradient ascent algorithm to search local optimal solutions. Given the current solution and corresponding market equilibrium in each iteration, we compute the gradient and update the solution with the following iteration rule:

$$\mathbf{y}^{(i+1)} = \mathbf{y}^{(i)} + \alpha \nabla R(\mathbf{y}^{(i)}), \tag{5.12}$$

where $\nabla R = \left[\frac{\partial R}{\partial f_s}, \frac{\partial R}{\partial f_p}, \frac{\partial R}{\partial \eta}\right]^T$ is the gradient of the revenue function with respect to $\mathbf{y}$ and α is a predefined step size.

Evaluating ∇R, however, is challenging as it requires to differentiate the current market equilibrium. A detailed derivation is provided in Section 5.5.3. It is worth noting that the method bears some similarities with the sensitivity-analysis-based algorithm for network design problems [e.g. Tobin and Friesz, 1988, Yang, 1995, Patriksson, 2004], that is, the movement towards the next solution is guided by the sensitivity of the equilibrium solution with the current solution.

To solve the constrained program M2 and M3, we implement the dual gradient ascent algorithm. Take M2 as an example. The dual problem is written as

$$\text{(M2')} \quad \min_{\lambda} \max_{\mathbf{y}} R(\mathbf{y}) - \lambda h(\mathbf{y}) \tag{5.13}$$

$$s.t.\ \lambda \geq 0,$$

where λ is the Lagrangian multiplier.

The problem is then solved by the following iterative rule:

$$\mathbf{y}^{(i+1)} = \arg\max\ R(\mathbf{y}) - \lambda^{(i)} h(\mathbf{y}), \tag{5.14a}$$

$$\lambda^{(i+1)} = \max\left(0,\ \lambda^{(i)} + \rho h(\mathbf{y}^{(i)})\right), \tag{5.14b}$$

where ρ is a constant penalty parameter. In each iteration, the maximization problem (5.14a) is first solved in the same way as M1, with the current estimate of the multiplier λ. Then (5.14b) is invoked to update λ.

5.3. Numerical experiments

In this section, we report the findings of numerical experiments on a monopoly e-hail market with pooling service. The demand and supply modes, as well as other parameters, are set in the same way as Section 4.6. Specifically, the passengers' mode choice is characterized by the Multinomial Logit (MNL) model and the driver's reservation rate is assumed to follows a uniform distribution in the range $[0, 2e_0]$, where e_0 is the average reservation rate. It is well known that in the MNL model the logsum term measures the expected utility of all alternatives [Small and Rosen, 1981, De Jong et al., 2007]. Since we treat transit as a benchmark mode, the expected general cost saving Δu in Eq. (5.10) is given as

$$\Delta u = \frac{1}{\theta_c} \log \sum_{i \in \mathcal{M}} \exp[\theta_c (u_t - u_i)],\ i \in \mathcal{M}. \tag{5.15}$$

Besides, the drivers' opportunity cost is reduced to

$$\int_0^N g^{-1}(n/S_0)\mathrm{d}n = \frac{\eta}{2}\left(Q_s \tau_s + \frac{1}{2} Q_p \tau_p\right) = \frac{e_0 N^2}{S_0}. \tag{5.16}$$

The default values and ranges of input parameters used in the following experiment are the same as those in Section 4.6, which are summarized in Table 4.2. The values of parameters newly defined in this chapter are reported in Table 5.2.

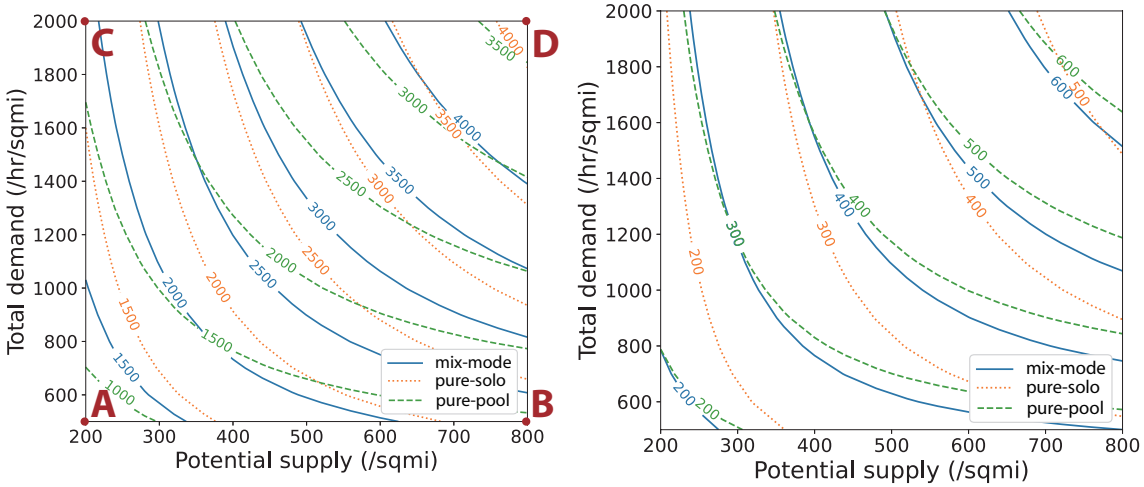
5.3.1. Performance of optimal pricing

We first examine the performance of the profit-maximization pricing problem M1 under various operational strategies and market conditions (i.e., combinations of total demand D_0 and potential supply S_0). The platform may use one of the following three operational strategies: (i) offering both solo and pooling rides, (ii) only offering solo rides, and (iii) only offering pooling rides. These strategies are referred to as "mix-mode", "pure-solo" and "pure-pool", respectively.

Figures 5.1 presents contours of the platform's profit and total trip output (i.e., $Q_s + Q_p$) gained by M1 under the three operational strategies. It can be read from Figures 5.1(a) that mix-mode always achieves the highest profit, followed by pure-solo strategy. This finding is predicted by the analytical solutions, as a platform serving both solo and pooling rides enjoys a higher market power (see Eqs. (5.2a) and (5.2b)). As expected, pure-pool is the least profitable because the platform has to keep the price sufficiently low to sustain a reasonable level of service for pooling (otherwise it would lose much of the market share to transit). On the other hand, mix-mode and pure-pool produce much more trips than pure-solo, and the difference increases as the market expands; see Figures 5.1(b). When the potential supply becomes scarce, pure-pool gradually achieves a leading position in terms of trip production. Overall, mix-mode achieves a more favorable balance between trip production and profitability than the other two strategies.

We further pick four corner cases (labeled in Figure 5.1(a)), namely, low-demand-and-low-supply (A: "low-low"), low-demand-and-high-supply, (B: "low-high"), high-demand-and-low-supply (C: "high-low"), and high-demand-and-high-supply (D: "high-high"), and examine more details of the system performance. Besides the three operational strategies, we also solve the second-best pricing problem (P3) under mix-mode, denoted as "SO", for comparison.

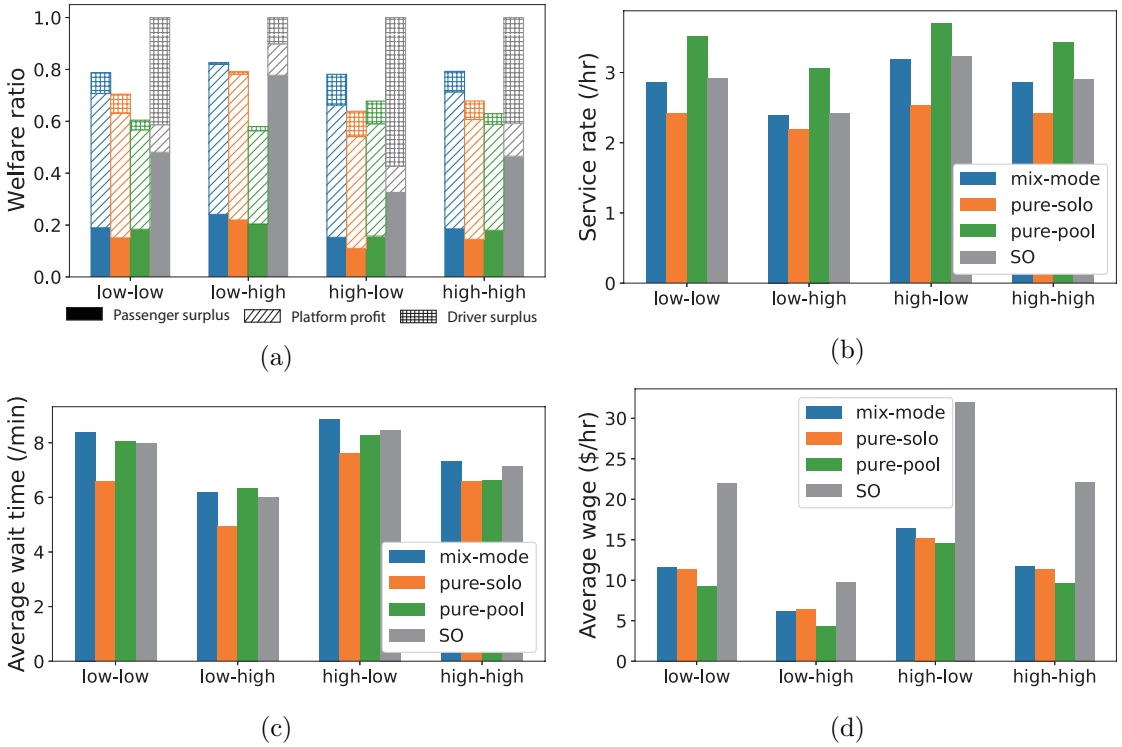
Figure 5.2(a) compares the social welfare obtained by the three operational strategies with the social optimal (SO) state. The different patterns indicate the passenger surplus (logsum term in Eq. (5.10)), the platform profit and the driver surplus (income less opportunity and congestion cost). We plot the ratio of these three components in each market scenario with respect to the



(a) Platform's profit

(b) Trip production

Figure 5.1. System performance under different operational strategies and market conditions.



(a)

(b)

(c)

(d)

Figure 5.2. System performance under different operational strategies and representative market conditions.

total welfare at SO (normalized as one). We first note that mix-mode consistently yields the highest social welfare, around 80% of the SO level. Pure-pool is the worst in three out of the four corner cases. It slightly outperforms pure-solo only when demand is high but supply is low (Case C: high-low), which is expected because pooling is most helpful in such a case. Combined with the results from Figure 5.1, we conclude that *serving a mixture of solo and pooling rides benefit both the platform and the society.* On the other hand, only serving pooling rides does not necessarily yield a higher social welfare than the regular ride-hail service, because the passengers suffer from a degraded level of service and the drivers earn a lower wage.

Under profit-maximizing pricing, the platform's profit dominates the total social welfare; see Figure 5.2(a). In contrast, the SO pricing clearly prioritizes the surplus of passengers and drivers. Interestingly, SO did not wipe out the platform's profit completely (even it is allowed to do so). This suggests that the only constraint in M3 (that the profit should not be negative) is inactive at the optimum. In other words, the solution obtained here is indeed a true SO, rather than a second-best optimum. Incidentally, this finding also implies that no subsidy is needed to sustain an SO solution, which seems at odds with previous findings about taxi markets [Douglas, 1972, Arnott, 1996, e.g.,]. Upon further inspection, we note that the discrepancy arises from the externality term (congestion cost $c_0 N$) included in the objective function. When that term is removed, a negative profit would indeed show up and the profitability constraint would be activated. Effectively, the congestion externality discourages over-supply in the ride-hail market.

Figure 5.2(b) examines the service rate, defined as the ratio between trip production and fleet size, i.e., $(Q_s + Q_p)/N$. As expected, pure-pool achieves the highest service rate while pure-solo has the lowest. The service rate under mix-mode is almost the same as that at SO, which suggests a profit-maximizing platform serving both solo and pooling rides could operate at a socially optimal service rate.

Figures 5.2(c) and (d) present, respectively, the average passenger wait time for ride-hail services (weighted by solo and pooling demand) and the average driver wage. In all cases, pure-solo offers

the shortest wait time because it serves fewer passengers with a relatively large fleet of drivers. Also, the wait time is the highest in the case of high-demand-and-low-supply (high-low) and the lowest in the case of low-demand-and-high-supply (low-high). Except for the low-high case, mix-mode always leads to the longest average passenger wait time but the highest driver wage. Yet, the difference from the second place is minor (less than half minus in wait time and $1/hr in earning rate). Pure-pool is the worst strategy for driver income in all cases. At SO, however, the drivers are treated much better than all operational strategies under profit-maximizing pricing. Their hourly income more than doubled in some case (e.g., high-low). This is clearly linked to higher driver surplus of SO solution shown in Figure 5.2(a).

In summary, the mix-mode strategy seems the ideal choice for a profit-maximizing platform. Compared to the other two strategies, it maximizes both profit and social welfare, and brings greater benefits to passengers and drivers. Interestingly, although a profit-maximizing platform would not operate at the socially optimal scale, it tends to achieves a service rate (the number of trips served per vehicle) close to the SO level.

5.3.2. Impact of regulations

5.3.2.1. Minimum wage. To assess the impact of the minimum wage policy, we first solve M3 to obtain the "socially optimal" earning rates. These rates are then imported in M2 to derive a profit-maximizing platform's pricing strategy with an SO minimum wage constraint. A moment of reflection suggests that the policy would encourage more drivers to enter the service, which, in turn, attracts more passengers and boosts trip production. Indeed the entire system could move closer to SO. However, such a policy could be potentially detrimental to profitability, as the platform is now obligated to guarantee a minimum earning rate to anyone entering the market. Note that the price is the only "legal" tool available to the platform to manage the fleet size in the short term, and consequently, it has little recourse to reduce the fleet size below the lower bound now dictated by the government-mandated minimum wage. In the long run, however, the platform can

reduce its driver pool S_0 to manage this downward pressure on profits. In fact, after New York City enacted the minimum wage policy, both Uber and Lyft have stopped hiring drivers [Edelstein, 2019b]. Over the time, such a hiring freeze, along with other tactics, could reduce S_0. To examine this effect, we assume that the platform seeks to achieve a profit-maximizing S_0 for the minimum wage imposed by the regulator. More specifically, by solving M_2 over a range of possible values for S_0, we identify the S_0 that yields the highest profit and the scenario is then used to represent the long-term impact of the minimum wage policy.

Below, we compare the system performance under the four representative market conditions. For each condition, four scenarios are examined: profit maximization without minimum wage ("MO"), profit maximization with minimum wage and a fixed potential supply ("short"), profit maximization with minimum wage and an "optimized" potential supply ("long"), and SO ("SO"). Note that in Scenario "long", the value of S_0 differs from those used the other three scenarios due to the platform's presumed reaction. In all scenarios, the platform is assumed to adopt the mix-mode strategy.

Figure 5.3(a) plots the normalized welfare under each market condition. As expected, the minimum wage policy significantly improves social welfare in the short term, especially when the potential supply is relatively small. This improvement can be attributed to prioritizing driver and passenger surplus at the expense of the platform's profit, which is related to the reduced market power predicted in Eqs. (5.8a) and (5.8b). Interestingly, the welfare distribution among the three stakeholders in Scenario "short" closely resembles that at SO. However, the result looks rather different in Scenario "long", where the platform is allowed to manipulate S_0 to its own advantage. By doing so, it manages to take back some of the lost profits, but unfortunately inflicts greater damages on driver and passenger surplus. *Under all conditions, the minimum wage policy ends up lowering the welfare in the long term, and the effect is more damaging when the supply pool is small to begin with.*

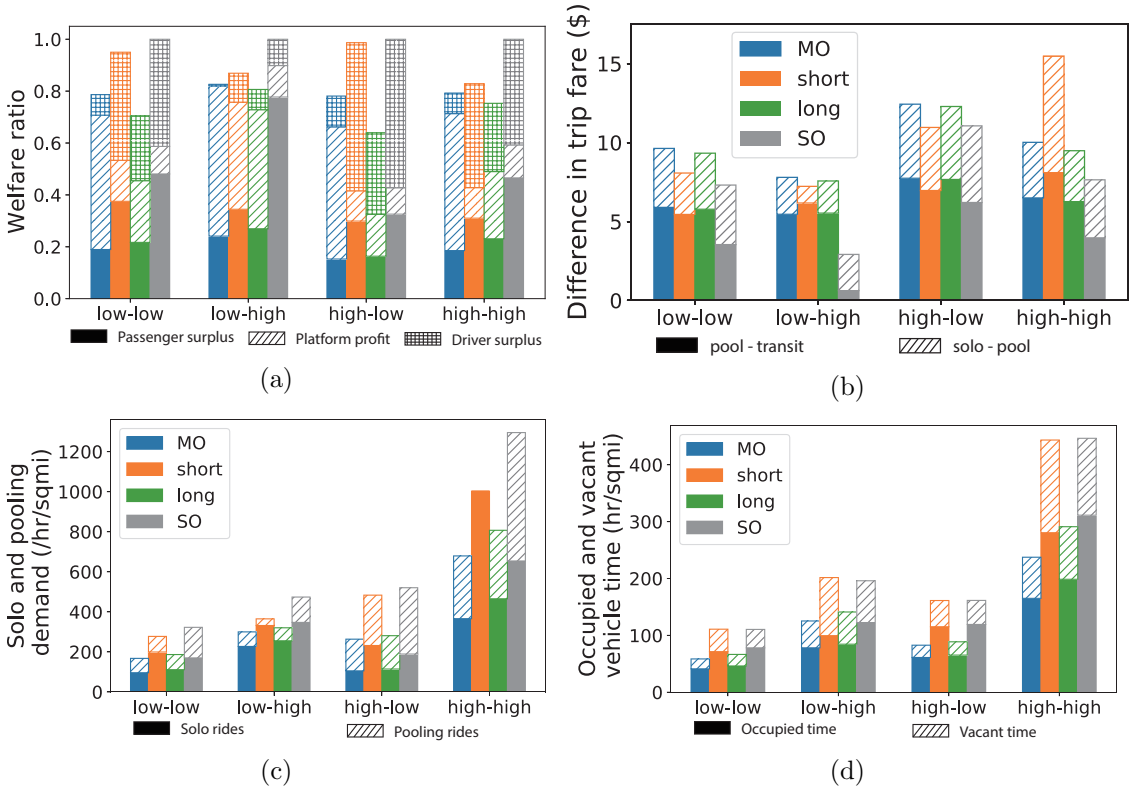


Figure 5.3. System performance under the minimum wage policy.

Figure 5.3(b) visualizes the platform's optimal pricing strategies. The bar chart in the plot represents the difference in trip fare between pooling and transit (solid filled) and between solo and pooling (pattern filled). In the short run, the platform tends to lower the trip fare of solo rides to attract passengers from transit. In this way, the platform could exhaust the additional service capacity induced by the minimum wage. An exception is the case of "high-high", where the platform raises both solo and pooling fares, but more for pooling than solo rides. As a result, no passenger would choose pooling at all; see Figure 5.3(c). Here, the rationale is to shift all pooling demand to solo rides, which not only keeps every driver busy but also generates higher revenue. These findings imply that *in a dense market, pooling may be completely eliminated by the minimum wage policy.* In the long-term, as the platform takes back more control on the supply side, it is able to keep the price close to the pre-regulation level for both pooling and solo rides.

Figures 5.3(c) and (d) show SO induces the highest demand and supply, followed by Scenario "short". This result again confirms, in the short run, imposing a socially optimal minimum wage will force the platform to scale up and discourage pooling by adjusting its pricing strategy. In addition, Scenario "short" significantly increases both occupied and vacant vehicle time. In fact, under all market conditions tested, the vacant vehicle time induced by the minimum wage policy in the short run is even higher than that at SO. In the long run, the ride-hail market is scaled back to the unregulated level. However, the pooling ratio does not fully recover. Instead, it remains modestly lower than what is achieved in both unregulated scenario and SO. Although the long-term adjustment made by the platform will largely eliminate the supply surge achieved by the minimum wage policy, the supply in Scenario "long" remains above the unregulated level.

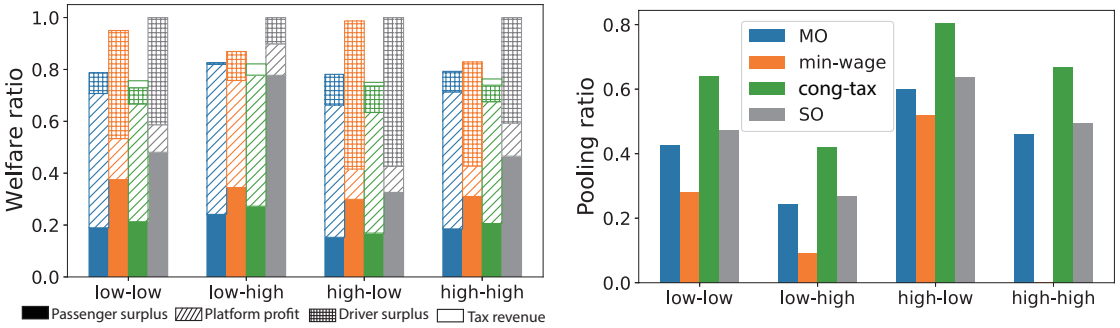
To summarize, although regulating the minimum wage does protect drivers from being unfairly exploited, it could create a host of problems. For one thing, the policy will surely draw opposition from the platform [e.g., Edelstein, 2019a]. More importantly, by maintaining the supply and demand of ride-hail at an artificially high level, it could depress the use of collective modes (transit and pooling), thus exacerbating traffic congestion. In the long run, the platform might limit the potential supply in order to recover the lost profits. As a result, *the regulation in the name of social justice might even hurt the social welfare.*

5.3.2.2. Congestion tax. We set the congestion tax $c_s = \$1$, which is in par with the actual policy implemented in Chicago[2]. Figure 5.4 compares the social welfare and pooling ratio in four scenarios: profit maximization without regulation ("MO"); profit maximization with minimum wage ("min-wage")[3]; profit maximization with congestion tax ("cong-tax"); and SO. The results show the congestion tax actually hurts the social welfare, though it slightly improves the passenger surplus; see Figure 5.4(a). Interestingly, the seemingly small congestion tax diverts a large number of passengers from solo to pooling rides. As shown in Figure 5.4(b), with the congestion tax, the

[2]The current policy charges congestion tax on all TNC trips, yet differently, based on the trip origin and destination as well as the time period. The price difference between solo and pooling rides is \$1.75 with downtown zone and \$0.6 otherwise. More details see `https://www.chicago.gov/city/en/depts/bacp/supp_info/city_of_chicago_congestion_pricing.html`

[3]This is equivalent to Scenario "short" in the previous section.

pooling ratio rises by more than 20% under all market conditions. In fact, a close look reveals that this effect is so dramatic that it significantly reduces the wait time for pooling rides, which is eventually translated to a greater passenger surplus.



(a) (b)

Figure 5.4. System performance under congestion tax.

Therefore, while the minimum wage policy improves the social welfare (at least in the short run) but undermines ride-sharing, the congestion tax has exactly the opposite effect. One wonders, naturally, whether or not jointly implementing these two policies would lead to a win-win solution. The results reported in Figure 5.5 offers a preliminary but promising answer to the question. The joint policy achieves a higher social welfare than each individual policy in all but one case. The exception is the case of "high-low", where the minimum wage policy delivers a slightly higher social welfare. Several factors contribute to the rise of social welfare. First, the higher earning rate attracts more drivers to enter the market. The improvement in LOS of ride-hail services thus retains the demand for solo rides, which translates into a significantly higher tax revenue compared to the congestion tax itself. Second, the increase in vehicle supply, along with the higher solo trip fare due to the congestion tax, makes pooling rides more appealing to passengers. Although the pooling ratio under the joint policy falls behind that only with the congestion tax policy, it is consistently higher than the pooling ratio at SO. It is reasonable to expect that the joint policies considered here is not "optimal" for a combined objective of minimizing social welfare, maximizing

toll revenue and promoting ride-sharing. We leave the problem of finding such an optimal policy to a future study.

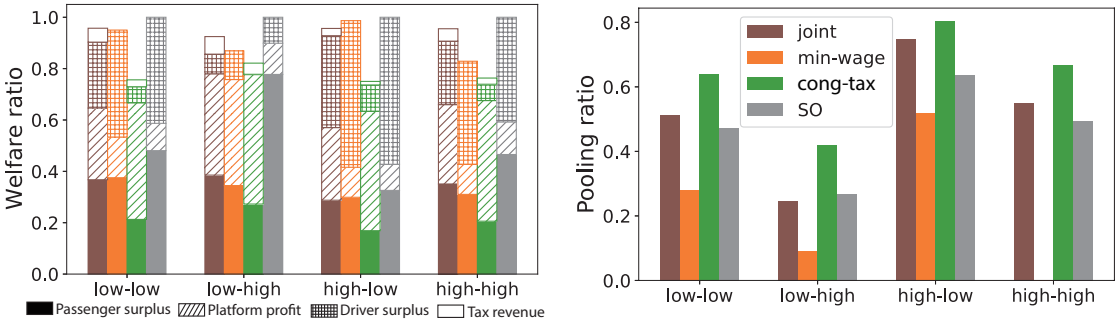


Figure 5.5. System performance under joint regulation of minimum wage and congestion tax.

5.4. Summary

In this chapter, we study the optimal pricing problem of monopoly platform serving both solo and pooling trips. The platform jointly determine the trip fares and the compensation rate paid to the drivers to maximize different objectives with and without regulatory constraints. We show the existence of optimal solution of the pricing problem and propose algorithms to solve for local optimal solutions. Numerical experiments are then conducted to compare different operational strategies and to examine the system perforamnces under various market conditions and regulations. Main findings are summarized as below:

- Without regulations, a mixed strategy, i.e., providing both solo and pooling rides, is the best choice for a profit-maximizing platform. Besides profit, it also achieves the highest social welfare compared to alternative strategies. Importantly, maintaining the system optimal output does not require subsidies if traffic congestion externality is considered in social welfare.
- The minimum wage policy can improve social welfare in the short term. However, in the long run, the platform might limit supply in an effort to recover the lost profits. As a

result, the policy could end up sacrificing social welfare, and the damage is greater when the potential supply is small. Moreover, by maintaining the supply and demand of ride-hail at an artificially high level, it could depress the use of collective modes (transit and pooling), and thus exacerbate traffic congestion.

- The congestion tax policy encourages pooling but undermines social welfare. Combining it with the minimum wage policy, however, achieves a desired balance between the two seemingly conflicting objectives in the short term.

In this chapter, we assume a single platform monopolizes the entire market, although it does not have full control on either side of the market. In reality, however, it is common to have multiple platforms competing with each other. In the next chapter, we will extend the current analysis to accommodate such platform competition. Another extension of the current model is establishing studying the pricing problem in a spatial ride-hail market, which will be main focus of Chapter 7.

A future study can also relax the assumptions made to simplify the matching process. Such an extension may endogenize the "matching parameters" (k and b) by linking them to such variables as matching interval/radius and maximum allowed detour. Accordingly, the platform may consider jointly optimizing the matching and the pricing decisions. As argued in Castillo et al. [2018], surge pricing can protect the system from the "catastrophic consequence" of WGC. The underlying logic is that certain amount of demand must be "priced out" so that the system can return to an efficient state of operation. Could pooling solve WGC without leaving a portion of demand unserved? This is also an intriguing question worthy of consideration in future studies.

5.5. Appendix

5.5.1. Notations

Table 5.1. List of notations

Variable	Description	Unit
Q_s (Q_p)	solo (pooling) demand rate	/hr/sqmi
r_Q	pooling ratio	
f_s (f_p)	trip fare of solo (pooling) rides	\$
η	compensation rate (payment per unit occupied time)	\$/hr
τ_s (τ_p)	travel time of solo (pooling) rides	hr
u_s	generalized cost of solo rides	\$
w_s	solo passenger wait time	hr
c_s	congestion tax on each solo ride	\$
ν	value of time	\$/hr
$\bar{u}$	average saving of each passenger due to switching from transit to ride-hail service	\$
θ_c	Mode choice uncertainty	/\$
S_0	potential supply	/sqmi
N	fleet size (number of drivers in operation)	/sqmi
e_0	average reservation rate	\$/hr
$\underline{e}$	minimum wage	\$/hr
c_0	congestion cost of each ride-hail vehicle	\$

5.5.2. Parameter setting

Table 5.2. Default values of additional parameters.

Parameter		Unit	Default value
Average reservation rate	e_0	\$/hr	19.84
Congestion cost per vehicle	c_0	\$/hr	2.9

5.5.3. Derivation of ∇R

Gradient ∇R is evaluated as

$$\frac{\partial R}{\partial f_s} = Q_s + (f_s - \eta\tau_s)\frac{\partial Q_s}{\partial f_s} + \left(f_p - \frac{1}{2}\eta\tau_p - \frac{1}{2}c_p\right)\frac{\partial Q_p}{\partial f_s}; \tag{5.17a}$$

$$\frac{\partial R}{\partial f_p} = Q_p + (f_s - \eta\tau_s)\frac{\partial Q_s}{\partial f_p} + \left(f_p - \frac{1}{2}\eta\tau_p - \frac{1}{2}c_p\right)\frac{\partial Q_p}{\partial f_p}; \tag{5.17b}$$

$$\frac{\partial R}{\partial \eta} = (f_s - \eta\tau_s)\frac{\partial Q_s}{\partial \eta} + \left(f_p - \frac{1}{2}\eta\tau_p - \frac{1}{2}c_p\right)\frac{\partial Q_p}{\partial \eta} - Q_s\tau_s - \frac{1}{2}Q_p\tau_p = 0. \tag{5.17c}$$

In what follows, we explain how to compute $\partial R/\partial f_s$ in each iteration. $\partial R/\partial f_p$ and $\partial R/\partial \eta$ can be computed similarly. The two components to be evaluated in Eq. (5.17a) are $\partial Q_s/\partial f_s$ and $\partial Q_p/\partial f_s$. Take $\partial Q_s/\partial f_s$ as an example. We expand it as

$$\frac{\partial Q_s}{\partial f_s} = D_0\left[\frac{\partial q}{\partial f_s} + \frac{\partial q}{\partial w_s}\frac{\partial w_s}{\partial f_s} + \frac{\partial q}{\partial w_{p1}}\frac{\partial w_{p1}}{\partial f_s} + \frac{\partial q}{\partial w_{p2}}\frac{\partial w_{p2}}{\partial f_s}\right]. \tag{5.18}$$

Here, the partial derivatives of the function q with respect to f_s, w_s, w_{p1} and w_{p2} can be evaluated numerically using automatic differentiation [Baydin et al., 2017].

To obtain the implicit partial derivatives $\partial w_s/\partial f_s$, $\partial w_{p1}/\partial f_s$ and $\partial w_{p2}/\partial f_s$, we first rewrite Eq. (4.5) as follows:

$$w_s = \frac{\delta}{2v\sqrt{k}}\sqrt{\frac{\Pi_{\text{eff}}}{V}}, \tag{5.19a}$$

$$w_{p1} = \frac{\delta}{2v\sqrt{k}}\sqrt{\frac{\Pi_{\text{eff}}}{V}\frac{\kappa + 4b\Pi_p}{2\kappa + 4b\Pi_p}}, \tag{5.19b}$$

$$w_{p2} = \frac{\delta}{2v\sqrt{b}}\frac{1}{\sqrt{\Pi_p}}, \tag{5.19c}$$

where Π_{eff} is defined in Eq. (3.8). Also, Π_{eff}, V and Π_p can be viewed as functions of f_s, w_s, w_{p1}, w_{p2} as per Eqs. (4.5a)-(4.5c).

Taking logarithm and then differentiating with respect of f_s on both sides of Eq. (5.19) yields

$$\frac{1}{w_s}\frac{\partial w_s}{\partial f_s} = \frac{1}{2\Pi_{\text{eff}}}\frac{\partial \Pi_{\text{eff}}}{\partial f_s} - \frac{1}{2V}\frac{\partial V}{\partial f_s}, \tag{5.20a}$$

$$\frac{1}{w_{p1}}\frac{\partial w_{p1}}{\partial f_s} = \frac{1}{2\Pi_{\text{eff}}}\frac{\partial \Pi_{\text{eff}}}{\partial f_s} - \frac{1}{2V}\frac{\partial V}{\partial f_s} + \frac{1}{2}\left(\frac{4b}{\kappa + 4b\Pi_p} - \frac{4b}{2\kappa + 4b\Pi_p}\right)\frac{\partial \Pi_p}{\partial f_s}, \tag{5.20b}$$

$$\frac{1}{w_{p2}}\frac{\partial w_{p2}}{\partial f_s} = -\frac{1}{2\Pi_p}\frac{\partial \Pi_p}{\partial f_s}. \tag{5.20c}$$

$\partial\Pi_{\text{eff}}/\partial f_s$ can be evaluated similarly as $\partial Q_s/\partial f_s$. Recall that $\Pi_{\text{eff}} = w_s Q_s + (w_{p1} + w_{p2})Q_p$. We may represent Π_{eff} as a function $\Pi_{\text{eff}} = \pi(w_s, w_{p1}, w_{p2}, Q_s, Q_p)$. Accordingly,

$$\frac{\partial \Pi'}{\partial f_s} = \frac{\partial \pi}{\partial w_s}\frac{\partial w_s}{\partial f_s} + \frac{\partial \pi}{\partial w_{p1}}\frac{\partial w_{p1}}{\partial f_s} + \frac{\partial \pi}{\partial w_{p2}}\frac{\partial w_{p2}}{\partial f_s} + \frac{\partial \pi}{\partial Q_s}\frac{\partial Q_s}{\partial f_s} + \frac{\partial \pi}{\partial Q_p}\frac{\partial Q_p}{\partial f_s}. \tag{5.21}$$

Again, the partial derivatives of π can be computed by automatic differentiation.

In other words, $\partial\Pi_{\text{eff}}/\partial f_s$ can be expressed as a linear function of $\partial w_s/\partial f_s$, $\partial w_{p1}/\partial f_s$ and $\partial w_{p2}/\partial f_s$. $\partial V/\partial f_s$ and $\partial \Pi_p/\partial f_s$ in Eq. (5.20) can be derived in the same way. Consequently, Eq. (5.20) turns into a linear equation system with respect to $\partial w_s/\partial f_s$, $\partial w_{p1}/\partial f_s$ and $\partial w_{p2}/\partial f_s$.

Plugging the solution of Eq. (5.20) into Eq. (5.18), we can obtain $\partial Q_s/\partial f_s$. The computation of $\partial Q_p/\partial f_s$ is similar and omitted here for brevity.

CHAPTER 6

Duopoly pricing in aggregate market

One of the most restrictive assumptions in Chapter 5 is that the market is dominated by a single platform. In reality, inter-platform competition is commonplace, with the duopoly of Uber and Lyft being a well-known example. Hence, in this chapter, we consider two platforms competing for both passengers and drivers in an aggregate ride-hail market. The competition is formulated as a simultaneous pricing game. That is, the platforms set the price for passengers and the compensation rate for drivers in order to maximize their own profits. The interactions between passengers, drivers and the platforms are characterized by the equilibrium model developed in Chapter 4 for the multi-platform market under either single-homing or multi-homing. Similar to Chapter 5, the platform's strategies, as well as the system performance, are investigated in both unregulated and regulated scenarios. All notations used in this chapter follow those in Chapter 4, which are also reported in Table 6.1.

6.1. Duopoly pricing game

6.1.1. Unregulated duopoly equilibrium

Consider an aggregate market with two platforms $\mathcal{P} = \{A, B\}$. Let $\mathbf{y}^j = \left[f_s^j, f_p^j, \eta^j\right]^T \in \mathbb{R}_+^3$ denote the pricing strategy of platform $j \in \mathcal{P}$. Without regulations, the feasible set of $\mathbf{y}^j$ is not affected by the other platform's strategy. In this case, the equilibrium state can be characterized as a Nash Equilibrium (NE), i.e., no platform can increase its profit by unilaterally changing its own pricing strategy.

Same as the monopoly scenario, we define each platform's profit as its total revenue less the direct expense of trip production. Hence, the profit of platform j, denoted by $R^j(\mathbf{y}^j, \mathbf{y}^{-j})$, is

evaluated as

$$R^j(\mathbf{y}^j, \mathbf{y}^{-j}) = f_s^j Q_s^j + f_p^j Q_p^j - \eta^j \left(Q_s^j \tau_s^j + \frac{1}{2} Q_p^j \tau_p^j \right). \tag{6.1}$$

Mathematically, the duopoly equilibrium strategy $\mathbf{y}^*$ satisfies the following conditions:

$$R^j(\mathbf{y}^{j*}, \mathbf{y}^{-j*}) \geq R^j(\mathbf{y}^j, \mathbf{y}^{-j*}),\ \forall \mathbf{y}^j \in \mathbb{R}^3_+,\ j \in \{A, B\}. \tag{6.2}$$

It is well known that the Nash equilibrium condition can be transformed into a variational inequality problem (VIP) [e.g., Harker, 1991]. Consider the optimal pricing problem of platform j

$$\max_{\mathbf{y}^j} R^j(\mathbf{y}^j, \mathbf{y}^{-j}). \tag{6.3}$$

The first-order necessary condition is given by

$$-\nabla_{\mathbf{y}^j} R^j(\mathbf{y}^{j*}, \mathbf{y}^{-j})^T (\mathbf{y}^j - \mathbf{y}^{j*}) \geq 0, \quad \forall \mathbf{y}^j \in \mathbb{R}^3_+, \tag{6.4}$$

where $\nabla_{\mathbf{y}^j} R^j(\mathbf{y}^{j*}, \mathbf{y}^{-j})$ is the gradient of R^j with respect to platform j's pricing vector and $\mathbf{y}^{j*}$ is the solution to (6.3) given the other platform's pricing strategy $\mathbf{y}^{-j}$. Note that Eq. (6.4) is sufficient only if $R^j(\cdot, \mathbf{y}^{-j})$ is concave. Otherwise, the solution to Eq. (6.4) is either a stationary point (i.e., $\nabla_{\mathbf{y}^j} R^j(\mathbf{y}^{j*}, \mathbf{y}^{-j}) = 0$) or a local maximum (i.e., $R^j(\mathbf{y}^{j^*}, \mathbf{y}^{-j}) \geq R^j(\mathbf{y}^j, \mathbf{y}^{-j})$, $\forall \mathbf{y}^j$ in a neighborhood of $\mathbf{y}^{j^*}$).

Since there is no binding constraints across players (i.e., the feasible set of the game is the full Cartesian product of individual player's strategy set), we can construct the VIP by summing up the above first-order condition of all players. Let $\mathbf{y} = [f_s^A, f_p^A, \eta^A, f_s^B, f_p^B, \eta^B]^T$ be the joint pricing strategy of the game. The VIP is stated as

$$\text{(D1)} \quad \text{Find } \mathbf{y}^* \in \mathbb{R}^6_+ \text{ such that} \quad -\nabla R(\mathbf{y}^*)^T (\mathbf{y} - \mathbf{y}^*) \geq 0,\ \forall \mathbf{y} \in \mathbb{R}^6_+, \tag{6.5}$$

where $\nabla R(\mathbf{y}) = [\nabla_{\mathbf{y}^A} R^A(\mathbf{y})^T, \nabla_{\mathbf{y}^B} R^B(\mathbf{y})^T]^T$ and $\nabla_{\mathbf{y}^j} R^j(\mathbf{y}) = \nabla_{\mathbf{y}^j} R^j(\mathbf{y}^j, \mathbf{y}^{-j})$.

6.1.2. Regulated duopoly equilibrium

In the unregulated game, the platforms only interact through their profit function, and the feasible sets of their pricing strategies are independent. This property implies that the feasible set of the pricing game is the full Cartesian product of each individual player's feasible set (i.e., $\mathbb{R}^3_+ \times \mathbb{R}^3_+ = \mathbb{R}^6_+$). However, with regulatory constraints, the feasible set of one platform is affected by the other. In this case, the feasible set of the regulated pricing game can be viewed as a point-to-set mapping $\Omega(\mathbf{y}) = \prod_{j\in\mathcal{P}} \Omega^j(\mathbf{y}^{-j})$, where $\Omega^j(\mathbf{y}^{-j}) = \{\mathbf{y}^j | h^j(\mathbf{y}^j, \mathbf{y}^{-j}) \leq 0, \mathbf{y}^j \in \mathbb{R}^3_+\}$. For a minimum wage policy, the constraint $h^j(\mathbf{y}^j, \mathbf{y}^{-j}) \leq 0$ can be further specified as

$$e^j = \frac{\eta^j}{N^j}\left(Q^j_s \tau^j_s + \frac{1}{2} Q^j_p \tau^j_p\right) \geq \underline{e} \tag{6.6}$$

for single-homing supply mode and

$$e = e^j = \frac{1}{N}\sum_j \eta^j \left(Q^j_s \tau^j_s + \frac{1}{2} Q^j_p \tau^j_p\right) \geq \underline{e} \tag{6.7}$$

for multi-homing, where $\underline{e}$ is the minimum wage rate imposed by the regulator. Note that, in addition to $\mathbf{y}^j$, $\mathbf{y}^{-j}$ is also involved in Eqs. (6.6) and (6.7) through passenger demands Q^j_s, Q^j_p and fleet size N^j.

This interaction through interdependent feasible sets gives rise to the concept of generalized Nash equilibrium (GNE) [Arrow and Debreu, 1954, Ichiishi, 1983]. In the context of regulated duopoly pricing game, it is characterized as

$$R^j(\mathbf{y}^{j*}, \mathbf{y}^{-j*}) \geq R^j(\mathbf{y}^j, \mathbf{y}^{-j*}),\ \forall \mathbf{y}^j \in \Omega(\mathbf{y}^*),\ j \in \{A, B\}. \tag{6.8}$$

Similar to the case of NE, GNE corresponds to a quasi-variational inequality problem (QVIP) [see e.g., Harker, 1991, Facchinei and Kanzow, 2010], which reads

$$\text{(D2) Find } \mathbf{y}^* \in \Omega(\mathbf{y}^*) \text{ such that} \quad -\nabla R(\mathbf{y}^*)^T(\mathbf{y} - \mathbf{y}^*) \geq 0,\ \forall \mathbf{y} \in \Omega(\mathbf{y}^*). \tag{6.9}$$

Again, because we cannot ensure that R is concave, a solution to the QVIP is not necessarily a GNE. However, the reverse statement must be true.

6.1.3. Single-homing vs multi-homing

As discussed in Section 4.4, switching from single-homing to multi-homing does not affect the passenger wait time if the platforms are symmetric and the inputs fed to the matching process (i.e., densities of vacant vehicles and waiting passengers) remain the same. However, the marginal effect of a platform's pricing strategy on these equilibrium variables does vary between the two supply modes. To elaborate the difference, consider two symmetric platforms $\mathcal{P} = \{A, B\}$ that only serve solo rides. The gradient $\nabla_{\mathbf{y}^A} R^A$ can be derived as

$$\frac{\partial R^A}{\partial f_s^A} = Q_s^A + (f_s^A - \eta^A \tau_s^A)\frac{\partial Q_s^A}{\partial f_s^A}; \tag{6.10a}$$

$$\frac{\partial R^A}{\partial \eta^A} = -Q_s^A \tau_s^A + (f_s^A - \eta^A \tau_s^A)\frac{\partial Q_s^A}{\partial \eta^A}. \tag{6.10b}$$

The terms $\partial Q_s^A / \partial f_s^A$ and $\partial Q_s^A / \partial \eta^A$ can be evaluated by

$$\frac{\partial Q_s^A}{\partial f_s^A} = \nabla_{u_s^A} q + \nu \left(\nabla_{u_s^A} q \frac{\partial w_s^A}{\partial f_s^A} + \nabla_{u_s^B} q \frac{\partial w_s^B}{\partial f_s^A} \right), \tag{6.11}$$

$$\frac{\partial Q_s^A}{\partial \eta^A} = \nu \left(\nabla_{u_s^A} q \frac{\partial w_s^A}{\partial \eta^A} + \nabla_{u_s^B} q \frac{\partial w_s^B}{\partial \eta^A} \right), \tag{6.12}$$

where $\nabla_{u_s^A} q$ and $\nabla_{u_s^B} q$ are the partial derivatives of the demand function $q(\cdot)$ with respect to u_s^A, u_s^B, respectively.

Without loss of generality, we may assume $\nabla_{u_s^A} q < 0$, $\nabla_{u_s^B} q > 0$ and $|\nabla_{u_s^A} q| > |\nabla_{u_s^B} q|$ (i.e., the general cost associated with one platform has a larger impact on itself than the other platform). Hence, the sign of $\partial Q_s^A / \partial f_s^A$, which indicates whether an increase in trip fare would increase or decrease the platform's market share, depends on $\partial w_s^A / f_s^A$ and $\partial w_s^B / f_s^A$. The same is true for the compensation rate η^A. As per Eq. (3.9), the solo wait time increases with passenger demand but

decreases with vehicle supply. Since a higher trip fare usually drags down the demand while a higher compensation rate always attracts more drivers to enter the system, we may assume $\partial w_s^A/\partial f_s^A < 0$ and $\partial w_s^A/\partial \eta^A < 0$, regardless of the supply mode.

Under single-homing, an increase in f_s^A would raise the wait time on platform B, i.e., $\partial w_s^B/\partial f_s^A > 0$, because it would increase the demand on platform B, thus intensifying competition there. Similarly, we expect $\partial w_s^B/\partial \eta^A > 0$ because an increase in η_A attracts more drivers to platform A, at the expense of platform B. On the other hand, as shown in Section 4.4.2, $w_s^A \propto w_s^B$ under multi-homing and the constant ratio is positive. Therefore, we have $\partial w_s^B/\partial f_s^A \propto \partial w_s^A/\partial f_s^A \to \partial w_s^B/\partial f_s^A < 0$. Since the passenger wait time depends on both platforms under multi-homing, the influence of each platform is expected to be less compared to single-homing. Hence,

$$\left(\frac{\partial w_s^B}{\partial f_s^A}\right)_{SH} > 0 > \left(\frac{\partial w_s^B}{\partial f_s^A}\right)_{MH} \propto \left(\frac{\partial w_s^A}{\partial f_s^A}\right)_{MH} > \left(\frac{\partial w_s^A}{\partial f_s^A}\right)_{SH},$$

where subscripts "SH" and "MH" denote single-homing and multi-homing, respectively.

Accordingly, it yields

$$\begin{aligned}(\partial Q_s^A/\partial f_s^A)_{MH} &= \nabla_{u_s^A} q + \nu \left[\nabla_{u_s^A} q \left(\partial w_s^A/\partial f_s^A\right)_{MH} + \nabla_{u_s^B} q \left(\partial w_s^B/\partial f_s^A\right)_{MH}\right] \\ &< \nabla_{u_s^A} q + \nu \left[\nabla_{u_s^A} q \left(\partial w_s^A/\partial f_s^A\right)_{SH} + \nabla_{u_s^B} q \left(\partial w_s^B/\partial f_s^A\right)_{SH}\right] = (\partial Q_s^A/\partial f_s^A)_{SH}.\end{aligned} \tag{6.13}$$

The inequality holds because $\nabla_{u_s^A} q < 0, \nabla_{u_s^B} q > 0$, as discussed above. The same reasoning can be used to derive $(\partial Q_s^A/\partial \eta^A)_{MH} < (\partial Q_s^A/\partial \eta^A)_{SH}$. Bringing this result to Eq. (6.10), we can conclude that, under multi-homing, the platforms are less inclined to raise trip fare and compensation rate because their profit tends to rise less when they do so than under single-homing.

To further explore the impact of a shift in compensation rate on the vehicle supply, we derive $\partial N/\partial \eta$ under single-homing and multi-homing modes. First, for platform A, it reads

$$\frac{\partial N^A}{\partial \eta^A} = S_0 \nabla_{e^A} g \frac{\partial e^A}{\partial \eta^A}, \tag{6.14}$$

where g is the supply function and $\nabla_e^A g$ is the derivative of g with respect to e^A. It is reasonable to assume $\nabla_{e^A} g > 0$, since the fleet size tends to increase with average wage rate. Hence, in general $\partial N^j/\partial \eta^j$ is positively proportional to $\partial e^j/\partial \eta^j$, for $j = A, B$.

For single-homing, first taking logarithm on both sides of Eq. (4.6) and then getting derivative with respect to η yield

$$\frac{1}{e^A}\frac{\partial e^A}{\partial \eta^A} = \frac{Q_s^A \tau_s^A + \eta^A \tau_s^A \frac{\partial Q_s^A}{\partial \eta^A}}{\eta^A Q_s^A \tau_s^A} - \frac{1}{N^A}\frac{\partial N^A}{\partial \eta^A}. \tag{6.15}$$

Plugging Eq.(6.14) into Eq. (6.15) gives

$$\left(\frac{\partial e^A}{\partial \eta^A}\right)_{SH} = \left(\frac{1}{e^A} + \frac{S_0^A}{N^A}\nabla_{e^A} g\right) \frac{Q_s^A \tau_s^A + \eta^A \tau_s^A \frac{\partial Q_s^A}{\partial \eta^A}}{\eta^A Q_s^A \tau_s^A}. \tag{6.16}$$

Following a similar derivation, $\partial e^A/\partial \eta^A$ under multi-homing can be derived form Eq. (4.9) as

$$\left(\frac{\partial e^A}{\partial \eta^A}\right)_{MH} = \left(\frac{1}{e^A} + \frac{S_0^A}{N^A}\nabla_{e^A} g\right) \frac{Q_s^A \tau_s^A + \eta^A \tau_s^A \frac{\partial Q_s^A}{\partial \eta^A} + \eta^B \tau_s^B \frac{\partial Q_s^B}{\partial \eta^A}}{\eta^A Q_s^A \tau_s^A + \eta^B Q_s^B \tau_s^B}. \tag{6.17}$$

When platforms are symmetric, we have $\eta^A = \eta^B, Q_s^A = Q_s^B, \tau_s^A = \tau_s^B$. In addition, it can be derived from Eq. (6.12) that $\partial Q_s^A/\partial \eta^A = \partial Q_s^B/\partial \eta^A > 0$[1]. Hence, we have

$$\left(\frac{\partial e^A}{\partial \eta^A}\right)_{MH} = \left(\frac{1}{e^A} + \frac{S_0^A}{N^A}\nabla_{e^A} g\right) \frac{\frac{1}{2}Q_s^A \tau_s^A + \eta^A \tau_s^A \frac{\partial Q_s^A}{\partial \eta^A}}{\eta^A Q_s^A \tau_s^A} < \left(\frac{\partial e^A}{\partial \eta^A}\right)_{SH}. \tag{6.18}$$

Thus, an increase in compensation rate would induce less supply under multi-homing than under simple-homing. In other words, when a platform attempts to squeeze profit by reducing the compensation to drivers, it only bears part of the cost caused by supply depression. In economics, such a phenomenon is widely known as the tragedy of the commons [Ostrom et al., 1994]. Here, it is the drivers that are the "commons" shared (and abused) by profit-driven platforms under the multi-homing mode.

[1]This is only true under symmetric multi-homing where $w_s^A = w_s^B$

6.2. Solution existence and algorithm

Due to the complex structure of the market equilibrium constraint, we cannot obtain a closed-form expression of $R^j(\cdot, \mathbf{y}^{-j})$ or establish its concavity. Hence, in theory a solution to D1 does not necessarily satisfy the equilibrium condition Eq. (6.2), though the reverse is evidently true. In what follows, we first establish the solution existence for D1, and then propose to solve it using a gradient ascent algorithm. Because a solution to D1 found by such an algorithm must be a local maximum, it must satisfy Eq. (6.2).

The existence of the unregulated duopoly equilibrium is hard to prove due to the nonconvexity of $R(\mathbf{y})$. Instead, we prove the solution existence of D1, as summarized in the following proposition. Accordingly, we may first solve D1 and check whether the solution satisfy the condition Eq.(6.2). In the experiments, we solve the problem with different initial solutions in case there are multiple duopoly equilibria.

Proposition 3. *The VI problem D1 has a solution.*

Proof. The proposition is proved by evoking Corollary 3.1 of Harker and Pang [1990].

Lemma 2. (Harker and Pang [1990] Corollary 3.1) *Let X be a nonempty, closed and convex subset of $\mathbb{R}^n$ and F be a continuous mapping $\mathbb{R}^n$ to itself. Then, the VI problem*

$$F(x^*)^T(x - x^*) \geq 0, \ \forall x \in X$$

has a solution if there exists a nonempty bounded set $D \subset X$ such that for every $x \in X \setminus D$ there is a $x_0 \in D$ with $F(x)^T(x_0 - x) \leq 0$, that is, no point outside D is a solution candidate for the VI problem.

Rewrite the objective function of each player:

$$R^j = (f_s^j - \eta^j \tau_s^j) Q_s^j + (f_p^j - \frac{1}{2}\eta \tau_p^j) Q_p^j. \tag{6.19}$$

Hence, the optimal price vector must satisfy $\eta^{j*} \leq \max\{f_s^{j*}/\tau_s^j, 2f_p^{j*}/\tau_p^j\}$. Under this condition, it is easy to show that R^j is bounded from below and has limit 0 one both sides (i.e., $\lim_{\mathbf{y}^j \to 0} R^j = 0$ and $\lim_{\mathbf{y}^j \to \infty} R^j = 0$). Since R is continuously differentiable, there must exist a point $\mathbf{y}_0 = \left[\mathbf{y}_0^A, \mathbf{y}_0^B\right]$ such that $\nabla_{\mathbf{y}_0^A} R^j(\mathbf{y}_0) \leq 0$ and $\nabla_{\mathbf{y}_0^B} R^j(\mathbf{y}_0) \leq 0$. Therefore, for all $\mathbf{y} \geq \mathbf{y}_0$, it yields $-\nabla R^j(\mathbf{y})^T(\mathbf{y}_0 - \mathbf{y}) \leq 0$. As per Lemma 2, there exists a solution to D1. □

Since the platforms always yield zero profit at the boundaries, the equilibrium price must lie inside of the feasible set (i.e., $\mathbf{y}^* > \mathbf{0}$). Hence, it is sufficient to solve $\nabla R(\mathbf{y}) = 0$ for D1. The solution procedure follows a similar iterative rule as the optimal pricing problem under monopoly:

$$\mathbf{y}^{(i+1)} = \mathbf{y}^{(i)} + \alpha \nabla R(\mathbf{y}^{(i)}). \tag{6.20}$$

To solve the regulated duopoly equilibrium Eq. (6.8), we implement the exact penalty method proposed by Facchinei and Kanzow [2010], summarized in Algorithm 6.1. The existence of the equilibrium is due to Theorem 2.5 in Facchinei and Kanzow [2010] associated with its solution algorithm. In brief, the theorem states that if the algorithm terminates in a finite number of iterations, then the limit point is a solution of the GNE.

In Algorithm 6.1, $h_+(\cdot) = \max\{0, h(\cdot)\}$ and the parameter ε is introduced to smooth the objective function of penalized problem. The algorithm terminates when the constraint is satisfied and ε is smaller than a predefined threshold $\bar{\varepsilon}$.

6.3. Numerical experiments

Same as Section 5.3, the numerical experiments in this section are constructed based on the TNC data collected in the City of Chicago (see Chapter 10). Specifically, we focus on four representative demand-supply conditions, dubbed as "low-low" (low-demand-low-supply), "low-high" (low-demand-high-supply), "high-low" (high-demand-low supply) and "high-high" (high-demand-high-supply). The corresponding total demand and potential supply are reported in Table 6.2. The

Algorithm 6.1 Solution algorithm for regulated duopoly equilibrium

Inputs: QVI problem Eq. (6.9), $\mathbf{y}^{(0)}$, $c \in (0,1)$, $\rho > 1$, $\varepsilon \in (0,1]$, $\bar{\varepsilon} \in [0,1)$.
Set $i = 0$, $\lambda = \mathbf{0}$.
while $\mathbf{y}^{(i)} \notin \Omega(\mathbf{y}^{(i)})$ or $\varepsilon \geq \bar{\varepsilon}$ **do**
Let $I = \{j | h^j(\mathbf{y}^{(i)}) \geq 0\}$, For each $j \in I$, if

$$||\nabla_{\mathbf{y}^j} R^j(\mathbf{y}^{(i)})|| > c\lambda^j ||\nabla_{\mathbf{y}^j} ||h_+^j(\mathbf{y}^{(i)})||_2||, \tag{6.21}$$

then set $\lambda^j = \lambda^j + \rho h^j(\mathbf{y}^{(i)})$.
If $||h(\mathbf{y}^{(i)})|| \leq \varepsilon$, set $\varepsilon = \varepsilon/2$
Solve penalized VI problem

$$F(\mathbf{y}^*)^T(\mathbf{y} - \mathbf{y}^*) \geq 0, \ \forall \mathbf{y} \geq 0, \tag{6.22}$$

where $F(\mathbf{y}^*) = \begin{bmatrix} F^A(\mathbf{y}^*)^T \\ F^B(\mathbf{y}^*)^T \end{bmatrix}$ and

$$F^j(\mathbf{y}) = -\nabla_{\mathbf{y}^j} R^j(\mathbf{y}) + \lambda^j \nabla_{\mathbf{y}^j} \left[\left(h_+^j(\mathbf{y}^{(i)}) \right)^2 + \varepsilon \right]^{1/2}. \tag{6.23}$$

Set $i = i + 1$.
end while
if $\mathbf{y}^{(l)}$ satisfies Eq. (6.8) **then**
Set $\mathbf{y}^* = \mathbf{y}^{(l)}$ as a GNE.
else
Report no GNE solution is found.
end if

default values of input parameters are the same as those used in Section 4.6, which are reported in Table 4.1.

For comparison, we also solve the monopoly pricing problem using the same inputs. The model and solution method has been thoroughly discussed in Chapter 5. Additionally, we define the social welfare of the aggregate market in the three scenarios as follows:

(1) Monopoly

$$\begin{aligned} W_{\text{MO}} = \frac{D_0}{\theta_c} \ln \left\{ 1 + \left\{ \sum_{m \in \mathcal{M}} \exp[\theta_c^r (u_t - u_m)] \right\}^{\theta_c/\theta_c^r} \right\} \\ + \frac{S_0}{\theta_d} \ln \{1 + \exp[\theta_d (e - e_0)]\} \\ + \sum_{j \in \mathcal{P}} \left[f_s^j Q_s^j + f_p^j Q_p^j - \eta^j \left(Q_s^j \tau_s^j + \frac{1}{2} Q_p^j \tau_p^j \right) \right] - c_0 N; \end{aligned} \tag{6.24}$$

(2) Duopoly under single-homing

$$W_{\text{single}} = \frac{D_0}{\theta_c} \ln \left\{ 1 + \left\{ \sum_{m \in \mathcal{M}} \sum_{j \in \mathcal{P}} \exp[\theta_c^r (u_t - u_m^j)] \right\}^{\theta_c/\theta_c^r} \right\} + \frac{S_0}{\theta_d} \ln \left\{ 1 + \left\{ \sum_{j \in \mathcal{P}} \exp[\theta_d (e^j - e_0)] \right\}^{\theta_d/\theta_d^r} \right\} + \sum_{j \in \mathcal{P}} \left[f_s^j Q_s^j + f_p^j Q_p^j - \eta^j \left(Q_s^j \tau_s^j + \frac{1}{2} Q_p^j \tau_p^j \right) \right] - c_0 N; \tag{6.25}$$

(3) Duopoly under multi-homing

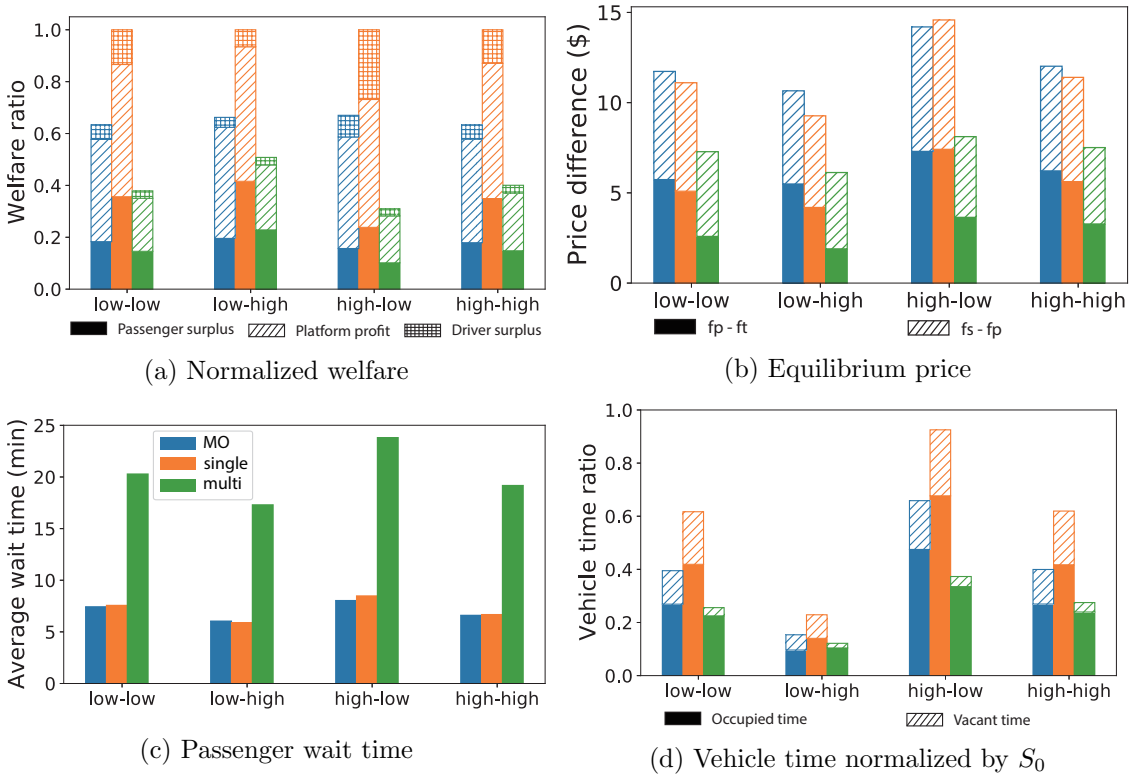
$$W_{\text{multi}} = \frac{D_0}{\theta_c} \ln \left\{ 1 + \left\{ \sum_{m \in \mathcal{M}} \sum_{j \in \mathcal{P}} \exp[\theta_c^r (u_t - u_m^j)] \right\}^{\theta_c/\theta_c^r} \right\} + \frac{S_0}{\theta_d} \ln \{1 + \exp[\theta_d (e - e_0)]\} + \sum_{j \in \mathcal{P}} \left[f_s^j Q_s^j + f_p^j Q_p^j - \eta^j \left(Q_s^j \tau_s^j + \frac{1}{2} Q_p^j \tau_p^j \right) \right] - c_0 N. \tag{6.26}$$

In all three welfare functions, the first term represents passenger surplus based on the logsum term derived from the NL or MNL model [see e.g., Kohli and Daly, 2006, De Jong et al., 2007]. Here, the utility of the transit trip is used as a benchmark to measure the extra utility contributed by the e-hail services. The second term similarly measures driver surplus, using the fallback option as a benchmark. The third term quantifies the total profits of all platforms and the last term approximates the congestion cost contributed by all vehicles employed by the e-hail market (c_0 is the marginal congestion cost per vehicle per unit operating time and set to be \$2.9 per vehicle in align with previous analysis).

6.3.1. Performance of unregulated duopoly game

Figure 6.1 compares the results of the duopoly pricing game under the two supply modes with those in monopoly. For clarity, the social welfare is normalized against the highest value under

each market condition. The normalized total welfare is compared in Figure 6.1(a), along with its three main components: consumer surplus, driver surplus and platform profit[2]. In all cases, the platform(s) is the winner, taking the lion share of the gains in the social welfare. The single- and multi-homing duopoly respectively achieve the highest and lowest social welfare under all conditions, with the monopoly lies right in the middle. The system performance under multi-homing is surprisingly bad, consistently lagging behind single-homing in all three components of the social welfare.



(a) Normalized welfare

(b) Equilibrium price

(c) Passenger wait time

(d) Vehicle time normalized by S_0

Figure 6.1. System performance without regulatory constraints: duopoly vs. monopoly. "MO" stands for monopoly; "single" stands for single-homing duopoly game; "multi" stands for multi-homing duopoly game.

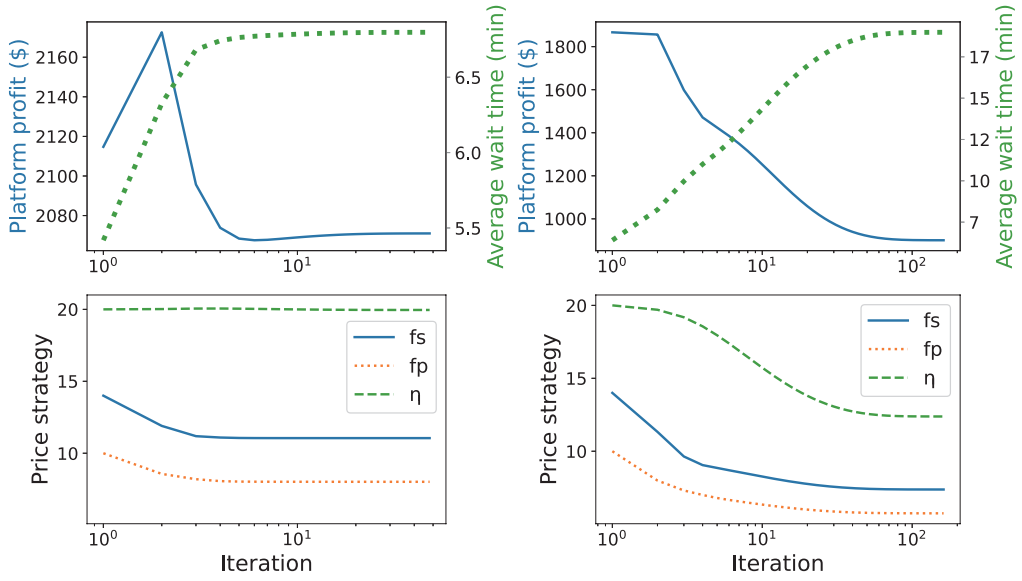
Figure 6.1(c) indicates that the lower passenger surplus in the multi-homing duopoly is likely due to the much longer wait time (which triples that in either the monopoly or the single-homing duopoly). It is further confirmed in Figure 6.1(d), which plots the ratio of vehicle time dedicated to

[2]The congestion externality cost is deducted from driver surplus.

ride-hail services. Even under the most favorable condition (high demand low supply), barely 40% of all potential workforce are working as ride-hail drivers in the multi-homing duopoly, compared to over 90% in the single-homing duopoly and about 70% in the monopoly. The vacant vehicle time is even lower, directly leading to the excessively long wait time in multi-homing, which in turn would depress the demand. Figure 6.1(b) shows why the platforms also suffer in the multi-homing duopoly: they have to charge a lower price to make up for their overall poorer LOS. Another noteworthy observation is the multi-homing duopoly tends to maintain a greater price spread between solo and pooling rides than the spread between pooling and transit, a choice that tends to favor pooling rides. For the monopoly, this difference between two spreads is mostly negligible.

Zha et al. [2016] and Nikzad [2017] pointed out that, despite the competition it introduces, a duopoly could set a higher price than a monopoly under certain demand-supply scenarios. This observation is confirmed in our results. As shown in Figure 6.1(b), when demand is high and supply is low (high-low), both solo and pooling trip prices are slightly higher in the single-homing duopoly than in the monopoly. This aligns with the finding of Nikzad [2017] especially well, who concludes that the duopoly price could be higher than the monopoly price if the market is not "sufficiently thick" (i.e., when the potential supply is low). Importantly, in most cases, the single-homing duopoly lowers the price for pooling, suggesting *the pressure of competition encourage the platforms to promote pooling.*

Why does the multi-homing duopoly perform so poorly? The discussion in Section 6.1.3 suggests that the culprit is the tragedy of the commons. To further examine the underlying mechanisms, we solve the duopoly game under the two supply modes with the same initial solution and plot the results over iterations in Figure 6.2. Specifically, the total demand, potential supply and initial price are set to be the default values in Table 4.2. The evolution can be seen as a repeated game where the two platforms slightly adjust their strategies along the best direction at each stage. Due to symmetry, the two platforms in each game follow exactly the same evolution path.



(a) Single-homing (b) Multi-homing

Figure 6.2. Solving unregulated duopoly equilibrium.

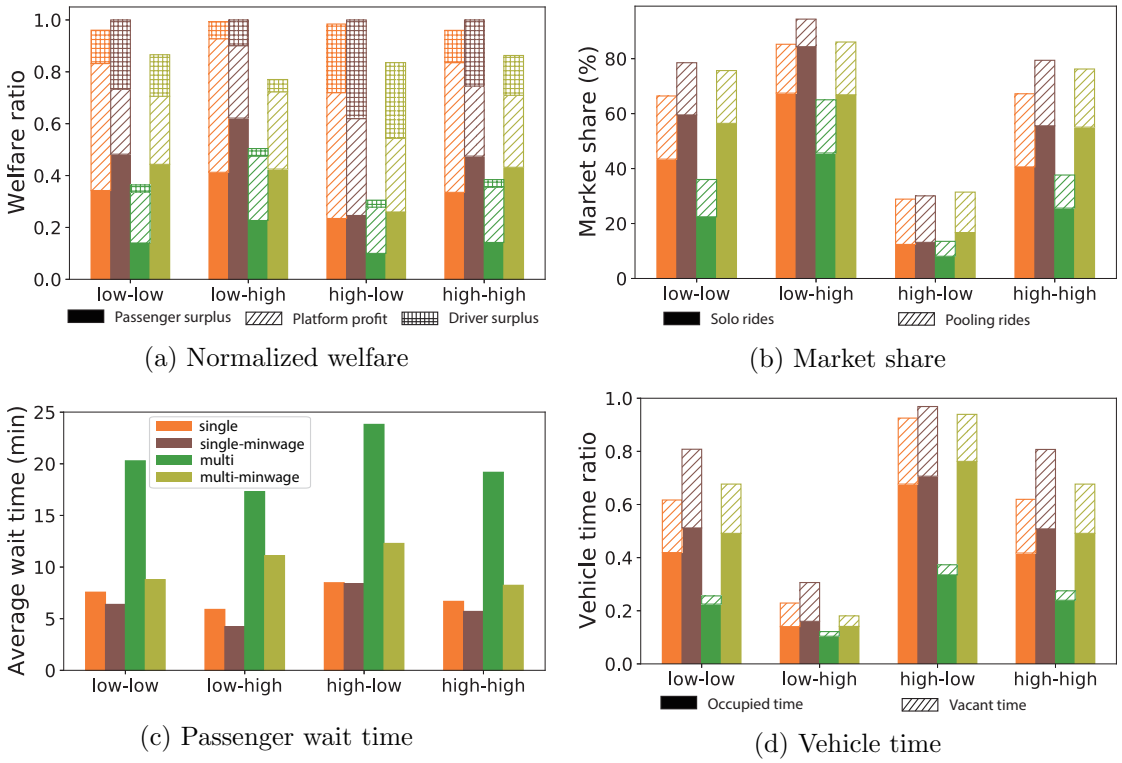
Comparing the bottom panels of Figure 6.2(a) and (b), one can find the platforms in the multi-homing duopoly tend to attract passengers with lower trip fares, rather than a higher LOS (i.e., shorter wait times). Since a multi-homing driver's decision to enter service hinges on the "market wage rate", it is more difficult for a platform to attract drivers by unilaterally raising the compensation rate η. Instead, the platform is better off by lowering η to reduce its operating cost, in exchange for a supply loss that is shared by its rival. This behavior is consistent with the prediction in Section 6.1.3. Consequently, *the pricing game in the multi-homing duopoly is stuck in a trap with lower fare, lower compensation and lower LOS.*

Interestingly, in reality, TNC platforms clearly dislike multi-homing——many actively discourage such behavior by implementing loyalty programs[3] or openly imposing penalty. Such a strong preference for an exclusive rather than shared workforce may be explained by the fact that multi-homing is as detrimental to the fundamentals of the platforms as to the efficiency and productivity of the entire system.

[3] `https://www.uber.com/newsroom/uberpro/https://www.uber.com/newsroom/uberpro/,https://www.lyft.com/rider/rewards`

6.3.2. Impact of regulations

In this section we repeat the experiments in the previous section but add a minimum wage constraint. The wage floor $\underline{e}$ in Eqs. (6.6) and (6.7) is selected as follows. We first solve a monopoly optimal pricing problem to maximize the social welfare Eq. (6.24). The "system optimal" wage rate obtained in this manner is then rounded to the whole dollar to get $\underline{e}$, as reported in Table 6.2. The main results are presented in Figure 6.3.



(a) Normalized welfare

(b) Market share

(c) Passenger wait time

(d) Vehicle time

Figure 6.3. System performance under the minimum wage policy: single-homing vs. multi-homing duopoly games.

As shown in Figure 6.3(a), the minimum wage policy does increase the social welfare under single-homing, but the net gain is rather modest. Although both passenger and driver surplus are improved substantially, these are largely gained at the expense of the platforms, and as a result, are likely to be wiped out over time. One option available to the platforms, as discussed in Section 5.3.2,

is to boost its profits by reducing the supply pool S_0. Such a response, however, would significantly undermine social welfare.

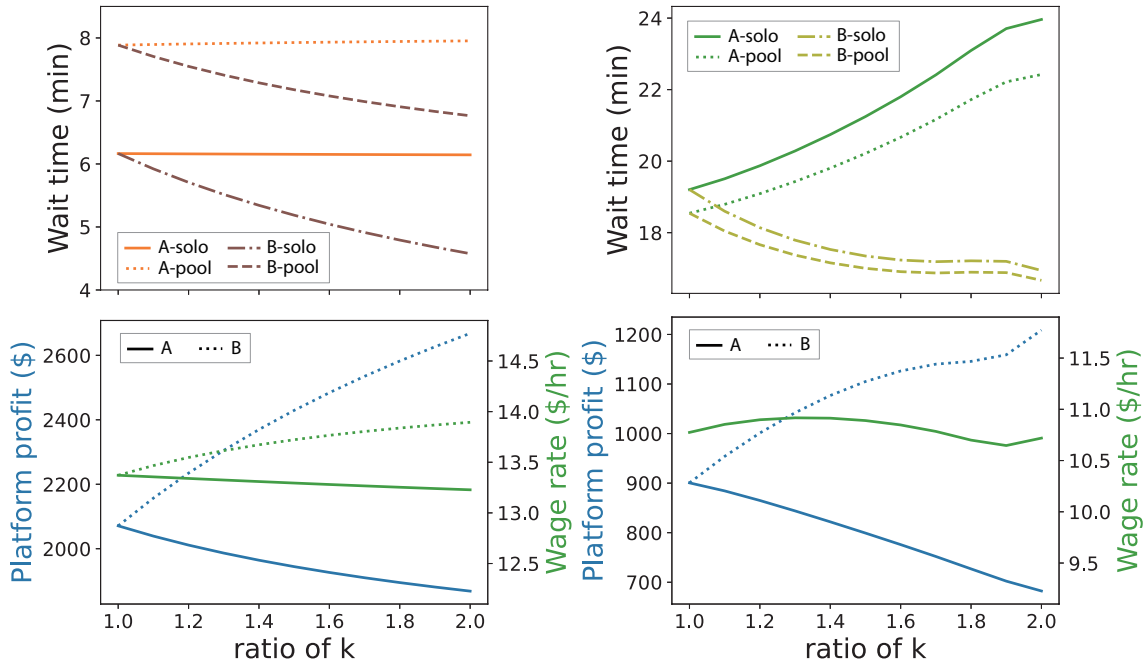
In the multi-homing duopoly, however, the minimum wage policy more than doubles the social welfare in most cases. Particularly, this improvement is achieved without sacrificing the platforms' profits. On the contrary, the platforms actually *benefit* from the policy, even though their gains are not as large as those of the passengers and drivers. The significant increase in the passenger surplus achieved by the minimum wage policy can be attributed to the dramatic improvement in passenger wait time; see Figure 6.3(c). Importantly, *the wage floor effectively prevents the market from the self-destructive price competition* found in Figure 6.2(b). As a result, more drivers are attracted to the ride-hail market by a higher average wage rate. In turn, the improved supply condition increases vacant vehicle time, lowers wait time, and finally grows the market share (Figure 6.3(b)). In a nutshell, *the minimum wage policy seems much more useful in a multi-homing duopoly than in a single-homing duopoly.*

As indicated in Figure 6.3(b), the effect of the minimum wage policy on pooling is detectable but not significant. For single-homing, the market share for pooling rides has a slight but clear dip in all but the case of high-demand-low-supply. This occurs despite the solo rides gain market share. In the multi-homing duopoly, both solo and pooling rides gain market share after the minimum wage policy is imposed, though most growth goes to solo rides. In a word, the minimum wage policy seems to favor solo over pooling rides, which echos the finding in Chapter 5.

We end this section by noting that the above analysis is limited to the short-term effect, when the platforms must cope with the extra supply induced by a minimum wage higher than the unregulated state. In the long term, the platforms could respond to this policy by reducing their driver pools (i.e., the potential supply S_0), as discussed in Section 5.3.2.

6.3.3. Asymmetric platforms

We now introduce asymmetric players into the duopoly game. First, we allow the two platforms in the unregulated pricing game to differ from each other on their competitive features: the matching efficiency k and the pooling efficiency b. For simplicity, we fix the parameter values of platform A while varying k^B and b^B, respectively. Figures 6.4 and 6.5 illustrate, respectively, the variation of the duopoly equilibrium as k^B/k^A and b^B/b^A increases from 1 (symmetric game) to 2 (highly asymmetric game).

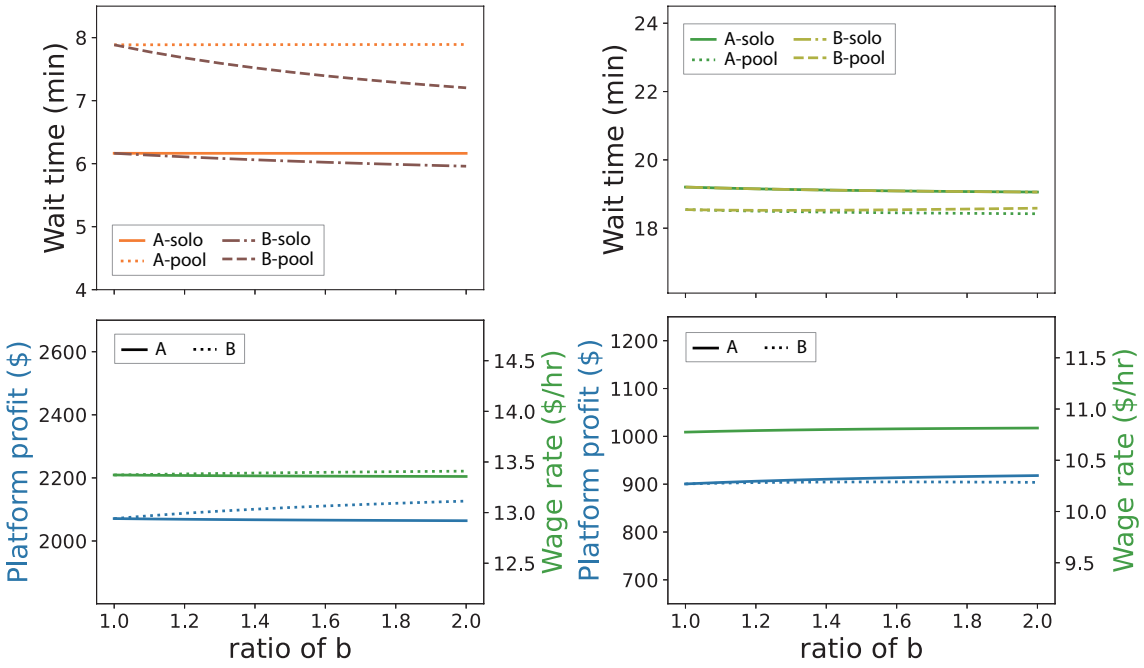


(a) Single-homing (b) Multi-homing

Figure 6.4. Sensitivity of the duopoly game to asymmetric matching efficiency. The ratio of k denotes k^B/k^A.

As expected, improving the matching efficiency substantially increases the platform's LOS and profit. However, its impact on the competing platforms varies with the supply mode. In the single-homing duopoly, both passenger wait time and driver wage rate on platform A are rather stable as platform B becomes increasingly more efficient. Platform A does gradually loses profits, but its loss amounts to merely 10% even when its rival becomes twice as efficient. On the contrary, a more

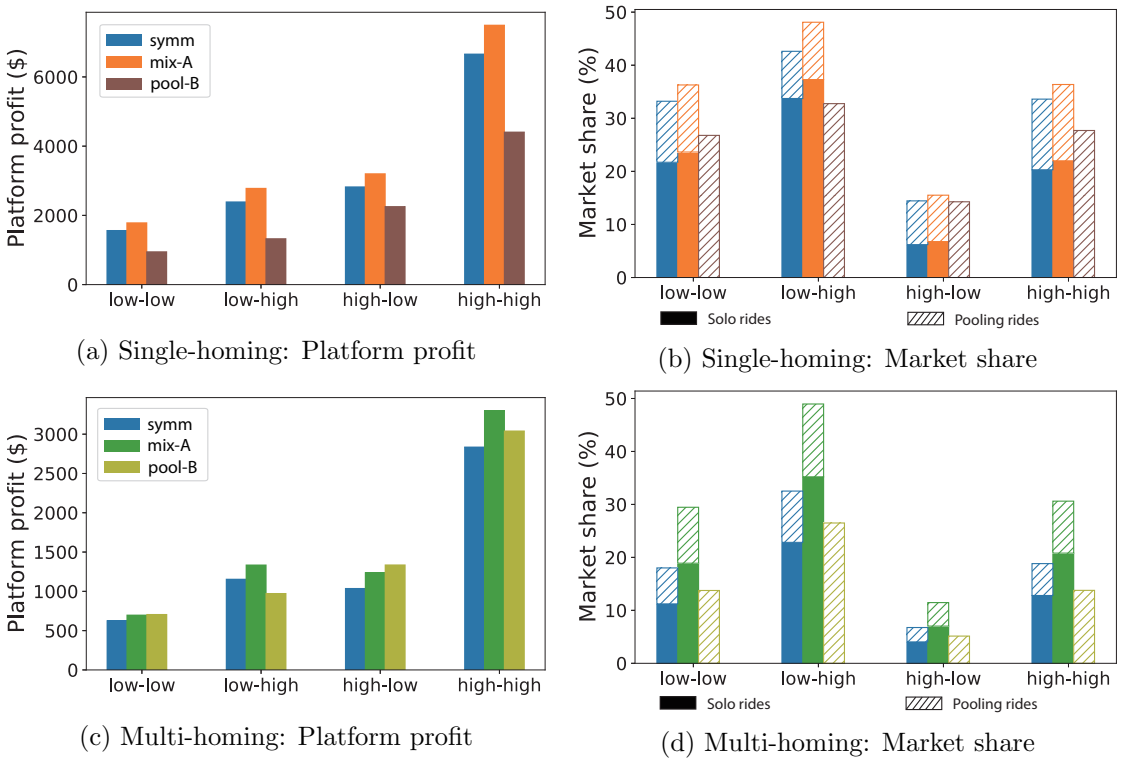
efficient competitor does far more harm to platform A in the multi-homing duopoly. As shown in Figure 6.4(b), the passenger wait time increases about 20% while the profit drops more than 20% on platform A when k^B/k^A doubles from 1 to 2. Because of the competitive edge it enjoys, platform B has less incentive to maintain sufficient vacant vehicle time or control inter-passenger congestion. Consequently, platform A, at a competitive disadvantage, suffers much more from the resulting supply-demand imbalance. Although it is similarly impeded by its relative inefficiency, *platform A in the single-homing duopoly manages to mitigate its losses because it has better control on the supply-demand relationship through pricing.*



(a) Single-homing (b) Multi-homing

Figure 6.5. Sensitivity of the duopoly game to asymmetric pooling efficiency b. The ratio of b is b^B/b^A.

Unlike k, the impact of the pooling efficiency b on the outcome of the duopoly game is rather small, as illustrated in Figure 6.5. For single-homing, when the pooling efficiency increases, the wait time on platform B decreases mildly for both pooling and solo rides, though the pooling rides enjoy a greater improvement. The higher pooling efficiency also slightly drives up profitability and wage rate for platform B. The effect of b in the multi-homing duopoly is almost negligible, likely



(a) Single-homing: Platform profit

(b) Single-homing: Market share

(c) Multi-homing: Platform profit

(d) Multi-homing: Market share

Figure 6.6. Duopoly equilibrium with asymmetric operational strategies.

because the first pickup time w_{p1} is much larger than the pickup detour w_{p2} (which decreases as b increases) due to the lower vacant vehicle density.

We finally consider asymmetry in terms of operational strategies. Specifically, we assume platform A offers both solo and pooling rides (i.e., a mixed-mode strategy) while platform B only serves pooling rides (a pure-pooling strategy). As some TNCs focus on providing pooling services (e.g., Via in the United States), it is interesting to examine how such a strategy fares against a competitor with a mixed service strategy. Figure 6.6 shows the performance of the two platforms (Mix-A vs. Pool-B) in an asymmetric duopoly game, and compares it with that of a platform in a symmetric game (with both platforms adopting a mixed strategy) .

In the single-homing duopoly, platform A always achieves a higher profit and market share, compared to platform B. The performance of the platform in a symmetric game lies right in

between the two. This finding is largely expected, as in Section 5.3 we have found the mixed strategy outperforms either the pooling or the solo strategy.

The results in the multi-homing duopoly, however, are far more intriguing. In the case of low-demand-low-supply, platform B achieves almost the same profit as platform A, while in the case of high-demand-low-supply, its profit is even higher than platform A. Moreover, platform B consistently earns a higher profit than the symmetric platform, except for the case of low-demand-high-supply. Interestingly, while platform B has a much lower market share than both platform A and the symmetric platform, the total market share of the asymmetric game is consistently higher than that of the symmetric game.

Because multi-homing effectively means the two platforms share the supply, platform B can afford a compensation rate much lower than that offered by platform A (remember drivers are assumed to focus on the average wage rate, rather than the wage rate of any individual platform). In this way, it can offer a reasonable pooling service at a rather low operating cost. On the other hand, since platform A essentially monopolizes the solo rides, it has a greater incentive to grow the market share, and attract more drivers to improve the LOS.

Therefore, multi-homing forges specialization in this game: platform B specializes in low-cost pooling service whereas platform A draws most revenues from high quality solo riders. This specialization helps explain why the asymmetric game tends to outperform its symmetric counterpart in terms of both market share and profits. It also suggests multi-homing likely encourages specialization in pooling under asymmetric conditions. We note that such specialized platforms do exist. For example, Via—an exclusive pooling service provider in the US—maintains a small but stable market share in big US cities (e.g., New York City[4]), where multi-homing is a prevalent phenomenon.

[4]See the market share of TNCs in New York City `https://toddwschneider.com/dashboards/nyc-taxi-ridehailing-uber-lyft-data/`

6.4. Summary

This chapter studies the inter-platform competition in an aggregate ride-hail market. The duopoly is built on the market equilibrium model developed in Chapter 4 with two supply modes, namely, single-homing and multi-homing, depending on whether or not each driver exclusively join one platform. The outcome of the duopoly pricing game is described as a Nash Equilibrium (NE) and solved by transforming it into a variational inequality problem (VIP). When a regulatory constraint is imposed, the duopoly equilibrium becomes a generalized NE, which corresponds to a quasi VIP. The main findings from the numerical experiments are summarized below:

- Without regulations, multi-homing may lead to a disastrous end. Specifically, passenger and driver surplus, as well as the platform profits, are all significantly lower in the multi-homing duopoly. This disaster arises because the platforms are locked in a self-destructive pricing war. Since drivers make decision to enter the market based on the average market (rather than platform) wage rate, the platforms soon discover lowering the payment to drivers to cut spending is better than raising it to attract more drivers. While the strategy makes senses individually, collectively it causes the collapse of the total supply and eventually the ruin of the business for all. This phenomenon may be interpreted as the tragedy of the commons, in which drivers are the "commons" shared and over-exploited by the platforms.
- While the minimum wage policy has a rather minor impact on the single-homing duopoly, it significantly improves the surplus for both passengers and drivers in the multi-homing duopoly, and does so without compromising platform profits. In essence, a suitable wage floor could resolve the pricing dilemma in the unregulated market, and thereby both platforms are able to attract a sufficient number of drivers to maintain a reasonable LOS. The minimum wage policy also tends to discourage pooling rides regardless of the supply mode, albeit the impact observed in our experiments is small.

- The matching efficiency seems a much more important asset in the competition compared to the pooling efficiency. In general, the platform with a higher matching efficiency ends up making more money and providing better LOS. Yet, the performance of the platform with a lower matching efficiency suffers much more in the multi-homing duopoly than in the single-homing duopoly.
- Offering both solo and pooling rides is a winning strategy in the monopoly and the single-homing duopoly. In the multi-homing duopoly, however, this winning strategy no longer holds a clear advantage. More surprisingly, having a platform only offer pooling rides in the multi-homing duopoly may improve the collective market share and profit. This counter-intuitive phenomenon can be attributed to the de-escalation of the pricing war achieved by service specialization.

In this chapter, we only consider two extreme supply modes: either all drivers are multi-homing, or none is. In reality, it is more likely that only a fraction of drivers are multi-homing. In Chicago, for instance, this fraction is about one quarter. Relaxing the definition multi-homing to allow drivers to join either one or both platforms hence constitutes an interesting direction for future research. Such a model not only moves one step closer to reality, but could also reveal a range of phenomena that have not arisen in the extreme cases studied herein. Another limitation of the present analysis has to do with the Nested Logit models used to represent passenger and driver choices. Despite their critical roles, these choice models have not been properly calibrated to match empirical data. Finally, the present study has left out other policies that have been discussed frequently in the literature and by policy makers, such as fleet cap and congestion tax. An interesting direction for future research is to design an "optimal" policy portfolio, similar to the joint implementation of minimum wage and congestion tax proposed in Chapter 5.

6.5. Appendix

6.5.1. Notations

Table 6.1. List of notations

Variable	Description	Unit
f_s^j (f_p^j)	trip fare of solo (pooling) rides of platform j	\$
w_s (w_p)	solo (pooling) passenger wait time of platform j	hr
τ_s (τ_p)	travel time of solo (pooling) rides	hr
η^j	compensation rate of platform j	\$/hr
e^j	driver's earning rate of platform j	\$/hr
$\underline{e}$	minimum wage	\$/hr
Q_s^j (Q_p^j)	solo (pooling) demand rate of platform j	/hr/sqmi
N^j	fleet size (number of drivers in operation) of platform j	/sqmi
D_0	total demand rate	/hr/sqmi
S_0	potential supply	/sqmi
$\theta_c(\theta_c^r)$	Mode choice uncertainty (among ride-hail options)	/\$
$\theta_d(\theta_d^r)$	Market entry uncertainty (among ride-hail platforms)	/\$
c_0	congestion cost of each ride-hail vehicle	\$
u_s^j (u_p^j)	generalized cost of solo (pooling) rides of platform j	\$
u_t	generalized cost of transit	\$
k^j	coefficient of matching efficiency of platform j	/sqmi
b^j	coefficient of pooling efficiency of platform j	

6.5.2. Parameter setting

Table 6.2. Representative market conditions.

Name	**Total demand** D_0 (/hr/sqmi)	**Potential supply** S_0 (/sqmi)	**Wage floor** $\underline{e}$ (\$/hr)
Low-low	500	200	18
Low-high	500	800	9
High-low	2000	200	26
High-high	2000	800	18

Part 3

Spatial market with pooling

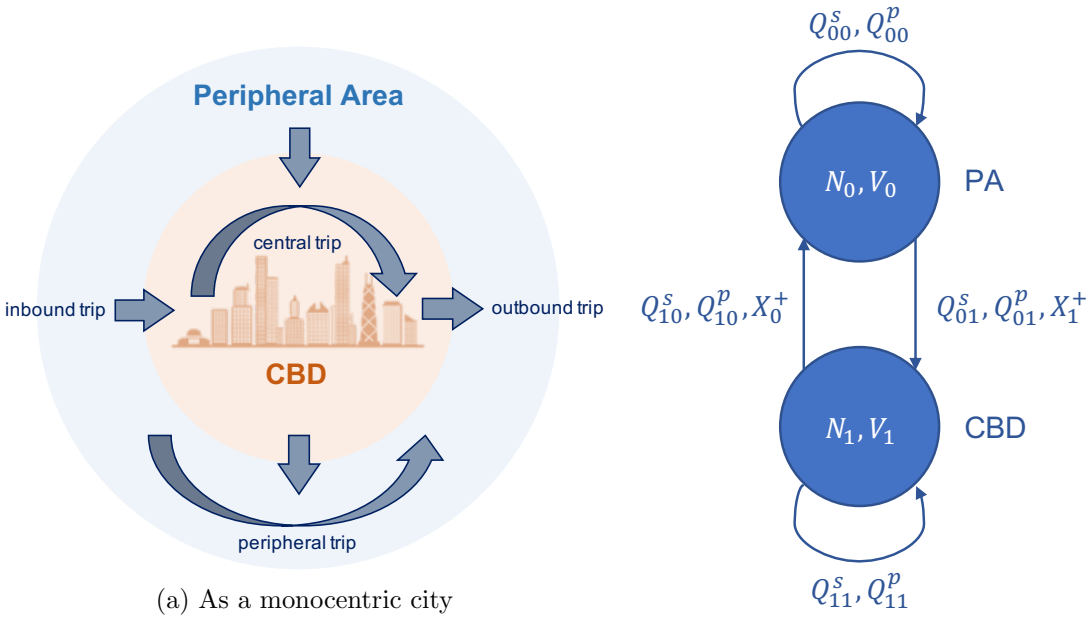
CHAPTER 7

A two-node model

In this chapter, we set out to consider a spatial ride-hail market. We start from a simple case: a city consists of two zones, namely, a central business district (CBD) and a peripheral area (PA), as shown in Figure 7.1(a). This simplification allows us to examine the spatial distribution of ride-hail vehicles and their congestion impact without modeling the sophisticated routing behaviors of drivers, which will be the main focus of the next chapter. For easy reference, we shall index PA and CBD as 0 and 1, and denote the set of zone indices as $\mathcal{I} := \{0, 1\}$. We abstract the city as a network of two zones connected by arterial roads, as illustrated Figure 7.1(b). Accordingly, four types of trips could take place in such a city: *central trips* that start and end within CBD, *peripheral trips* that start and end within PA, *inbound trips* that start in PA and end in CBD, and *outbound trips* that start in CBD and end in PA.

Assume that, for each market segment, the origins and destinations of all trips are uniformly distributed in the respective zone. Also, we assume a fixed portion of travelers drive alone, and the rest choose between transit and ride-hail service provided by a single TNC platform (*platform* hereafter) that monopolizes the market. Same as the analysis in previous chapters, we consider the platform offers both solo and pooling rides at different prices and each pooling ride is shared by two passengers.

In what follows, we first proposes a congestion model that links the trip duration and vehicle relocation time to ride-hail vehicle flows, and then adjust the spatial matching model developed in Chapter 3 to adapt to the spatial setting. After that, we will establish the market equilibrium and investigate the platform's pricing strategy in response to different congestion mitigation schemes. All notations used in this chapter are summarized in Table 7.1.



(a) As a monocentric city

(b) As two-node

Figure 7.1. Illustration of the spatial ride-hail market.

7.1. Congestion model

Let d^l_{ij}, $l \in \{s, p\}$ denote the average trip distance for mode m from zone i to zone j. Similarly, we define the average relocation distance as d^r_{ij}. The travel times are then given as

$$\tau^s_{ij} = d^s_{ij}/v_{ij} + \varepsilon_s, \quad i, j \in \mathcal{I}, \tag{7.1a}$$

$$\tau^p_{ij} = d^p_{ij}/v_{ij} + \varepsilon_p, \quad i, j \in \mathcal{I}, \tag{7.1b}$$

$$\tau^r_{ij} = d^r_{ij}/v_{ij}, \qquad i, j \in \mathcal{I};\ i \neq j, \tag{7.1c}$$

where ε_s and ε_p are additional travel time independent of trip distance and the matching process (e.g., the time spent in finding the TNC vehicle at the pickup location and the drop-off time). We assume $d^p_{ij} > d^s_{ij}$ because of the detour required to visit two destinations in pooling rides. While the en-route detour $d^p_{ij} - d^s_{ij}$ is expected to vary with pooling demand [Ke et al., 2021], we have shown the market equilibrium is generally insensitive to the impact of such dependency (see Section 4.8.3).

Hence, we shall treat d_{ij}^m as exogenous and calibrate them from empirical data. Similarly, ε_p and ε_s (with $\varepsilon_p > \varepsilon_s$) are also estimated from data.

We use Greenshields' fundamental diagram [Greenshields et al., 1935] and BPR function [US Bureau of Public Roads, 1964] to represent the dependence of speed on traffic within and between zones, respectively. Let v_f be the free-flow speed, $\boldsymbol{\rho}_i^{\text{jam}}$ be the jam density in zone i, and C_p be an aggregate road capacity between the two zones. The intra- and inter-zonal speeds are given by

$$\text{Intra-zone:} \quad v_{ii} = v_f \left(1 - \frac{\rho_i + \bar{\rho}_i}{\rho_i^{\text{jam}}}\right), \qquad i \in \mathcal{I}, \tag{7.2a}$$

$$\text{Inter-zone:} \quad v_{ij} = v_f \left[1 + 0.15 \left(\frac{z_{ij} + \bar{z}_{ij}}{C_p}\right)^4\right]^{-1}, \quad i, j \in \mathcal{I};\ i \neq j, \tag{7.2b}$$

where ρ_i ($\bar{\rho}_i$) is the ride-hail vehicle (background) density in zone i and z_{ij} ($\bar{z}_{ij}$) the ride-hail (background) vehicle flow from zone i to zone j. Note that all background traffic is produced by private motorists (i.e., those who do not choose of the three public modes). Using historical traffic and TNC data, we may calibrate the parameters used in Eq. (7.2) and estimate the magnitude of background traffic (i.e., $\bar{\rho}_i$ and $\bar{z}_{ij}$). More details are included in Chapter 10.

7.2. Adjusted matching model

Since each zone in the two-node model typically covers a large area, we assume waiting passengers are only matched vehicles in the same zone and passengers who shared one pooling trip must have the same origin and destination zone. Accordingly, the wait times of a passenger traveling from zone i to zone j are estimated as follows:

$$w_{ij}^s = \frac{\delta}{2v_{ii}} \sqrt{\frac{\Pi_i^{\text{eff}}}{k\Lambda_i}}, \qquad i, j \in \mathcal{I}, \tag{7.3a}$$

$$w_{ij}^p \simeq w_{ij}^s \sqrt{\frac{\kappa + 4b\Pi_{ij}^p}{2\kappa + 4b\Pi_{ij}^p}} + \frac{\delta}{2v_{ii}} \frac{1}{\sqrt{b\Pi_{ij}^p}}. \quad i, j \in \mathcal{I}, \tag{7.3b}$$

Here, we use w_{ij}^p to represent the sum of first pickup time and pick detour defined in Chapter 3 for the simplicity of notations. Also, since no variable is subscripted with j in Eq. (7.3a), we have $w_{ii}^s = w_{ij}^s$ for $i, j \in \mathcal{I};\ i \neq j$. Note that the effective waiting passenger density Π_i^{eff} and vacant vehicle density Λ_i are indexed by origin zone i whereas the pooling waiting passenger density Π_{ij}^p is indexed by the origin-destination (OD) pair. This is due to the assumption that the pooling trips are only assigned to passengers with same OD pair. In contrast, the competition for vacant vehicles is among all passengers in the same zone.

7.3. Market equilibrium

Same with the aggregate model (see Chapter 4), we assume the passenger mode choice depends on the generalized cost of solo and pooling trips, as well as transit. The costs are defined in the same way as Eq. (4.1) except that now the travel times are no longer exogenous but given by Eq. (7.1). Yet, the transit travel time is still set to be fixed and thus the generalized cost of transit is a constant.

The key difference of a spatial model from an aggregate model is the inter-zonal movement of vehicles. At equilibrium, drivers associated with each zone should operate at the same earning rate, meanwhile the vehicle inflow must equal the vehicle outflow for each zone. These two conditions dictate the vehicle flows between the two zones. At a stationary state, the net loss of occupied vehicle flow of zone i is given by

$$X_i = \left(Q_{ij}^s + \frac{1}{2}Q_{ij}^p\right) - \left(Q_{ji}^s + \frac{1}{2}Q_{ji}^p\right), \quad i, j \in \mathcal{I};\ i \neq j, \tag{7.4}$$

where Q_{ij}^l denotes the demand for mode $l \in \{s, p\}$ from zone i to zone j. Hence, a relocation vehicle flow $X_i^+ = \max(0, X_i)$ is needed to make up the loss such loss if necessary and is considered as part of vehicle supply of zone i.

With the same definition of demand and supply models in Chapter 4, the equilibrium of two-node model is characterized by the following system of equations:

$$\text{Mode choice:} \quad Q_{ij}^m = D_{ij}^0 q(u_{ij}^m; u_{ij}^{-m}), \qquad i,j \in \mathcal{I}, \tag{7.5a}$$

$$\text{Fleet size:} \quad N = S_0 g(e), \tag{7.5b}$$

$$\text{Flow conservation:} \quad N_i = V_i + \sum_{j' \in \mathcal{I}} Q_{ij'}^s \tau_{ij'}^s + \frac{1}{2} \sum_{j' \in \mathcal{I}} Q_{ij'}^p \tau_{ij'}^p + X_i^+ \tau_{ji}^r, \qquad i,j \in \mathcal{I};\ i \neq j, \tag{7.5c}$$

$$X_i = \left(Q_{ij}^s + \frac{1}{2} Q_{ij}^p \right) - \left(Q_{ji}^s + \frac{1}{2} Q_{ji}^p \right), \qquad i,j \in \mathcal{I};\ i \neq j, \tag{7.5d}$$

$$\text{Passenger density:} \quad \Pi_i^{\text{eff}} = \frac{1}{A_i} \sum_{j \in \mathcal{I}} \left(Q_{ij}^s w_{ij}^s + \frac{1}{2} Q_{ij}^p w_{ij}^p \right), \qquad i \in \mathcal{I}, \tag{7.5e}$$

$$\Pi_{ij}^p = \frac{1}{A_i} Q_{ij}^p w_{ij}^p, \qquad i,j \in \mathcal{I}, \tag{7.5f}$$

$$\text{Vehicle density:} \quad \Lambda_i = \frac{1}{A_i} V_i, \qquad i \in \mathcal{I}, \tag{7.5g}$$

$$\text{Wait time:} \quad w_{ij}^s = \frac{\delta}{2 v_{ii}} \sqrt{\frac{\Pi_i^{\text{eff}}}{k \Lambda_i}}, \qquad i,j \in \mathcal{I}, \tag{7.5h}$$

$$w_{ij}^p = w_{ij}^s \sqrt{\frac{\kappa + 4b\Pi_{ij}^p}{2\kappa + 4b\Pi_{ij}^p}} + \frac{\delta}{2 v_{ii}} \frac{1}{\sqrt{b\Pi_{ij}^p}}, \qquad i,j \in \mathcal{I}, \tag{7.5i}$$

$$\text{Traffic speed:} \quad v_{ii} = v_f \left[1 - \frac{1}{\rho_i^{\text{jam}} A_i} \left(Q_{ii}^s \tau_{ii}^s + \frac{1}{2} Q_{ii}^p \tau_{ii}^p + V_i + \bar{Q}_{ii} \tau_{ii}^s \right) \right], \quad i \in \mathcal{I}, \tag{7.5j}$$

$$v_{ij} = v_f \left[1 + 0.15 \left(\frac{Q_{ij}^s + \frac{1}{2} Q_{ij}^p + X_j^+ + \bar{Q}_{ij}}{C_p} \right)^4 \right]^{-1}, \qquad i,j \in \mathcal{I};\ i \neq j, \tag{7.5k}$$

where A_i denotes the area of zone i and $\bar{Q}_{ij}$ denote the background travel demand.

Eq.(7.5) shares the same structure as Eq.(4.5) for the aggregate market equilibrium, except for the additional flow conservation for each zone (Eq. (7.5d)) and the equations of traffic speed (Eqs. (7.5j) and (7.5k)). Specifically, Eqs. (7.5j) and (7.5k) rewrite Eqs. (7.2a) and (7.2b) by specifying vehicle density and flow. For the intra-zonal speed of zone i, the vehicle time consists of four parts: $Q_{ii}^s \tau_{ii}^s$ (contributed by solo rides in zone i), $0.5 Q_{ii}^p \tau_{ii}^p$ (contributed by pooling rides in zone i), V_i (contributed by vacant ride-hail vehicles), and $\bar{Q}_{ii} \tau_{ii}^s$ (contributed by background trips). Here, we assume the average travel time of background traffic is identical to that of solo rides.

Also, since the analysis period is normalized to a unit time, the total vehicle time divided by zone area A_i yields the vehicle density in the zone. Likewise, the cross-zone vehicle flow from i to j consists of the following: Q_{ij}^s (solo trip flows from i to j), $0.5Q_{ij}^p$ (pooling trip flows from i to j), X_j^+ (relocation trip flow from i to j), and $\bar{Q}_{ij}$ (background trip flow).

Due to the similar structure, it is easy to show Eq. (7.5) can be reduced to a fixed-point system $\mathbf{x} = F(\mathbf{x})$, where $\mathbf{x} = (\mathbf{w}_s, \mathbf{w}_p, \mathbf{v}) \in \mathbb{R}^{12}$. The proof of existence also follows Proposition 2 since the newly added variable $\mathbf{v}$ naturally satisfies the boundedness condition ($v_{ij} \in [0, v_f]$).

7.4. Platform pricing without regulations

Compared to the pricing problem in an aggregate market, the decision variables expand to be the trip fares between each OD pair for each mode f_{ij}^l, $l \in \{s, p\}$, along with the compensation rate η. For a profit-maximizing platform, the optimal pricing problem can be formulated as

$$\max_{\mathbf{f}_s, \mathbf{f}_p, \eta} \quad R = \frac{1}{\sum_i A_i} \left[\mathbf{f}_s^T \mathbf{Q}_s + \mathbf{f}_p^T \mathbf{Q}_p - \eta \left(\boldsymbol{\tau}_s^T \mathbf{Q}_s + \frac{1}{2} \boldsymbol{\tau}_p^T \mathbf{Q}_p \right) \right]. \tag{7.6a}$$

If the platform aims to maximize social welfare, the objective function turns to be

$$W = \frac{1}{\sum_i A_i} \left(\mathbf{D}_0^T \bar{\mathbf{u}} + \mathbf{f}_s^T \mathbf{Q}_s + \mathbf{f}_p^T \mathbf{Q}_p - \int_0^N g^{-1}(n/S_0) \mathrm{d}n - \nu \bar{\mathbf{Q}}^T (\boldsymbol{\tau}_s - \boldsymbol{\tau}_0) + T \right). \tag{7.7}$$

Recall that, in Chapters 5 and 5, we use a constant c_0 to represent the congestion externality of each ride-hail vehicle. Here, we explicitly model the congestion effect by the term $\nu \bar{\mathbf{Q}}^T (\boldsymbol{\tau}_s - \boldsymbol{\tau}_0)$, which computes the total loss of background travelers due to the operation of ride-hail service. Specifically, ν denotes the value of time and $\boldsymbol{\tau}_0$ is the travel time experienced by private motorists if the ride-hail service is not operated in the city.

The last term T in Eq. (7.7) refers to the tax revenue from the congestion policies, when applicable.

Above two unconstrained optimization problem can be solved in the same way as that proposed in Chapter 5, i.e., differentiate the equilibrium at the current solution, use the partials to construct

the gradient, then update the solution via gradient ascent. However, there are issues the original solution method. First, the equation of relocation flow Eq. (7.5d) is not smooth, thus the gradient tends to have a considerable jump when X_i approaches to zero. Second, as the number of decision variables increases, a uniform moving step may not be appropriate given that the gradient with respect to each variable could vary a lot.

To deal with the first problem, we introduce a SoftPlus function to approximate X_i^+, i.e.,

$$X_i^+ \simeq \hat{X}_i^+ = \ln\left(1 + e^{X_i}\right). \tag{7.8}$$

It is easy to see that $\hat{X}_i^+ \to 0$ as $X_i \to -\infty$ while $\hat{X}_i^+ \to X_i$ as $X_i \to +\infty$. The evaluation of $\hat{X}_i^+$ may become numerically unstable when the magnitude of X_i^+ is too large. This problem is resolved by introducing an adjustable scalar γ to scale down X_i^+ whenever necessary, i.e.,

$$\hat{X}_i^+ = \gamma \ln(1 + e^{X_i/\gamma}). \tag{7.9}$$

In the solution procedure, γ is adjusted according to X_i such that $X_i/\gamma \leq 10$.

To stabilize and fasten the solution procedure, we implement Adam [Kingma and Ba, 2014], an adaptive optimization algorithm that makes use of moment information to adaptively scale the "learning rate" for each decision variable. With $\nabla R^{(n)}$ obtained in the nth iteration, the Adam algorithm updates the current solution as follows:

$$m_1^{(n+1)} = \beta_1 m_1^{(n)} + (1 - \beta_1)\nabla R^{(n)}, \tag{7.10a}$$

$$m_2^{(n+1)} = \beta_2 m_2^{(n)} + (1 - \beta_2)\left(\nabla R^{(n)}\right)^2, \tag{7.10b}$$

$$\hat{m}_1 = \frac{m_1^{(n+1)}}{1 - \beta_1^n}, \tag{7.10c}$$

$$\hat{m}_2 = \frac{m_2^{(n+1)}}{1 - \beta_2^n}, \tag{7.10d}$$

$$\mathbf{y}^{(n+1)} = \mathbf{y}^{(n)} + \alpha \frac{\hat{m}_1}{\sqrt{\hat{m}_2} + \varepsilon}, \tag{7.10e}$$

where $\nabla R^{(n)}$ is the gradient with respect to the solution in nth iteration, β_1 and β_2 are exponential decay rates for the moment estimates m_1 and m_2, $(\cdot)^2$ is an element-wise square operator, α is the step size, and ε is a scalar to avoid zero denominator. In words, Adam maintains m_1 and m_2 as exponential moving averages of ∇R and $(\nabla R)^2$, respectively (Eqs. (7.10a) and (7.10b)). The two moment estimates are then rescaled (Eqs. (7.10c) and (7.10d)) to obtain the next solution (Eq. (7.10e)). In this study, we set β_1 and β_2 as 0.9 and 0.999, respectively, and $\varepsilon = 10^{-8}$, following the common practice. The step size α, however, is problem specific and has to be set in a trial-and-error fashion to optimize performance.

Finally, to avoid being trapped at a "bad" solution, the solution procedure is usually repeated with multiple initial points, as we did in the previous analysis.

7.5. Platform pricing under regulations

In this section, we discuss three congestion mitigation policies and explain how to model their interactions with the platform's operation. The first two policies impose trip-based or cordon-based congestion fees, whereas the third implements a cruising cap [Schaller, 2017a, NYC Taxi and Limousine Commission, 2019].

7.5.1. Trip-based congestion fee

In order to encourage pooling and transit, major U.S. cities like New York City and Chicago have started to charge a congestion fee on TNC trips starting and/or ending in a designated zone during peak hours. The charge is set higher for solo rides than pooling rides. To simplify the analysis, we assume a congestion fee c is charged on solo rides only and is internalized into the trip fare of each solo ride[1]. Hence, the generalized cost for a solo ride becomes

$$u_{ij}^s = f_{ij}^s + \nu(w_i^s + \tau_{ij}^s) + c_{ij}^s, \quad i, j \in \mathcal{I}, \tag{7.11}$$

[1]Passing such fees to customers is a common practice among TNC platforms, see e.g., `https://www.uber.com/blog/new-york-city/congestion-surcharge/` (Accessed: 2020-09-28).

where $c^s_{ij} = 0$ if $i = 0, j = 0$ (i.e., when the trip both starts and ends in the PA zone), otherwise it equals c.

Since the congestion pricing has no impact on drivers' revenue, the supply model does not change. Besides, the tax revenue is $T = \sum_{ij\in\mathcal{I}} c_{ij} Q^s_{ij}$.

7.5.2. Cordon-based congestion fee

A commonly used congestion pricing strategy, though rarely implemented in ride-hail, is the cordon-based congestion fee. Under this policy, each vehicle must pay a toll equal to c if (i) it relocates from PA to CBD without carrying any passengers, or (ii) it delivers a solo trip from PA to CBD. For Case (ii), c is passed on to the trip fare. Hence, c^s_{ij} in Eq. (7.11) becomes c when $i = 0, j = 1$ and 0 otherwise. For Case (i), the congestion fee becomes part of drivers' operation cost. Therefore, the expected earning of a driver becomes

$$e = \frac{1}{N_i}\left[\eta \sum_{j\in\mathcal{I}} \left(Q^s_{ij}\tau^s_{ij} + \frac{1}{2}Q^p_{ij}\tau^p_{ij}\right) - c^d_i X^+_i\right], \quad i \in \mathcal{I}, \tag{7.12}$$

where $c^d_i = c$ if $i = 1$, and 0 otherwise. Accordingly, the tax revenue becomes $T = \sum_{ij\in\mathcal{I}} c^s_{ij} Q^s_{ij} + \sum_{i\in\mathcal{I}} c^d_i X^+_i$.

7.5.3. Cruising cap

Vacant vehicle cruising is considered an important contributor to the worsening traffic congestion in city centers [e.g., Erhardt et al., 2019, Beojone and Geroliminis, 2021]. In New York City, the total vacant vehicle hours had grown by 81% from 2013 to 2017, the majority of which is attributed to the operation of TNC vehicles [Schaller, 2017a]. In an attempt to curtail further growth, the city proposed a regulation in 2019 that requires any TNC platform to ensure its fleet spend no more than 31% of total operating time without passenger. However, the policy was challenged in the court and later revoked by the state judge [Guse, 2019].

For the purpose of comparison, we implement a version of the cruising cap policy as follows. We first calculate the vehicle utilization rate in CBD as

$$\mu_1 = 1 - \frac{V_1}{V_1 + Q_{11}^s \tau_{11}^p + \frac{1}{2} Q_{11}^p \tau_{11}^p}. \tag{7.13}$$

Let $\underline{\mu}$ be the lower bound on the utilization rate. Thus, the cruising cap is given by $1 - \underline{\mu}$.

Since the cruising cap imposes a constraint on the equilibrium variables, the platform's pricing problem becomes a constrained optimization problem. The regulatory constraint can be written as $h(\mathbf{y}) = \underline{\mu} - \mu_1(\mathbf{y}) \leq 0$ and the resulting problem can be solved by a dual gradient ascent algorithm as follows:

$$\mathbf{y}^{(j+1)} = \arg\max_{\mathbf{y}} \mathcal{L}(\lambda^{(j)}, \mathbf{y}) = \arg\max_{\mathbf{y}} R(\mathbf{y}) - \lambda h(\mathbf{y}), \tag{7.14}$$

$$\lambda^{(j+1)} = \max\left(0, \lambda^{(j)} + \rho h(\mathbf{y}^{(j+1)})\right), \tag{7.15}$$

where λ represents the Lagrangian multipliers and ρ is a constant penalty parameter. In each iteration, the subproblem Eq. (7.14) is first solved using the method discussed above. Then, λ is updated using Eq. (7.15).

7.6. Numerical experiments

We continue to use the empirical data collected in the City of Chicago for the numerical experiments. In Figure 7.2, the orange area shows the identified CBD zone, which covers 29 census tracts and aligns with the congestion zone defined by the city's recent congestion pricing policy [2]. The remaining area is then classified as the PA zone. Based on this construction, the CBD zone has an area of 6.4 square mile (sqmi) and the PA zone's area is about 22.3 sqmi. The default values of parameters are reported in Table 7.2, while details about the parameter estimation are included in Section 10.2.

[2]See `https://www.chicago.gov/city/en/depts/bacp/supp_info/city_of_chicago_congestion_pricing.html`.

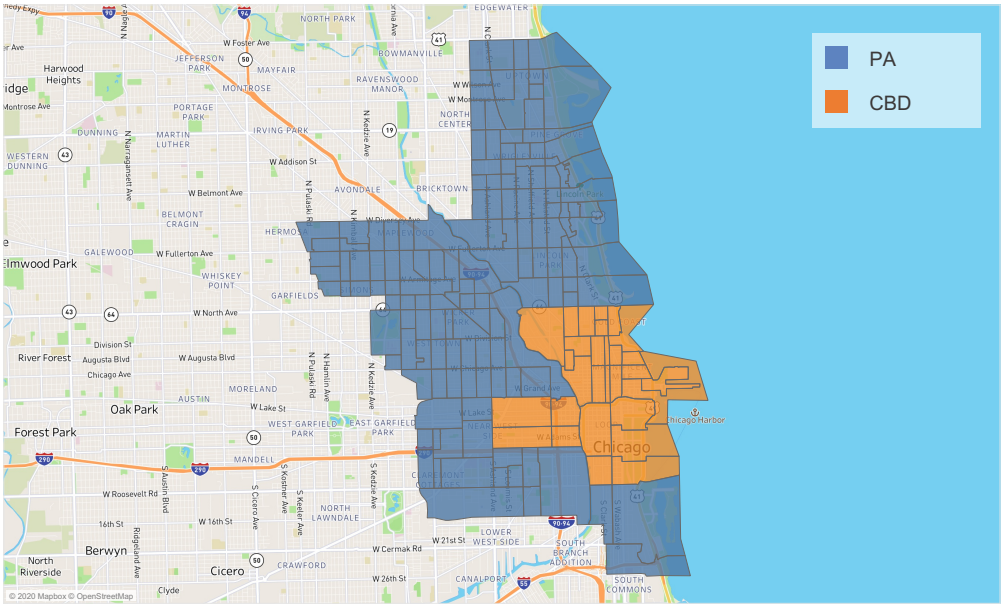


Figure 7.2. Study area.

Similar to the analysis of aggregate market, we characterize passenger's mode choice with a Multinomial Logit (MNL) model and assume driver's reservation rate follows a uniform distribution. Differently, in the current experiment, we have the reservation rate varies over a range $[\underline{e}_0, \overline{e}_0]$. Accordingly, Eq. (7.5b) can be rewritten as

$$(\overline{e}_0 - \underline{e}_0)N^2 + \underline{e}_0 S_0 N - eS_0 = 0, \tag{7.16}$$

from which $N = \sum_{i\in\mathcal{I}} N_i$ can be easily solved as the positive root. Finally, the driver's opportunity cost defined in Eq. (5.10) is reduced to

$$\int_0^N G^{-1}(n/S_0)\mathrm{d}n = \frac{\overline{e}_0 - \underline{e}_0}{2S_0}N^2 + \underline{e}_0 N. \tag{7.17}$$

7.6.1. Unregulated market

We first compare the system performance under profit- and welfare-maximization pricing in an unregulated market. For simplicity, they are referred to as monopoly (MO) and social optimum (SO) pricing, respectively. For MO pricing, we consider another case where the platform fails to

anticipate the impact of its operation on traffic congestion. That is, it optimizes the pricing policy based on a default, flow-independent speed. However, the actual effect of this "sub-optimal" pricing policy will be evaluated according to a market equilibrium that does characterize traffic congestion. We are particularly interested in the following question: *should the platform "care" about traffic congestion even if it is completely motivated by self-interest (i.e., profit)?*

Figure 7.3 illustrates the overall ride-hail market share and supply level in the three studied scenarios: "w/ traffic" as MO pricing considering traffic congestion, "w/o traffic" as MO pricing without considering traffic congestion, and "SO" as SO pricing considering traffic congestion. The market share is defined as the percentage of the potential passengers ($\sum_{i,j\in\mathcal{I}} D_{ij}^0$) who opt for ride-hail service, whereas the supply level is the percentage of potential drivers (S_0) who join the workforce of the platform. As expected, when traffic congestion is ignored, the platform tends to hire more drivers and serves more trips. In contrast, at system optimum, the ride-hail market is much smaller in both demand and supply. The drop in vehicle supply is especially steep: almost half of the drivers would leave the market under an SO pricing regime.

Figure 7.4 plots the ride-hail market share by OD pair and mode. First, under no circumstance is pooling observed in the peripheral trips. Pooling within the PA zone is unattractive because the demand is low and scattered. The former leads to long wait time, while the latter results in long en-route detour. Second, the share is the lowest for central trips, thanks to a much more competitive transit service in this area. This is especially true under SO pricing. In this case, almost no central trips are served by ride-hail vehicles. Evidently, given the quality of transit services available in Chicago downtown, the net utility generated by switching from transit to ride-hail is barely enough to offset the negative traffic impact. Third, for inter-zonal markets, ride-hail performs much better, in some case capturing well over 50% of the market share (e.g., for outbound trips). Moreover, the majority of these inter-zonal trips are pooled. On the other hand, the pooling ratio is lower under MO pricing without traffic, especially for trips originated in CBD. Since solo rides require more vehicle time, failing to anticipate the traffic impact is likely to inadvertently worsen traffic

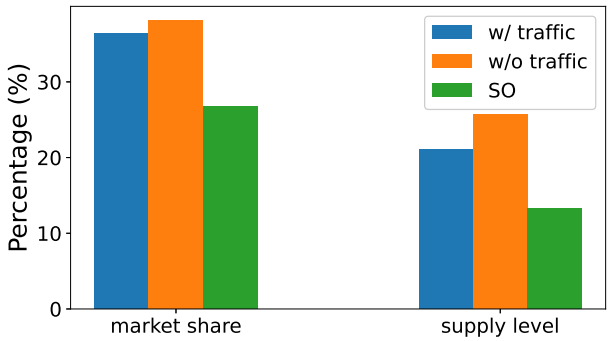


Figure 7.3. Market share and supply level.

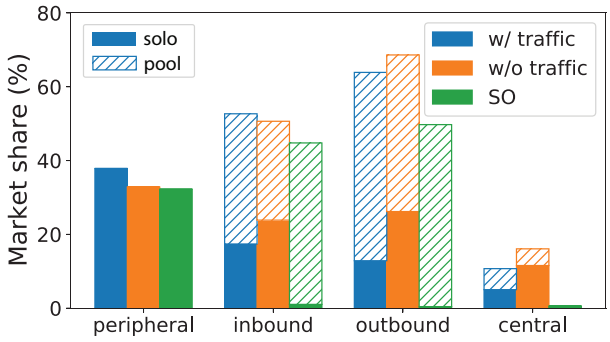


Figure 7.4. Share by market segments.

congestion in CBD. Finally, the SO pricing strongly encourages pooling. In fact, in both inter-zonal market segments, nearly all trips are pooled.

Figure 7.5 reports the four components of social welfare under the three pricing policies. As expected, MO pricing with traffic delivers the highest profit to the platform, while the profit generated by SO pricing is the lowest. Passengers and drivers both benefit the most from MO pricing without traffic, because the platform tends to hire more drivers in this case, which in turn improves the ride-hail LOS. SO pricing, on the other hand, minimizes the traffic impact of ride-hail at the expense of other stakeholders. The biggest losers are the drivers, whose surplus is reduced to but a fraction of what they would receive under MO pricing.

As illustrated in Figure 7.5, in all scenarios, traffic congestion plays the major role in the social welfare: its value is comparable to that of the other three combined in all three cases. In

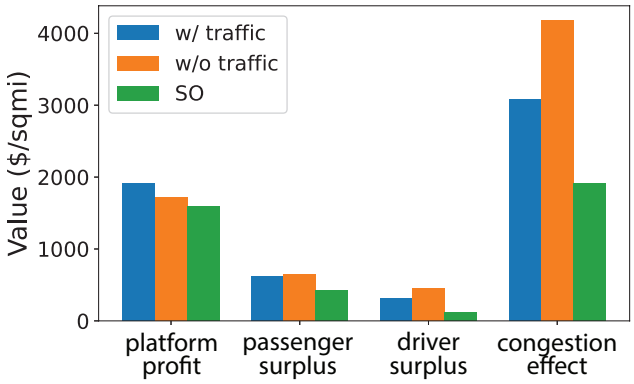


Figure 7.5. Components of social welfare.

particular, for MO pricing without traffic, the sum of net benefits gained by the platform, passengers and drivers is not nearly enough to make up for the loss due to worsened congestion.

Figure 7.6 compares travel speed in the three scenarios. It shows the operation of ride-hail service lowers the travel speed by up to 5 mph. Although the overall speed reduction seems modest, the cumulative effect is nonetheless considerable because of the large background traffic ($\bar{\mathbf{Q}}$ in Eq. (7.7)). Figure 7.6 also indicates that the largest reduction occurs inside CBD, which agrees with recent empirical findings [Erhardt et al., 2019, Diao et al., 2021]. Compared to MO pricing, SO pricing has the most modest impact on traffic speed. It thus causes a significantly smaller congestion effect in social welfare (see Figure 7.5).

Figure 7.7 plots both occupied and vacant VMT inside each zone and between zones. It suggests that a significant amount of vacant VMT in both CBD and PA. This is expected, as a certain amount of vacant vehicles is needed in each zone to maintain a reasonable LOS. On the other hand, the vacant VMT between zones is due to vehicle relocation. It is interesting to see that relocation trips are only observed in the case of MO pricing without traffic. This phenomenon indicates that the pricing policy fails to balance the demand and supply across zones. Specifically, by overlooking congestion, the platform attracts more drivers into CBD than what the demand in the zone could adequately consume. Consequently, at equilibrium, some drivers would relocate to PA, suffering a dead-weight loss in the process. This problem is rooted in the sub-optimal nature of MO pricing without traffic.

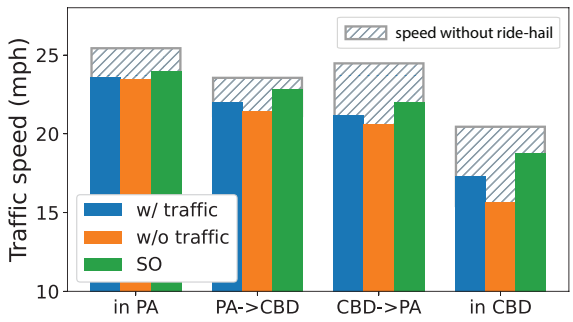


Figure 7.6. Travel speed inside and between zones.

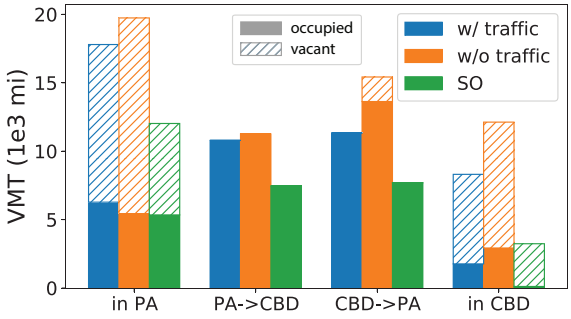


Figure 7.7. Ride-hail vehicle miles traveled (VMT) inside and between zones.

Note that, in Figure 7.7, the occupied VMT inside CBD and PA only accounts for the intra-zonal trips. Hence, when computing the vehicle occupancy rate associated with each zone market, one need to include the occupied VMT of inter-zonal trips as well. Consequently, the occupancy rate in the CBD zone is greater than the PA zone, though this is not clearly illustrated in the figure. Overall, SO pricing produces the lowest total vacant VMT whereas MO pricing without traffic leads to the highest. The difference is more significant in CBD. As shown in Figure 7.4, a profit-maximization platform (i.e., MO pricing) tends to serve more trips originated from CBD. It thus requires more vacant vehicle time inside CBD to support a desired LOS.

7.6.2. Regulated market

We proceed to investigate the impact of the three congestion mitigation policies discussed in Section 7.5. For both trip-based and cordon-based policies, the congestion fee varies from \$0.2 to \$2. The cruising cap ranges between 0.3 and 0.6, which is equivalent to a minimum occupancy between 0.4 and 0.7. The results without regulations are also included to benchmark the performance of tested policies. Specifically, the result of MO pricing with traffic represents the status quo, while that of SO pricing represents the "desirable outcome".

7.6.2.1. Market share. Figure 7.8 shows the total market share of ride-hail service rises under the two congestion pricing policies, though the growth is rather mild (less than 2%). With the cruising cap, the market share first increases in parallel with the congestion pricing policies, but

begins to drop precipitately when the minimum occupancy exceeds 0.55. However, it remains well above the level achieved by SO pricing, even when a minimum occupancy of 0.7 is imposed.

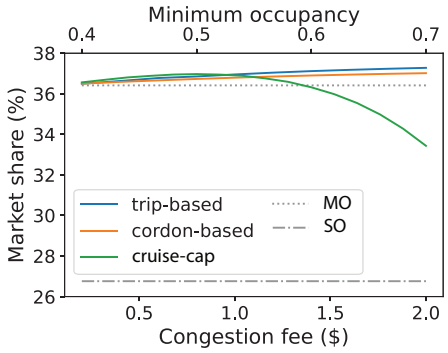
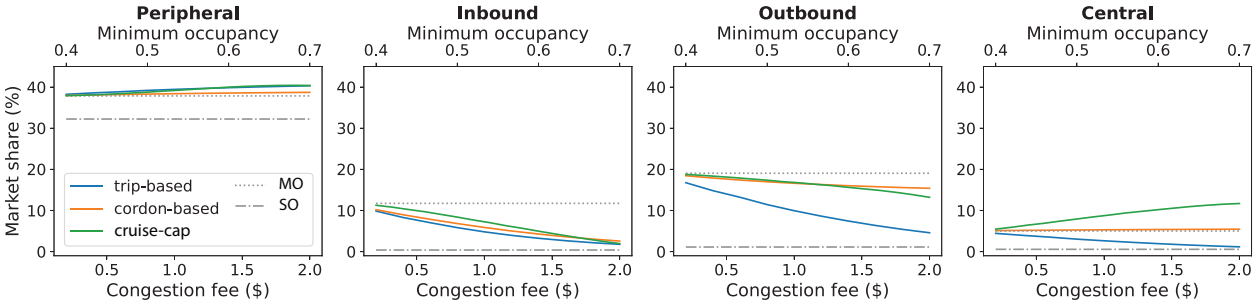


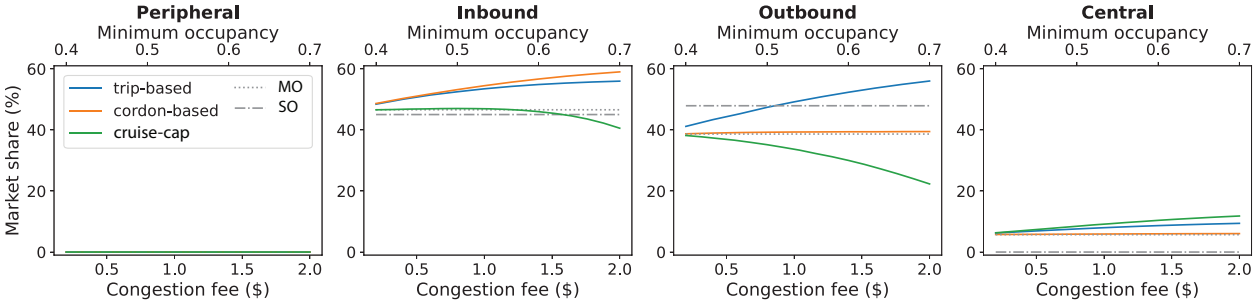
Figure 7.8. Ride-hail market share

While the total market share is relatively stable, the shares between different OD pairs vary a lot with service mode and congestion mitigation policies, as reported in Figure 7.9. The share of the peripheral trips hardly changes because these trips are not hit by any of the policies. For all other three O-D pairs, a large portion of solo rides are replaced by pooling rides under the trip-based policy. The cordon-based policy, however, only produces a similar effect on inbound trips as its trip-based counterpart. It does cause a small number of trips to switch from solo to pooling for outbound trips, but has almost no impact at all on central trips. These findings are expected, because the trip-based congestion fee targets solo passengers who travel to, from and in CBD whereas the cordon-based fee only affects inbound travelers. The performance of the cruising cap is peculiar. The regulation forces the platform to sacrifice its market share in the inter-zonal trips in exchange for an uptick in both solo and pooling rides in CBD. *This is not a desired outcome because the extra trips it attracts could have used a descent transit service.*

7.6.2.2. Level of service and congestion relief. Figures 7.10 and 7.11 plot passenger wait time and vacant vehicle density in PA and CBD, respectively. The wait time is an average weighted by demand for each mode and OD pair. The most intriguing finding here is that passengers have to endure a much longer wait time when a cruising cap is imposed. Those traveling from CBD see their wait times jump from around six minutes to well above 10 minutes as the minimum



(a) Solo trips



(b) Pooling trips

Figure 7.9. Market share by service mode and OD pair.

occupancy grows from 0.4 to 0.7. The increase in wait time is accompanied by a decline in vacant vehicle density, which is steeper as the cap becomes more restrictive (see Figure 7.11).

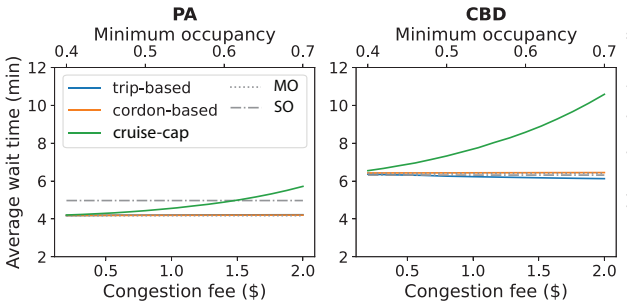


Figure 7.10. Passenger wait time by zone.

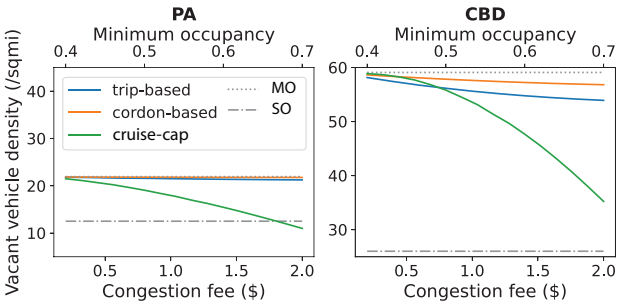


Figure 7.11. Vacant vehicle density by zone.

Interestingly, the average wait time remains almost intact under both congestion pricing policies. The vacant vehicle density, however, had a more noticeable dip, especially in CBD. At first glance, this result is somewhat puzzling, because one would expect a lower vacant vehicle density to prolong waiting. It also appears to contradict with Li et al. [2020], who find trip-based fee hurts the LOS in the peripheral area. We believe the discrepancy can be explained by the availability of pooling

in our model. As revealed in Figure 7.9, more inbound passengers end up choosing to pool, which helps maintain a sufficient vehicle supply in PA. The lower vacant vehicle density in CBD under the two congestion pricing policies are also attributed to a greater share of pooling rides.

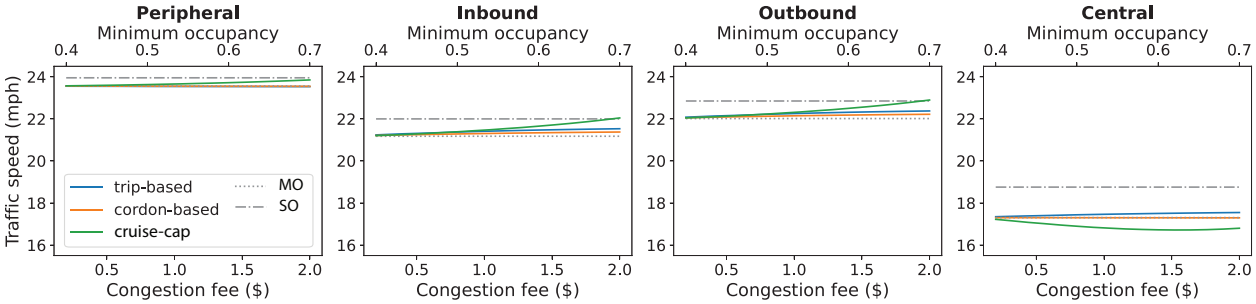


Figure 7.12. Travel speed by OD pair.

Another observation from Figure 7.11 is how little the two congestion pricing policies had done to move vacant vehicle density towards the SO level. This is true in both PA and CBD, although the gap is much larger in the latter. *Thus, these policies may not be effective instruments to reduce vacant vehicle cruising.* As suggested in Figures 7.12 and 7.13, their contribution to traffic relief is also minor. Under both trip- and cordon-based pricing policies, the improvement in travel speed across all market segments is barely visible. Similarly, vacant VMT were little affected, though occupied VMT were reduced much more thanks to the greater reliance on pooling. As for the cruising cap, it achieves vacant VMT at the SO level when the minimum occupancy requirement reaches 0.7. Yet, the occupied VMT is still far away from that at SO. Furthermore, the reduced vacant VMT only helps improve the inter-zonal travel speed. As shown in Figures 7.12, *the congestion in CBD becomes even worse.*

7.6.2.3. Fare and compensation. Figure 7.14 reveals how the platform's pricing strategy varies with the congestion mitigation policies[3]. We can see that the fares under the two congestion pricing policies are quite similar to those under MO pricing. Hence, the shift in demand is largely caused by the congestion fee directly charged on passengers. In contrast, to meet the cruising cap requirement, the platform has to more actively manipulate the price. Notably, it aggressively

[3]The pooling trip fare for peripheral trips is not plot because no trip is pooled.

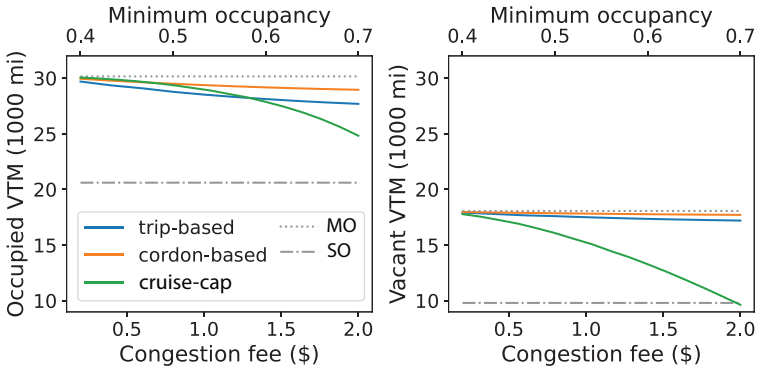
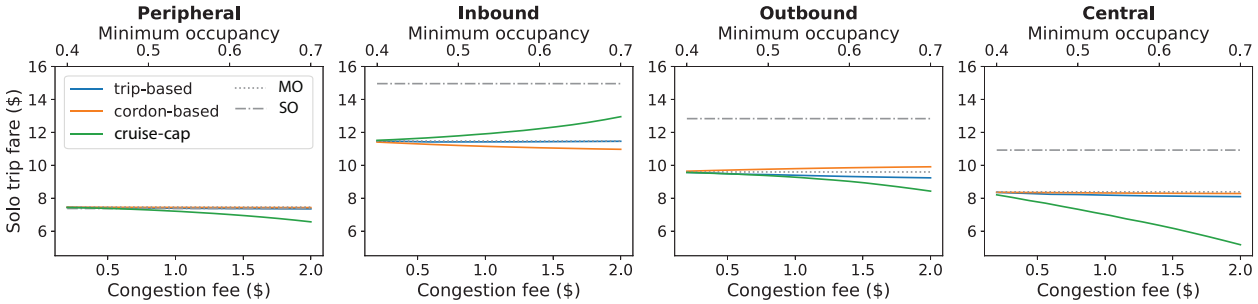


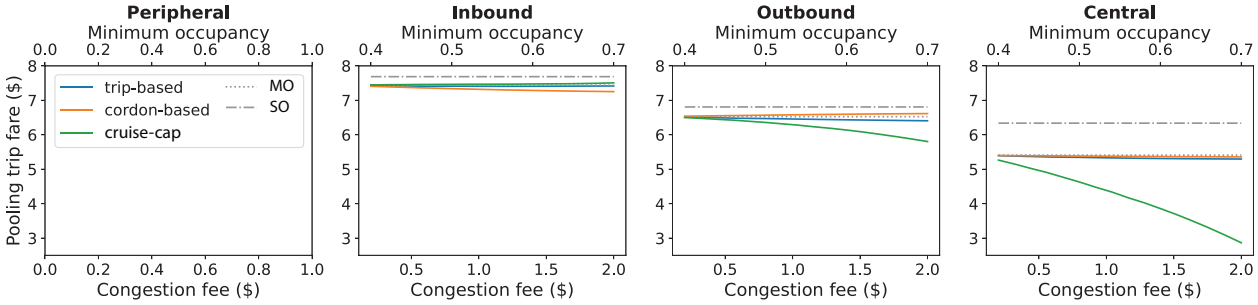
Figure 7.13. VMT by vehicle status.

lowers the price for both solo and pooling trips in CBD to make them more attractive such that the occupancy rate could be sustained at the required level. Meanwhile, it increases the fare for inbound trips—especially the solo ones—and reduces that for outbound trips. This maneuver is meant to discourage inbound travel and encourage outbound travel. Intuitively, moving vehicles out of CBD on occupied trips could help reduce vehicle cruising in CBD and thus release the pressure of cruising cap. However, this strategy does not attract more demand (see Figure 7.9), due primarily to the deteriorating LOS (see Figure 7.10).

Figures 7.15 and 7.16 plot the compensation rate paid to drivers and the vehicle supply associated with each zone, respectively. It is clear that drivers are paid much less at SO, which is the direct cause of a smaller vehicle supply in the market. While the driver's earning is reduced slightly by the two congestion pricing policies, it is still much higher than the SO level. Under the cruising cap, however, the earning plunges. Despite this suffering, the cruising cap does not effectively control vehicle supply in CBD except when the cap is highly restrictive. A mild cap actually induces more vehicles to enter CBD because the platform seeks to serve more trips there under the circumstance (see Figure 7.9). In contrast, under SO pricing, almost all trips inside CBD are served by transit. As indicated in Figure 7.14, such an outcome requires the platform to raise price on both solo and pooling rides, while cutting the compensation to drivers. However, none of the three mitigation policies could achieve such an outcome, not even close.



(a) Solo trips



(b) Pooling trips

Figure 7.14. Trip fare by service mode and OD pair.

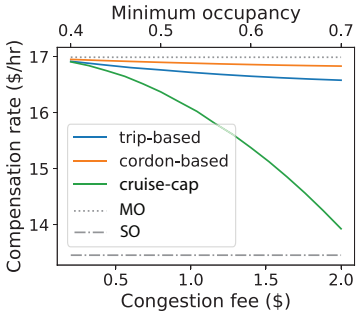


Figure 7.15. Compensation rate.

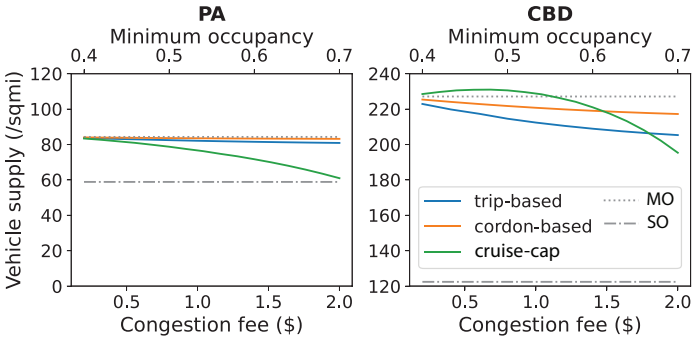
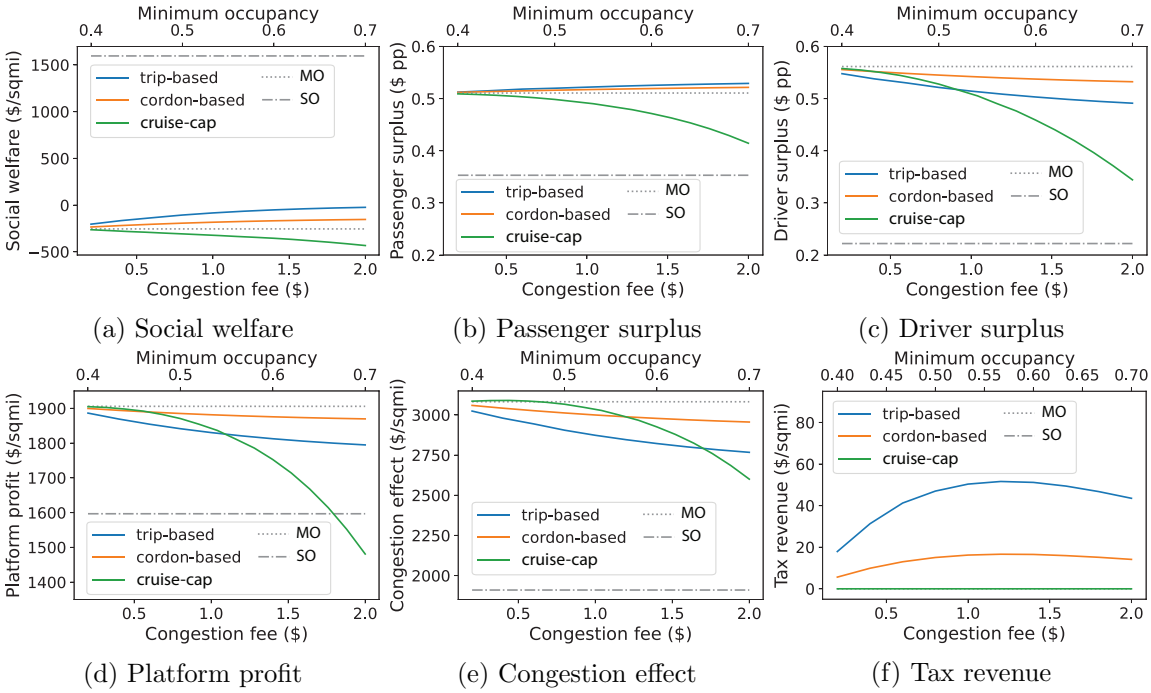


Figure 7.16. Vehicle supply by zone.

7.6.2.4. Discussions. In this section, we further discuss the performance of the three tested policies by comparing their welfare effects. As seen in Figure 7.17(a), both trip- and cordon-based pricing policies yield a net gain in social welfare, whereas the cruising cap leads to a net loss. Although the trip-based fee stands out as a clear winner, its overall welfare outcome still falls far behind that under SO pricing.



(a) Social welfare (b) Passenger surplus (c) Driver surplus

(d) Platform profit (e) Congestion effect (f) Tax revenue

Figure 7.17. Social welfare by part.

A closer look at different welfare components reveals that the two congestion pricing policies give ride-hail passengers a modest benefit at the expense of drivers and the platform (see Figures 7.17(b)-(d)). The cruising cap, however, hurts all three stakeholders. With the most restrictive cap, the driver surplus decreases almost 40% and the platform loses over 20% of its profit. Indeed, in this case the platform earns even less than under SO pricing.

At 2$/ride, the trip-based fee realizes about a quarter of the reduction in the congestion cost achieved by SO pricing, more than doubling the improvement obtained by the cordon-based fee (see Figure 7.17(e)). This finding seems surprising at first glance given the small improvement in travel speed (see Figure 7.12). Yet, it is important to note that the traffic generated by TNC rides is a rather small fraction of the total vehicular traffic (around 5%). Hence, whatever gains in travel speed would be magnified by a large number of background trips. Indeed, even the SO pricing can only increase the average speed by 2 mph. Despite its seemingly more dramatic impact on vacant VMT and travel speed, the cruising cap policy did not bring about an overwhelming congestion relief. Even at the extreme (minimum occupancy = 0.7), its congestion savings is still well below

50% of what is obtained by SO pricing. With milder restrictions on vehicle occupancy rate (below 0.65), the cruising cap actually trails behind the trip-based policy. This seemingly counter-intuitive result is likely due to the negative impact of cruising cap on travel speed in CBD (see Figure 7.12). Since most trips are concentrated in the city center, each unit of decrease in average speed in CBD inflicts a large external cost. Accordingly, the extra cost induced by the more intense congestion in CBD offsets the savings from PA and inter-zonal trips.

Finally, the trip-based policy is good at generating tax revenues. At every congestion fee tested, it collects more than twice as much revenue as the cordon-based policy does (see Figure 7.17(f)). This finding is also expected as trips between more OD pairs are charged under trip-based pricing compared to cordon-based policy. Interestingly, for both policies, the tax revenue is maximized when the congestion fee is around $1.25.

To summarize, *trip-based pricing seems to be the best of the three regulations considered herein. It improves social welfare, benefits passengers, reduces the congestion cost and generates a substantial amount of tax revenues.* Curiously, drivers and the platform are always losers, regardless of which congestion mitigation policies is adopted. Their losses are the smallest under cordon-based pricing. On the other hand, *the cruising cap policy delivers the worst outcomes. Even though it improves traffic speed outside CBD, the policy does not benefit anybody or generate any tax revenue.* Furthermore, the cruising cap hits drivers and the platform much harder than the other two, especially when a highly restrictive cap is implemented. As a result, the system is worse off in terms of social welfare, compared to the unregulated benchmark.

7.7. Summary

In this chapter, we analyzed the impact of three congestion mitigation policies on an idealized two-node model. Similar to the analysis of the aggregate ride-hail market, we assume a single platform serving both solo and pooling trips and passengers choose among the two ride-hail services and transit. The market equilibrium is build upon two sub-models. One is the matching model

extended from what is presented in Chapter 3 to account for the spatial heterogeneity in demand pattern and restriction in matching. The other is an aggregate traffic model, which describe the congestion impact of ride-hail vehicles on all travelers in the urban transport system. Based on the equilibrium model, we formulate the platform's optimal pricing problem and propose a solution algorithm to address the additional complexity due to the spatial equilibrium model. Main findings from the numerical experiments are summarized as follows:

- TNC operations have a considerable impact on traffic congestion, even though they contribute a rather small fraction (below 5%) to the the total traffic. As a result, failing to anticipate congestion in its pricing decisions leads to sub-optimal decisions that not only worsen traffic congestion by inducing extra supply, but also hurt the platform's profitability.
- Welfare-maximization pricing strongly encourages pooling in all but peripheral trips. In the city center, it raises the price on both solo and pooling rides in favor of transit. This is because the net utility generated by switching from transit to TNC in the city center, where the transit service has a high LOS, is hardly enough to offset the negative traffic impact. In contrast, profit-maximization pricing tends to serve more central trips that pool less. To do so, more vehicles must be induced to the city center, which exacerbates overall traffic conditions.
- The trip- and cordon-based congestion mitigation policies substantially reduce the extra congestion cost caused by unregulated TNC operations, even though the positive impact on average travel speeds is barely discernible. They also promote pooling, which helps cut the occupied vehicle miles traveled. While the cruising cap imposed in the city center effectively brings down the total vacant vehicle miles traveled, it does not offer an overwhelming congestion relief. Paradoxically, the policy could worsen traffic conditions in the city center, where most trips originate, as it pushes the platform to serve more trips there.

- Of the three policies considered herein, the trip-based policy delivers the best overall performance. It lifts the system's social welfare while keeping the LOS of ride-hail almost intact. Although it slightly disadvantage the platform and drivers, the benefits it creates for all travelers, in the form of congestion relief, consumer surplus and tax revenue, more than make up for this loss. The cordon-based policy demonstrates similar behaviors as the trip-based policy, but trails behind in nearly all metrics. The cruising cap policy is a surprising disappointment. While the private motorists benefit from the extra congestion relief the policy promises to deliver, the other stakeholders together suffer a much greater loss.

In this study, we simplify the city as a two-node model since most congestion mitigation policies in practice classify the congestion region in the same way. To better analyze the spatial variations of demand and supply, our next step is to construct a more general network model. Such an extension would complicate the modeling of vehicle relocation behaviors, whether it is driven by drivers' self-interest or coordinated by the platform. The following chapter aims to tackle precisely this problem. Once a network equilibrium model is established, we could set out to design congestion mitigation policies with a more flexible structure (e.g., allow the congestion fee to vary spatially and/or temporally).

7.8. Appendix

7.8.1. Notations

Table 7.1. List of notations

Variable	Description	Unit
τ_{ij}^s (τ_{ij}^p)	average trip duration of solo (pooling) rides	hr
τ_{ij}^r (τ_{ij}^0)	vehicle relocation time (baseline vehicle travel time)	hr
v_{ij}	traffic speed of trips	mph
d_{ij}^s (d_{ij}^p)	average trip distance of solo rides (pooling rides)	hr
v_f	free-flow traffic speed	mph
C_p	capacity of arterial roads between zones	/hr
ε_{ij}^s (ε_{ij}^p)	distance-independent trip duration of solo (pooling) rides	hr
ρ_i ($\bar{\rho}_i$, ρ_i^{jam})	ride-hail (background, jam) vehicle density	/sqmi
z_{ij} ($\bar{z}_{ij}$)	ride-hail (background) vehicle flow	/hr
w_{ij}^s (w_{ij}^p)	wait time of solo (pooling) passengers	hr
Λ_i	vacant vehicle density	/sqmi
Π_i^s (Π_{ij}^p, Π_i^{eff})	solo (pooling, effective) waiting passenger density	/sqmi
k	coefficient of matching efficiency	/sqmi
b	coefficient of pooling efficiency	
κ	approximation parameter in function of w_{ij}^p	/sqmi
δ	detour ratio of road network	
Q_{ij}^s (Q_{ij}^p)	solo (pooling) demand rate	/hr
$\bar{Q}_{ij}$	background traffic flow	/hr
X_i	difference between outbound and inbound occupied vehicle flow of zone i	/hr
X_i^+ ($\hat{X}_i^+$)	(approximated) vehicle relocation flow to zone i	/hr
N_i	fleet size (number of drivers in operation)	
V_i	vacant vehicle time	hr
D_{ij}^0	total demand rate	/hr
S_0	potential supply	
A_i	area of zone i	sqmi
f_{ij}^s (f_{ij}^p)	trip fare of solo (pooling) rides	\$
η	compensation rate (payment per unit occupied time)	\$/hr
u_{ij}^s	generalized cost of solo rides	\$
ν	value of time	\$/hr
c_{ij}^s	congestion fee on solo ride	\$
c_i^d	congestion fee on relocation vehicles to zone i	\$
$\bar{u}_{ij}$	average saving of a trip due to switching from transit to ride-hail service	\$
e	driver's earning rate	\$/hr
$\bar{e}_0$ ($\underline{e}_0$)	upper (lower) bound of reservation rate	\$/hr
μ_i	vehicle occupancy rate	

[a] Subscription ij means "from zone i to zone j" when not specified

[b] Subscription i means "in zone i" when not specified

7.8.2. Parameter setting

Table 7.2. Default values of parameters.

Parameter		Unit	Default value
Total demand	$\mathbf{D}_0$	pax/hr	$34440 \times \begin{bmatrix} 0.18 & 0.23 \\ 0.21 & 0.38 \end{bmatrix}^*$
Average solo trip distance	$\boldsymbol{d}_s$	mi	$\begin{bmatrix} 2.6746 & 3.6353 \\ 3.7550 & 1.5392 \end{bmatrix}$
Average pooling trip distance	$\boldsymbol{d}_p$	mi	$\begin{bmatrix} 3.8519 & 4.1357 \\ 4.2476 & 2.0211 \end{bmatrix}$
Average transit trip duration	$\boldsymbol{\tau}_t$	hr	$\begin{bmatrix} 0.34 & 0.48 \\ 0.47 & 0.31 \end{bmatrix}$
Transit trip fare	f_t	\$/ride	2.69
Passengers' value of time	ν	\$/hr	27.69
Relative disutility rate of transit	ζ	\$/hr	6.92
Mode choice uncertainty	θ		1
Minimum reservation rate	$\underline{e}_0$	\$/hr	7.25
Maximum reservation rate	$\bar{e}_0$	\$/hr	31.37
Potential supply	S_0	veh	15785
Free-flow speed	v_f	mph	40
Jam density	$\boldsymbol{\rho}_{\text{jam}}$	veh/sqmi	$\begin{bmatrix} 2000 & 3000 \\ 3000 & 4000 \end{bmatrix}$
Aggregate road capacity	C_p	veh/hr	30000.
Background traffic	$\bar{\boldsymbol{Q}}$	veh/hr	$\begin{bmatrix} 81219 & 44034 \\ 43055 & 73690 \end{bmatrix}$
Traffic speed without ride-hail vehicles	$\bar{\boldsymbol{v}}_0$	mph	$\begin{bmatrix} 25.4330 & 23.5815 \\ 24.4450 & 20.4262 \end{bmatrix}$
Default traffic speed in MO pricing without congestion model	$\bar{\boldsymbol{v}}$	mph	$\begin{bmatrix} 25 & 22 \\ 23 & 19 \end{bmatrix}$
Detour ratio of road network	δ		1.3
Matching efficiency	k	/sqmi	0.16
Pooling efficiency	b		0.05
Approximation parameter	κ		4
Zone area	$\mathbf{A}$	sqmi	$[22.3, 6.4]$

* All parameters in vector form follow index {0: PA; 1: CBD} and those in matrix form follow index {(0,0): peripheral; (0,1): inbound; (1,0): outbound; (1,1): central}.

CHAPTER 8

Routing game of strategic drivers

In the two-node model presented in Chapter 7, drivers are assumed to relocate between the two local markets such that their expected earnings are equal. At equilibrium, the relocation flow is uni-directional and can be easily solved from flow conservation. However, it becomes a challenge to model the strategic relocation behaviors of drivers in a general spatial market, particularly when the time dynamics are also introduced.

In this chapter, we study the RIde-hail VEhicle Routing (RIVER) problem in a dynamic spatial ride-hail market. We consider a group of homogeneous drivers traveling across a network of local markets to search and deliver passengers. The routing decision is characterized by a search strategy that gives the search destination of each idle driver in each time period, and the return of a certain strategy is computed as the expected reward earned over time. Specifically, the expected reward in each time period is dictated by two factors. The first is the trip fare from the search destination, which is set by the ride-hail operator and considered exogenous in the current analysis. The second is the probability of successfully picking up a passenger by the end of the time period, referred to as *meeting probability*, which depends on both the passenger demand and vehicle supply in the same local market. In other words, each driver's reward is affected by the behaviors of other drivers. This fact builds up the nature of a game, where, as players, drivers select the optimal strategy in consideration of others' strategies to maximize their own utilities.

In what follows, we first present the RIVER model and discuss its connection to cooperative routing, where drivers make routing decisions to maximize total return. Using the physical matching model discussed in Chapter 3, we then derive the meeting probability for two common ride-hail modes and calibrate them through simulations. In this way, we are able to compare the service

performance of different service modes and analyze the loss of efficiency due to selfish routing. All notations used in this chapter are summarized in Table 8.5.

8.1. System dynamics

We represent a ride-hail market as a set $\mathcal{N}$ of zones and discretize the analysis horizon into T time periods with an identical length Δ. We further assume the travel time from zone i to zone j, denoted as τ_{ij}, is a multiple of Δ. Let q_i^t be the demand rate (i.e., the number of passenger arrivals per unit area per unit time) originated from zone i during time period $t = 1, \ldots, T$, and α_{ij} be the fraction of the demand from i destined for j. Thus, $\sum_j \alpha_{ij} = 1,\ \forall i$.

The market is served by a set of drivers, denoted as $\mathcal{M}$ (with $M = |\mathcal{M}|$), who travel across zones to search and deliver passengers. We assume a driver always drop off passengers at the end of a period and thus they become idle when the next period starts. At the beginning of each period, idle drivers choose their next search zone. They could either stay in the current zone or relocate to a *neighboring* zone. To simplify the analysis, we assume the time spent to move from one zone to an adjacent zone is much less than Δ and hence can be rounded to zero. In other words, relocation can be completed instantaneously[1]. Let x_{ij}^t be the vacant vehicular flow from zone i to zone j, and y_j^t be the number of vacant vehicles in zone j at the beginning of time period t. Thus, the first flow conservation condition dictates

$$y_j^t = \sum_i x_{ij}^t, \quad t = 1, \cdots, T. \tag{8.1}$$

We model *matching* as a process that assigns idle drivers in each zone to passengers originating from the same zone. It starts at the beginning of each period, taking the numbers of arriving passengers and vacant vehicles as inputs, and yields the matched and idle drivers at the end of the period according to a matching mechanism. The meeting probability during period t in zone j,

[1]Relocation differs from delivering a passenger between adjacent zones in that it is completed once the vehicle crosses the zonal boundary. In contrast, a passenger trip ends when the vehicle finally reaches the passenger's destination.

denoted as m_j^t, is thus a function of q_j^t, y_j^t and Δ, whose form depends on the characteristics of specific ride-hail matching technology. While the mathematical form of m_j^t could vary, it is safe to assume, everything else equal, a larger number of vacant vehicles always leads to a lower meeting probability and a higher demand rate always yields a higher meeting probability. Besides, it is reasonable to assume the vacant vehicle time in a zone has a diminishing marginal effect on the pickup number my in that zone. These properties are formally stated in the following assumption, while the functional form of m will be specified in Section 8.4 for different ride-hail modes.

Assumption 7. *The meeting probability function $m(q, y)$ satisfies $\partial m/\partial y \leq 0$, $\partial m/\partial q \geq 0$, and $\partial^2(my)/\partial y^2 \leq 0$.*

At the end of period t, $m_j^t y_j^t$ drivers in zone j are successfully matched and ready to deliver passengers to their destinations. The others will make a new search decision at the beginning of time $t+1$, along with the drivers dropping off passengers in the zone at the end of t. Hence, the second flow conservation condition requires

$$\sum_k x_{jk}^t = x_j^t = \begin{cases} d_j^0, & t = 1 \\ d_j^{t-1} + (1 - m_j^{t-1}) y_j^{t-1}, & t = 2, \cdots, T \end{cases}, \tag{8.2}$$

where x_j^t denotes the total vacant vehicular flow departing from zone j at the beginning of time t, d_j^t denotes the flow of vehicles completing trips in zone j by the end of t. Specifically, d_j^0 denotes the number of vacant vehicles in zone i at the beginning of the analysis horizon.

Since the trip duration depends on the origin-destination (OD) pair, d_j^t equals the sum of occupied vehicular flows departing from each zone $i \in \mathcal{N}$ at the beginning of period $t - \tau_{ij}$, i.e.,

$$d_j^t = \sum_{t'=0}^{t} \sum_{i \in \mathcal{I}(j,t,t')} m_i^{t'} y_i^{t'} \alpha_{ij}, \quad t = 1, \cdots, T-1, \tag{8.3}$$

where the set of trip origin zones is defined as $\mathcal{I}(j, t, t') := \{i : \tau_{ij}/\Delta = t - t'\}$.

For convenience, we assume OD travel pattern α_{ij}, travel time τ_{ij} and trip fare p_{ij} are time-independent. However, it is straightforward to allow temporal variations in travel demand, traffic

conditions and pricing policies in our framework. This assumption is also relaxed in the numerical experiments presented in Section 8.5.

8.2. Individual routing

8.2.1. MDP formulation

We model each vehicle's movements across zones as a Markov decision process (MDP). The MDP framework is commonly used in the recommendation of optimal search path for ride-hail drivers [e.g., Shou et al., 2020, Yu et al., 2019b] as it accounts for the payoff that may realize *later* due to *current* decisions. An MDP is defined by a tuple $(\mathcal{S}, \mathcal{A}, P, R, \gamma)$ as follows:

- $\mathcal{S}$ is a set of states in which the vehicle may settle in a given period. $\mathcal{S}$ can be divided into vacant states and occupied states. When the vehicle is vacant, the state is determined by its current zone i, i.e., $s = (\text{vac}, i)$. When the vehicle is occupied, the state is determined by the OD pair of its passenger, and the time elapsed (measured in Δ) since it starts the trip, i.e., $s = (\text{occ}, i, j, \delta)$, where i (j) is the origin (destination) zone and $\delta = 0, \dots \tau_{ij} - 1$.
- $\mathcal{A}$ is a set of actions. For vacant vehicles, each action a is defined as the next zone to visit. Hence, a vehicle at state $s = (\text{occ}, i)$ has an action set consisting of zone i and its neighbor zones, denoted as $\mathcal{A}_i$. For an occupied vehicle, the only action available is to continue the current trip, denoted as $a = 0$.
- P denotes the state transition probability distribution. Based on the state types, there are four groups of transition probabilities, including

  vacant $\rightarrow$ vacant:

$$P(s_{t+1} = (\text{vac}, j) | s_t = (\text{vac}, i), a_t = j) = \begin{cases} 1 - m_j^t, & j \in \mathcal{A}_i \\ 0, & \text{otherwise} \end{cases}, \tag{8.4a}$$

  vacant $\rightarrow$ occupied:

$$P(s_{t+1} = (\text{occ}, j, k, 0) | s_t = (\text{vac}, i), a_t = j) = \begin{cases} m_j^t \alpha_{jk}, & j \in \mathcal{A}_i \\ 0, & \text{otherwise} \end{cases}, \tag{8.4b}$$

occupied $\rightarrow$ occupied:

$$P(s_{t+1} = (\text{occ}, i, j, \delta + 1) | s_t = (\text{occ}, i, j, \delta), a_t = 0) = 1, \quad \delta < \tau_{ij} - 1, \tag{8.4c}$$

occupied $\rightarrow$ vacant:

$$P(s_{t+1} = (\text{vac}, j) | s_t = (\text{occ}, i, j, \tau_{ij} - 1), a_t = 0) = 1. \tag{8.4d}$$

Note any transition starting at an occupied state is deterministic as a vehicle must complete its current trip once started.

- R defines the reward for the current state s_t, an action a_t, and a next state s_{t+1}. Here, the trip fare p_{ij} is set as the reward for a transition from vacant to occupied states, formally defined as

$$R(s_t, a_t, s_{t+1}) = \begin{cases} p_{jk} & , \ s_t = (\text{vac}, i), a_t = j, s_{t+1} = (\text{occ}, j, k, 0) \\ 0 & , \ \text{otherwise} \end{cases}. \tag{8.5}$$

- $\gamma \in (0, 1]$ is a discount factor that accounts for the preference for a present gain over a future one.

Let π be a *search strategy* that maps the current state s to a vector of choice probabilities, each corresponding to a feasible action a_t. For occupied states, π is a singleton as there is only one feasible action, i.e., continue the current trip ($a_t = 0$). For vacant states, π can be seen as a probability distribution over the feasible actions. Accordingly, we have $\pi_{ij}^t = Pr(a_t = j | s_t = (\text{vac}, i)) \geq 0$ and $\sum_j \pi_{ij}^t = 1$.

We use $V^{\pi}(s_t)$ to denote the expected return of a driver from state s_t following a strategy π. The Bellman equation is then written as

$$V^{\pi}(s_t) = E_{a_t \sim \pi} \left\{ \sum_{s_{t+1}} P(s_{t+1}|s_t, a_t) \left[R(s_t, a_t, s_{t+1}) + \gamma V^{\pi}(s_{t+1}) \right] \right\}, \quad t = 1, \ldots, T, \tag{8.6}$$

where $E_{a_t \sim \pi}[\cdot]$ denotes the expectation over the outcomes corresponding to all possible actions prescribed by π. Without loss of generality, we set the final value $V^{\pi}(s_{T+1}) = 0, \forall s_{T+1}$ under any strategy π.

Another important variable is Q-value $Q^{\pi}(s_t, a_t)$, which is defined as the expected return achieved by a driver who starts at state s_t, takes action a_t, and then follows strategy π thereafter. The corresponding Bellman equation is written as

$$Q^{\pi}(s_t, a_t) = \begin{cases} \sum_{s_{t+1}} P(s_{t+1}|s_t, a_t) \left[R(s_t, a_t, s_{t+1}) + \gamma V^{\pi}(s_{t+1}) \right], & t = 1, \ldots, T \\ 0, & t = T+1 \end{cases}. \tag{8.7}$$

The objective of a driver is thus to determine the optimal search strategy that maximizes the value function from its initial state s_0. Given the probability distribution of initial state ζ_0, the optimal strategy is solved as

$$\pi^* = \arg\max_{\pi} E_{s_0 \sim \zeta_0}[V^{\pi}(s_0)]. \tag{8.8}$$

It is worth emphasizing that Eq. (8.8) assumes the vacant vehicle distribution $\mathbf{y}$ is given by other drivers' search strategies.

8.2.1.1. Solution algorithm. Recall that the transitions from occupied states are deterministic. Hence, the Q-value for any state, either occupied or vacant, can be represented by the Q-values of all vacant states. For notation simplicity, we denote the Q-values of vacant state $s_t = (\text{vac}, i)$ and

action $a_t = j$ as $Q(t, i, j)$. Accordingly, the Q-value function is simplified as

$$Q(t,i,j) = \begin{cases} m_j^t \sum_k (pjk + \gamma^{\tau_{jk}+1} V(t+\tau_{jk}+1, i)) + \gamma(1-m_j^t)V(t+1,j), & t = 1, \ldots, T \\ 0, & t = T+1 \end{cases}. \tag{8.9}$$

In Eq. 8.9, $V(t, i)$ is the value function of vacant state $s_t = (\text{vac}, i)$ and it is computed as

$$V(t,i) = \sum_{j \in \mathcal{A}_i} \pi_{ij}^t Q(t,i,j). \tag{8.10}$$

Given the vacant vehicle distribution $\mathbf{y}$, the optimal routing strategy π^* is solved through backward induction, as summarized in Algorithm 8.1.

Algorithm 8.1 Backward induction

1: **Inputs:** Demand rate $\mathbf{q} = \{q_i^t, \forall t, i\}$; trip fare $\mathbf{p} = \{p_{ij}, \forall i, j\}$; vacant vehicle distribution $\mathbf{y}$; time horizon T; discount factor γ.
2: **Outputs:** Q-values of vacant states $Q(t, i, j)$ and the optimal routing strategy $\tilde{\pi}$.
3: Initialize $Q(t, i, j) = 0,\ \forall i, j$.
4: **for** $t = T, \ldots, 1$ **do**
5: Compute meeting probability $m_t^j,\ \forall j \in \mathcal{N}$ according to the ride-hail mode and market condition (i.e., passenger demand $\mathbf{q}$ and vehicle supply $\mathbf{y}$).
6: Compute state transition probabilities $P(s'_{t+1}|s_t, a_t),\ \forall s'_{t+1}, s_t, a_t$ by Eq. (8.4).
7: Compute reward $R(s_t, a_t, s'_{t+1}),\ \forall s'_{t+1}, s_t, a_t$ by Eq. (8.5).
8: Update Q-values $Q(t, i, j),\ \forall i, j$ by Eq. (8.9).
9: For each zone i, determine $\mathcal{Z}_i$ such that $\forall j \in \mathcal{Z}_i,\ j = \arg\max_{j'} Q(t, i, j')$.
10: Set the optimal search strategy for zone i as $\tilde{\pi}_{ij}^t = 1/|\mathcal{Z}_i|$ if $j \in \mathcal{Z}_i$, otherwise 0.
11: **end for**

8.3. RIVER

8.3.1. Collective routing

Individually, each driver routes through the spatial ride-hail market to maximize their own expected return. Collectively, their routing decisions produce the relocation flow $\mathbf{x} = \{x_{ij}^t, \forall t, i, j\}$ and idle drivers distribution $\mathbf{y} = \{y_j^t, \forall t, j\}$. In turn, the spatiotemporal distribution of idle drivers affects state transition probabilities and reward through the meeting probability $\mathbf{m} = \{m_j^t, \forall t, j\}$ (see Eqs. (8.4a), (8.4b) and (8.5)). As each driver's expected return depends on not only their own

decision, but also those of others, the evolution of the system may be described as an *MDP congestion game* [Calderone and Sastry, 2017, Calderone and Shankar, 2017]. Formally, the collective routing problem concerned in this study may be defined as follows.

Definition 1 (RIVER)**.** *The RIde-hail VEhicle Routing (RIVER) problem seeks to determine the spatiotemporal distribution of vacant vehicles and vacant vehicular flow between any pair of zones in the ride-hail market, assuming each driver $l \in \mathcal{M}$ aims to maximize their expected return following an MDP defined by $(\mathcal{S}, \mathcal{A}, P, R, \gamma)$.*

Since drivers are homogeneous per assumption, they must follow the same strategy when the system arrives at a steady state, or equilibrium. Also, since the number of drivers in a ride-hail market is typically very large, it is safe to assume the impact of a single driver's decision on the overall vehicle distribution is negligible. Thus, RIVER essentially describes a mean field game [Lasry and Lions, 2007]. To characterize its equilibrium, let us first define $Q(s_t, a_t, \pi)$ as the expected return achieved by a driver who starts at state s_t, takes action a_t, and then follows strategy π thereafter, when *all other drivers also follow* π. Note that $Q(s_t, a_t, \pi)$ differs from $Q^\pi(s_t, a_t)$ in Eq. (8.7) in that the former is defined on a strategy shared by everyone, whereas the latter concerns the strategy of an individual driver given other drivers' strategies.

We now formally define the stable state of the RIVER problem as a Wardrop equilibrium [Wardrop, 1952, Beckmann et al., 1956].

Definition 2 (Wardrop equilibrium for RIVER)**.** *An aggregate strategy π^* corresponds to be a Wardrop equilibrium of the RIVER problem if for any state $s_t = (\text{vac}, i)$ and action $a_t = j$ such that $\pi_{ij}^{t*} > 0$,*

$$Q(s_t, a_t, \pi^*) \geq Q(s_t, a_t', \pi^*),\ \forall a_t' \in \mathcal{A}. \tag{8.11}$$

At an equilibrium, if any driver chooses action $a_t = j$ at s_t, no other feasible actions at that state would yield a higher expected return. In other words, no driver could improve their expected return by unilaterally switching to a different strategy.

In RIVER, each driver solves the MDP problem given a vacant vehicle distribution $\mathbf{y}$, which is, in turn, determined by the aggregate strategy of all vehicles. We can express these dependencies as follows:

$$\begin{cases} \pi = F_1(\mathbf{y}) \\ \mathbf{y} = F_2(\pi) \end{cases}, \tag{8.12}$$

where the first equation restates the MDP problem (Eq. (8.8)) given $\mathbf{y}$ and the second describes the system dynamics given π. Note that the vacant vehicle distribution $\mathbf{y}$ is aggregated from the relocation vacant vehicular flows $\mathbf{x}$ (see Eq. (8.1)), each relocation flow x_{ij}^t is dictated by the search strategy (i.e., $\pi_{ij}^t x_i^t$), and the total relocation flow x_i^t is determined by vacant vehicle distribution at an earlier time (see Eqs. (8.2) and (8.3)). Thus, by induction, we can represent $\mathbf{y}$ as a function of π given the fleet size M and initial vehicle distribution ζ_0. This enables us to reduce Eq. (8.12) to a fixed point system $\pi^* = F(\pi^*)$ with $F(\cdot) = F_1(F_2(\cdot))$. The following proposition connects such a fixed point to the Wardrop equilibrium defined above.

Proposition 4. *Each solution to the fixed point system $\pi^* = F(\pi^*)$ is a Wardrop equilibrium, and vice versa.*

Proof. "⇒": Let $(\pi^*, \mathbf{y}^*)$ be a solution to Eq. (8.12). The Bellman optimality equation reads

$$V^{\pi^*}(s_t) = \max_{a_t} Q^{\pi^*}(s_t, a_t). \tag{8.13}$$

By definition, $V^{\pi^*}(s_t) = E_{a_t \sim \pi^*} Q^{\pi^*}(s_t, a_t)$. Hence, it must hold that $\pi_{ij}^{t*} > 0$ if and only if $Q^{\pi^*}(s_t, a_t) = \max_{a_t} Q^{\pi^*}(s_t, a_t)$. In other words, for any state $s_t = (\text{vac}, i)$ and action $a_t = j$ such that $\pi_{ij}^{t*} > 0$, $Q^{\pi^*}(s_t, a_t) \geq Q^{\pi^*}(s_t, a_t'),\ \forall a_t' \in \mathcal{A}$. Up to this point, we assume the vehicle distribution $\mathbf{y}^*$ is given at the fixed point. According to Eq. (8.12), $\mathbf{y}^*$ corresponds to π^*, i.e., the

same strategy adopted by all drivers. Therefore, a driver's Q-value can be rewritten as $Q(s_t, a_t, \pi^*)$, leading to the Wardrop equilibrium as stated in Definition 2.

"$\Leftarrow$": Since all drivers adopt strategy π^*, the equilibrium property also holds for any individual driver. That is, $\pi^* = \arg\max_\pi E_{a_t\sim\pi} Q^\pi(s_t, a_t) = \arg\max_\pi V^\pi(s_t)$ for any state $s_t = (\text{vac}, i)$. This leads to $\pi^* = \arg\max_\pi E_{s_0\sim\zeta_0}[V^\pi(s_0)]$. Recall from Eq. (8.8) that each driver optimizes their strategy based on the vacant vehicle distribution. In this case, the distribution $\mathbf{y}^*$ is determined by the shared strategy π^* at the equilibrium. Thus, π^* is a solution of a fixed point problem. □

Next, we prove the existence of the Wardrop equilibrium by invoking the fixed point theorem.

Proposition 5. *There exists at least one Wardrop equilibrium for the RIVER problem.*

Proof. Given Proposition 4, it suffices to prove the existence of solution to the fixed point system. Let Ω denote the feasible region of π. Given a strategy π, a vehicle distribution is uniquely determined. Although the MDP problem may not have unique solution, all solutions must fall in Ω. Hence, F is a set-valued function that maps from Ω to 2^Ω.[2]

As per the Kakutani's fixed point theorem, a set-valued function $\phi : \mathcal{Z} \to 2^\mathcal{Z}$ has a fixed point if the following conditions are satisfied: (i) $\mathcal{Z}$ is non-empty, compact and convex, (ii) $\phi(z)$ is non-empty and convex for all $z \in \mathcal{Z}$, and (iii) ϕ has a closed graph.

The first condition holds as per the definition of π, i.e., $\pi_{ij}^t \geq 0,\ \forall i, j, t$ and $\sum_j \pi_{ij}^t = 1,\ \forall i, t$. Recall that F maps from one strategy π to a set of optimal strategies, denoted by $\mathcal{S}(\pi)$, based on the induced vehicle distribution $F_2(\pi)$. Since the set of feasible strategies (i.e., Ω) is closed, $\mathcal{S}(\pi)$ is non-empty. Consider two strategies $\pi_1, \pi_2 \in \mathcal{S}(\pi)$. Then, we have $V^{\pi_1}(s_t) = V^{\pi_2}(s_t) = V^*(s_t),\ \forall s_t$, where V^* refers to the optimal value for state s_t. Let $\pi_3 = \lambda\pi_1 + (1-\lambda)\pi_2$ for any $\alpha \in (0, 1)$. For any $s_{T-1} = (\text{vac}, i)$, it yields

$$V^{\pi_3}(s_T) = E_{a_T\sim\pi_3}\left\{\sum_{s_{T+1}} P(s_{T+1}|s_T, a_T)\left[R(s_T, a_T, s_{T+1}) + \gamma V^\pi(s_{T+1})\right]\right\} \tag{8.14}$$

[2] 2^Ω denotes the set of all subsets of Ω.

$$
\begin{aligned}
&= \sum_j \left[(\pi_3)_{ij}^T \sum_{s_{T+1}} P(s_{T+1}|s_T, j) R(s_T, j, s_{T+1}) \right] \\
&= \sum_j \left\{ \left[\lambda (\pi_1)_{ij}^T + (1-\lambda)(\pi_2)_{ij}^T \right] \sum_{s_{T+1}} P(s_{T+1}|s_T, j) R(s_T, j, s_{T+1}) \right\} \\
&= \lambda V^{\pi_1}(s_T) + (1-\lambda) V^{\pi_2}(s_T) = V^*(s_T).
\end{aligned}
$$

The second equality above holds because $V^{\pi}(s_T) = 0$. By induction, we can show $V^{\pi_3}(s_0) = V^*(s_0)$, $\forall s_0$. Therefore, π_3 is also an optimal strategy. In other words, $\mathcal{S}(\pi)$ contains any convex combinations of its elements and thus it is convex. Accordingly, the second condition is satisfied.

Since Ω is a compact set in a Hausdorff space, the third condition holds if F is upper hemicontinuous and $F(\pi)$ is a closed subset of Ω for any π [Theorem 17.11 in Aliprantis and Border, 2006]. The latter condition is easily satisfied by the definition of F. Since $F(\cdot) = F_1(F_2(\cdot))$ and F_2 is single-valued and continuous with π (see Eqs. (8.1)–(8.3)), we only need to prove F_1 is upper hemicontinuous. To this end, we first state the Berge's maximum theorem [Theorem 17.31 in Aliprantis and Border, 2006] as follows. Let $\mathcal{Z}$ and Θ be metric spaces, $C : \Theta \rightrightarrows \mathcal{Z}$ be a compact-valued and non-empty correspondence, and $f : \mathcal{Z} \times \Theta \to \mathbb{R}$ be a continuous function. The theorem states that, if C is continuous, then $f^*(\theta) = \max_{z \in C(\theta)} f(z, \theta)$ is continuous and $z^*(\theta) = \arg\max_{z \in C(\theta)} f(z, \theta)$ is upper hemicontinuous, compact-valued and non-empty.

To apply the Berge's maximum theorem in our setting, we note that $\mathcal{Z}$ and Θ correspond to the feasible sets of strategy π (i.e., Ω) and vacant vehicle distribution $\mathbf{y}$, respectively, and C maps $\mathbf{y}$ to a set of feasible strategies. Since the feasible strategies do not depend on $\mathbf{y}$[3], $C(\mathbf{y}) = \Omega$ for all feasible $\mathbf{y}$. It is easy to verify C defined in this way is continuous, compact and non-empty, as required by theorem. The function f is the objective function of the MDP problem (i.e., $E_{s_0 \sim \zeta_0}[V^{\pi}(s_0)]$), which takes $\mathbf{y}$ as inputs. Let f be the objective function of the MDP problem (i.e., $E_{s_0 \sim \zeta_0}[V^{\pi}(s_0)]$), which takes $\mathbf{y}$ and $C(\mathbf{y})$ as inputs. Then, $F_1(\cdot)$ corresponds to $z^*(\cdot)$ above, i.e., the optimal solution to

[3]This implicitly requires one can always relocate to a neighbor zone regardless how many vacant vehicles are there, which is not a restrictive assumption.

the MDP problem given $\mathbf{y}$. Therefore, F_1 is upper hemicontinuous, compact-valued and non-empty as per the Berge's maximum theorem. It thus concludes that F is also upper hemicontinuous, compact-valued and non-empty.

Having confirmed all three conditions required by the Kakutani's fixed point theorem are satisfied, the proof is completed. □

8.3.2. Extension to cooperative routing

So far, we have assumed drivers compete with each other to maximize their own revenue. However, a ride-hail operator may be interested in the total revenue of the entire fleet, which may not be maximized by the selfish routing decisions of individual drivers. This gap between the best possible collective outcome and the outcome of an equilibrium achieved by selfish agents is well known, and often referred to as the price of anarchy (PoA) [Roughgarden and Tardos, 2002]. A common remedy to this conflict is to amend the personal reward such that self-interest individuals are incentivized to reach collective good, or system optimum (SO). This approach is often referred to as *decentralization*, in that it achieves the system-level objectives without directly controlling all agents. Agents who adopt these amended rewards are called cooperative agents, because they are no longer driven by self-interest.

Let the total system return be defined as $\Phi_c(\pi) = \sum_t \gamma^t \left(\sum_j y_j^t m_j^t \bar{p}_j \right)$, where $\bar{p}_j = \sum_k \alpha_{jk} p_{jk}$ is the expected fare of a trip originated in zone j. The following result asserts that finding an aggregate strategy that maximizes $\Phi_c(\pi)$ is equivalent to solving the RIVER problem with a cooperative reward function.

Proposition 6. *Under Assumption 7, a cooperative routing strategy that maximizes the total system return $\Phi_c(\pi)$ corresponds to a Wardrop equilibrium for the RIVER problem when each driver*

maximizes the following cooperative reward function

$$R_c(s_t, a_t, s_{t+1}) = \begin{cases} p_{jk}\left(1 + \frac{y_j^t}{m_j^t}\,\nabla_y m_j^t\Big|_{y=y_j^t}\right) & , \; s_t = (\text{vac}, i), a_t = j, s_{t+1} = (\text{occ}, j, k, 0) \\ 0 & , \; \textit{otherwise} \end{cases} . \quad (8.15)$$

Proof. The cooperative routing strategy is obtained by solving the following problem:

$$\max_{\pi} \quad \Phi_c(\pi) \tag{8.16a}$$

$$s.t. \quad \pi \in \Omega. \tag{8.16b}$$

Problem (8.16) can be reformulated as

$$\max_{\pi, \mathbf{y}} \quad \Phi_c(\mathbf{y}) \tag{8.17a}$$

$$s.t. \quad \mathbf{y} = F_2(\pi), \tag{8.17b}$$

$$\pi \in \Omega. \tag{8.17c}$$

The dual of (8.17) reads

$$\min_{\lambda} \quad \mathcal{L}(\lambda) = \max_{\pi \in \Omega, \mathbf{y}} \Phi_c(\mathbf{y}) - \lambda^T(\mathbf{y} - F_2(\pi)), \tag{8.18a}$$

where λ is the multiplier associated with Constraint (8.17b).

For a given λ, Problem (8.18) can be decomposed into the following two subproblems

$$\pi = \arg\max_{\pi' \in \Omega} \lambda^T F_2(\pi'), \tag{8.19}$$

$$\mathbf{y} = \arg\max_{\mathbf{y}'} \Phi_c(\mathbf{y}') - \lambda^T \mathbf{y}'. \tag{8.20}$$

Proposition 5 proves a Wardrop equilibrium of the RIVER problem exists with the original reward function Eq. (8.5). Following the same reasoning, we can show the existence also holds with the cooperative reward function Eq. (8.15). Let $(\check{\pi}, \check{\mathbf{y}})$ be such a Wardrop equilibrium. In what

follows, we will show a solution to the dual problem (8.18) can be constructed such that $(\check{\pi}, \check{\mathbf{y}})$ satisfy the conditions (8.19) and (8.20).

Under Assumption 7, Φ_c is a concave function of $\mathbf{y}$, because

$$\frac{\partial^2 \Phi_c(\mathbf{y})}{\partial (y_j^t)^2} = \gamma^t \frac{\partial^2 (m_j^t y_j^t)}{\partial (y_j^t)^2} \bar{p}_j \leq 0. \tag{8.21}$$

Therefore, the optimality condition of Subproblem (8.19) dictates that

$$\frac{\partial \Phi_c(\mathbf{y})}{\partial \mathbf{y}} = \lambda. \tag{8.22}$$

Let

$$\check{\lambda}_j^t = \frac{\partial}{\partial y_j^t} \Phi_c(\check{\mathbf{y}}) = \gamma^t \check{m}_j^t \left(1 + \frac{\check{y}_j^t}{\check{m}_j^t} \nabla_y m_j^t \big|_{y=\check{y}_j^t}\right) \bar{p}_j, \tag{8.23}$$

where $\check{m}_j^t$ denotes the meeting probability associated with a vehicle distribution $\check{y}_j^t$. Then, $\check{\mathbf{y}}$ satisfies Eq. (8.20).

To show Eq. (8.19) also holds for $\check{\pi}$ given $\check{\lambda}$, we define u_t as the location of a vacant vehicle after its relocation in time period t. Clearly, u_t is a random variable and depends on the driver's strategy. With all drivers taking strategy π, we have $Pr(u_t = j|\pi) = y_j^t/M$ (recall M is the fleet size).

Let $r^\pi(s_t)$ be the stage-reward at state s_t under strategy π (i.e., the expected reward obtained in time period t by following π). Since the reward is always zero except for transitions from vacant to occupied states, it follows that

$$\begin{aligned} E_{s_t}[r^\pi(s_t)] &= E_{s_t}\left\{ E_{a_t \sim \pi}\left[\sum_{s_{t+1}} P(s_{t+1}|s_t, a_t) R_c(s_t, a_t, s_{t+1}) \right] \right\} \\ &= \sum_i Pr(s_t = (\text{vac}, i)) \sum_j \pi_{ij}^t \sum_k m_j^t \alpha_{jk} R_c(s_t = (\text{vac}, i), a_t = j, s_{t+1} = (\text{occ}, j, k, 0)) \\ &= \sum_j Pr(u_t = j|\pi) m_j^t \sum_k \alpha_{jk} p_{jk} \left(1 + \frac{y_j^t}{m_j^t} \nabla_y m_j^t \big|_{y=y_j^t}\right), \end{aligned} \tag{8.24}$$

$$= \sum_j Pr(u_t = j|\pi) m_j^t \left(1 + \frac{y_j^t}{m_j^t} \nabla_y m_j^t \big|_{y=y_j^t}\right) \bar{p}_j.$$

Here, it is safe to remove the index i because the reward function is the same regardless the relocation starting zone.

Given $\check{\lambda}$ set in Eq. (8.23), the equivalence between the solution to Subproblem (8.19) and $\check{\pi}$ is established as follows:

$$\begin{aligned}
\pi &= \arg\max_{\pi'} \check{\lambda}^T F_2(\pi) \qquad (8.25)\\
&= \arg\max_{\pi'} \sum_t \sum_j \check{\lambda}_j \frac{y_j^t}{M}\\
&= \arg\max_{\pi'} \sum_t \sum_j \gamma^t Pr(u_t = j|\pi') \check{m}_j^t \left(1 + \frac{\check{y}_j^t}{\check{m}_j^t} \nabla_y m_j^t \big|_{y=\check{y}_j^t}\right) \bar{p}_j\\
&= \arg\max_{\pi'} \sum_t \gamma^t E_{s_t}[r^{\pi'}(s_t)] = \check{\pi}.
\end{aligned}$$

The last equality states that the equilibrium strategy $\check{\pi}$ maximizes the sum of discounted expected stage-reward, which can be established by expanding Eq. (8.8) at each stage.

Therefore, we established that $(\check{\pi}, \check{\mathbf{y}}, \check{\lambda})$ is a solution to the dual problem (8.18). As per weak duality, we have $\mathcal{L}(\check{\lambda}) \geq \Phi_c(\pi)$ for all feasible strategy π. Because the equilibrium condition dictates $\check{\mathbf{y}} = F_2(\check{\pi})$, we have $\mathcal{L}(\check{\lambda}) = \Phi_c(\check{\mathbf{y}}) \geq \max_{\pi\in\Omega} \Phi_c(\pi)$. Because by definition $\Phi_c(\check{\mathbf{y}})$ cannot be larger than $\max_{\pi\in\Omega} \Phi_c(\pi)$, $\check{\pi}$ must be the cooperative routing strategy that maximizes $\Phi_c(\check{\mathbf{y}})$. This completes the proof. □

8.3.3. Solution algorithm

Since the RIVER problem is highly nonlinear and nonconvex, devising an exact convergent algorithm is difficult. In this study, we propose to solve it using a heuristic inspired by Proposition 4, i.e., creating a fixed-point iteration of the search strategy π. To stabilize the process, π is updated using the method of successive average (MSA). The origin of MSA can be tracked back to the work

of Robbins and Monro [1951] and Blum [1954], and it has been widely used to solve congestion games in general Miller et al. [2020] and various traffic assignment problems Sheffi and Powell [1982], Powell et al. [1995] in particular.

Algorithm 8.2 MSA for the RIVER problem

1: **Inputs:** Demand profile $\mathbf{q} = \{q_i^t, \forall t, i\}$ and $\boldsymbol{\alpha} = \{\alpha_{ij}, \forall i, j\}$; travel time $\boldsymbol{\tau} = \{\tau_{ij}, \forall i, j\}$ and trip fare $\mathbf{p} = \{p_{ij}, \forall i, j\}$; fleet size M; initial vehicle distribution ζ_0; number of time periods T; discount factor γ; maximum iteration I; gap tolerance ε.
2: **Outputs:** Equilibrium relocation flow pattern $\mathbf{x}^*$.
3: **Initialization**: Randomly generate an initial routing strategy $\pi^{(0)}$; set $g = \infty$ and $n = 0$.
4: **while** $g > \varepsilon$ and $n < I$ **do**
5: Update $\mathbf{x}$ and $\mathbf{y}$ via ***Forward Induction***.
6: Update Q-values and compute the optimal strategy $\tilde{\pi}$ via ***Backward Induction***.
7: Update gap $g = \sum_{(t,i,j)\in\mathcal{X}} |Q(t,i,j) - \max_j Q(t,i,j)| / \sum_{t,i,j} |Q(t,i,j)|$, where $\mathcal{X} = \{(t,i,j) : x_{i,j}^t > 0\}$.
8: Update the search strategy $\pi^{(n+1)} = (1-\eta)\pi^{(n)} + \eta\tilde{\pi}$ with step size $\eta = 1/(n+1)$.
9: Set $n = n + 1$.
10: **end while**
11: Set $\mathbf{x}^* = \mathbf{x}$.

Algorithm 8.2 describes the main steps of the solution procedure. Starting with an initial routing strategy $\pi^{(0)}$, we first load vehicular flows through forward induction (see Algorithm 8.3). Then, we find the optimal strategy based on the loading result by solving the MDP through backward induction (see Algorithm 8.1). Finally, we compute the gap and update the search strategy in an MSA manner. The iterations terminate when the gap is below a predefined threshold or a maximum iteration number is reached. Here, the gap is defined as the relative deviation from the equilibrium, that is, the total difference between the Q-value of any vacant state-action tuple with positive relocation flow and the maximum Q-value at the same state over the total absolute Q-values (see line 7 in Algorithm 8.2).

8.4. Meeting probability

In this section, we specify the meeting probability m_i^t for two ride-hail modes: street-hail and e-hail. Commonly used by conventional taxis in densely populated urban areas, street-hail allows passengers to hail vacant vehicles off street as they cruise by. In contrast, passengers and drivers

Algorithm 8.3 Forward induction

Inputs: Demand profile $\mathbf{q} = \{q_i^t, \forall t, i\}$ and $\boldsymbol{\alpha} = \{\alpha_{ij}, \forall i, j\}$; travel time $\boldsymbol{\tau} = \{\tau_{ij}, \forall i, j\}$; search strategy π; fleet size M; initial vehicle distribution ζ_0; number of time periods T.
Outputs: Vacant vehicular flows $\mathbf{x}$ and vacant vehicle distribution $\mathbf{y}$.
Initialization: Set $d_i^0 = MPr(s_0 = (\text{vac}, i))$, $\forall i$, where $Pr(s_0 = (\text{vac}, i))$ is dictated by ζ_0.
for $t = 1, \ldots, T$ **do**
    Set the total vacant vehicular flow originated from zone i, x_i^t, using Eq. (8.2).
    Set the idle flow from zone i to zone j $x_{ij}^t = \pi_{ij}^t x_i^t$.
    Update vacant vehicle distribution y_i^t in each zone i by Eq. (8.1).
    Update future vehicle arrivals in each zone i by adding new arriving vehicles $m_i^t y_i^t \alpha_{ij}$ to $d_j^{t+\tau_{ij}+1}$.
end for

in e-hail mode always interact on-line through an intermediate (e.g., TNC) before they meet. To simplify notation, we remove the subscription i and superscription t from the symbols in this section.

8.4.1. Spatial matching model

Consider a vacant vehicle searching for passengers. Assume waiting passengers are uniformly distributed in space with density Π and the vehicle can search an area a in Δ. The number of passengers encircled by the vehicle's search area, denoted by N_a, may be described using a spatial Poisson process with parameter Π Larson and Odoni [1981]. The meeting probability is thus equivalent to the probability that $N_a > 0$, which is given by

$$m = 1 - P(N_a = 0) = 1 - \exp(-\Pi a). \tag{8.26}$$

Let us first derive the passenger density for the two ride-hail modes. Note that we may estimate the number of waiting passengers per unit area as qw per the Little's formula Little [1961], where w is the average passenger wait time. The physical matching model presented in Section 3.1 implies that $w \propto 1/\Lambda$ for street-hail and $w \propto \sqrt{\Pi/\Lambda}$ for e-hail, where Λ is the vacant vehicle density. Following the above analysis, we replace Π with qw and rewrite Λ as y/A, where A denotes the zone area. With some algebra, we can derive Π in Eq. (8.26) as

$$\text{Street-hail:} \quad \Pi = k_s' \frac{qA}{y}, \tag{8.27}$$

$$\text{E-hail:} \quad \Pi = k'_e \frac{q^2 A}{y}, \tag{8.28}$$

where k'_s and k'_e are parameters associated with zone properties and matching mechanism. They may be calibrated from empirical or simulated data, as illustrated later in this section.

In street-hail, the search area a equals the cruising area, which may be estimated as $v\Delta/\rho$, where ρ is the road density. For e-hail, the search area can be approximated by A/y, assuming vacant vehicles are uniformly distributed in space and passengers are assigned to the closest vehicle. Accordingly, the meeting probabilities for street-hail and e-hail are given by

$$\begin{aligned} \text{Street-hail:} \quad m_s &= 1 - \exp\left[-\left(k'_s \frac{qA}{y}\right)\left(\frac{v\Delta}{\rho}\right)\right] = 1 - \exp\left(-k_s \frac{q\Delta A}{y}\right) \\ &= 1 - \exp\left(-k_s \varphi\right), \end{aligned} \tag{8.29}$$

$$\begin{aligned} \text{E-hail:} \quad m_e &= 1 - \exp\left[-\left(k'_e \frac{q^2 A}{y}\right)\left(\frac{A}{y}\right)\right] = 1 - \exp\left[-k_e \left(\frac{q\Delta A}{y}\right)^2\right] \\ &= 1 - \exp(-k_e \varphi^2), \end{aligned} \tag{8.30}$$

where $\varphi = q\Delta A/y$ is the demand-supply ratio, $k_s = k'_s v/\rho$ and $k_e = k'_e/\Delta^2$.

The above derivation of the meeting probability function for e-hail implicitly assumes the passenger demand and the vehicle supply are comparable in the market so that the passenger wait time depends on Π. However, we note the result no longer holds when the market is severely oversupplied. In such a case, there is most one passenger waiting in the market at any moment and m_e needs to be derived from a different physical matching process.

Suppose N_p passengers arrive in a market with N_v vacant vehicles in sequence and $N_p \ll N_v$. Then, the probability of each driver picking up any of the N_p passengers is given by

$$\begin{aligned} Pr(\text{pickup a passenger}) &= \frac{1}{N_v} + \frac{N_v - 1}{N_v}\frac{1}{N_v - 1} + \cdots + \left(\Pi_{n=1}^{N_p} \frac{N_v - n}{N_v - n + 1}\right)\frac{1}{N_v - N_p} \\ &= \frac{N_p}{N_v}. \end{aligned} \tag{8.31}$$

The nth term in Eq. (8.31) is the probability that the vehicle picks up the nth arriving passenger giving that they fails to pick up the previous $n-1$ passengers. For example, if the vehicle successfully picks up the second arriving passenger, then the probability is computed as $\frac{N_v-1}{N_v}$ (the probability that they did not pick up the first passenger) times $\frac{1}{N_v-1}$ (the probability that the second passenger is matched to them, assuming the vehicle matched to the first passenger is no longer available and all remaining vehicles have the equal opportunity to be matched with the new arriving passenger). There are N_p terms in Eq. (8.31) and each of them can be reduced to $1/N_v$. Thus, we specify the e-hail meeting probability in the oversupply case as

$$m_e = k_{\tilde{e}} \frac{q\Delta A}{y}, \tag{8.32}$$

where $k_{\tilde{e}}$ is the parameter associated with this case and may be calibrated from data as well.

A remaining question is when Eq. (8.32), instead of Eq. (8.30), should be used to estimate the meeting probability. We hypothesize that there should exist a critical demand-supply ratio φ^*, below which the market enters the oversupply regime. As shown later, the value of φ^* may be estimated from data. Thus, m_e may be written as the following piece-wise function

$$m_e = \begin{cases} 1 - \exp(-k_e \varphi^2) & \varphi \geq \varphi^* \\ k_{\tilde{e}} \varphi & \varphi < \varphi^* \end{cases}. \tag{8.33}$$

8.4.2. Model calibration

To validate and calibrate the spatial matching model, we simulate street-hail and e-hail services on grid networks with various zone sizes, road densities, travel speeds and a wide range of supply-demand profiles. Two sets of simulations are conducted to calibrate the meeting probability function in both normal and oversupply regimes. The control parameters and the range of their values are reported in Table 8.1.

The simulation is developed in MATSim, an agent-based transportation simulation framework Horni et al. [2016], and will be introduced in more detail in Section 9. At the beginning

of each simulation, a fleet of empty vehicles are randomly placed in the network all at once, while passengers arrive following a Poisson process. In all runs, the simulation horizon is set to 0.25 hr, which is considered to be a reasonable length for Δ in the RIVER problem.

Table 8.1. Range of values for control parameters.

control parameter	regular	oversupply
number of block	[4, 40]	[4, 40]
block length (m)	[100, 500]	[100, 500]
speed factor	[0.5, 1.5]	1.0
vehicle supply (veh)	[50, 250]	[100, 300]
passenger demand (pax/hr)	[250, 500]	[50, 150]

The matching process presented in the previous sections suggests that the parameters k_s and k_e, as well as $k_{\tilde{e}}$, mainly depend on traffic speed v and road density ρ in the zone. Thus, we propose to model these parameters as a function of the two variables, i.e.,

$$k_l = e^{\beta_{l0}} \rho^{\beta_{l1}} v^{\beta_{l2}},\ l = \{s, e, \tilde{e}\}. \tag{8.34}$$

Replacing k_l in Eqs. (8.29) and (8.30) with the right-hand-side of Eq. (8.34) leads to the following linear regression model that can be fitted with data generated from simulations.

$$\log(-\log(1 - m_l)) = \beta_{l0} + \beta_{l1} \log \rho + \beta_{l2} \log v + \beta_{l3} \log \varphi, \tag{8.35}$$

where β_{l3} is the exponent on the term $q\Delta A/y$ in Eqs. (8.29) and (8.30). Here, we turn the exponent into a parameter to test if this prediction of the matching theory matches simulation results.

For e-hail under oversupply, we instead fit the following linear model:

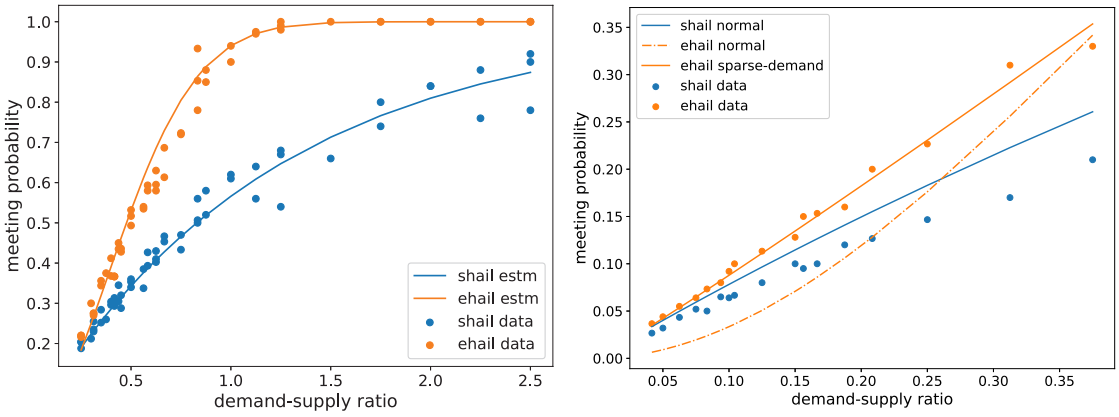
$$\log(m_e) = \beta_{\tilde{e}0} + \beta_{\tilde{e}1} \log \rho + \beta_{\tilde{e}2} \log v + \beta_{\tilde{e}3} \log \varphi. \tag{8.36}$$

Table 8.2. Main regression results.

model	street-hail ($l = s$)	e-hail normal ($l = e$))	e-hail oversupply ($l = \tilde{e}$)
β_{l0}	0.1959**	-0.0751***	-0.1148**
	(0.028)	(0.026)	(0.050)
β_{l1}	-0.3732***	0.0767***	0.0284***
	(0.004)	(0.004)	(0.007)
β_{l2}	0.2949***	0.2789***	0.0043
	(0.008)	(0.008)	(0.014)
β_{l3}	0.9919***	1.8977***	1.0538***
	(0.005)	(0.005)	(0.008)
# obs.	24750	24750	4121
df. model	3	3	3
R^2	0.761	0.919	0.841
adj. R^2	0.761	0.919	0.841
Standard errors in parentheses. Significance levels: *** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$.			

The regression results are reported in Table 8.2. First and foremost, the estimated β_{l3}, 0.99 for s-hail, 1.90 for e-hail in normal regime and 1.05 for e-hail in oversupply regime, agree well with the model's prediction, that is, 1 for street-hail and e-hail with oversupply, and 2 for e-hail in normal regime. The results also indicate traffic speed has a similar impact on the meeting probability in both modes (as measured by β_{l2}), while it hardly affects the meeting probability for e-hail with oversupply. As expected, road density ρ negatively affects the meeting probability in street-hail because a more dense road network would reduce the search area within a given time (see Eq. (8.29)). In contrast, ρ is positively correlated with the e-hail meeting probability, though the effect seems rather minor ($\beta_{e2} = 0.08$ and $\beta_{\tilde{e}2} = 0.03$). This is also expected because a larger ρ usually leads to a more "direct" pickup route, which improves the matching probability. Overall, the fitness of three models with the simulated data is considered satisfactory, though the e-hail model fits better with a R^2 greater than 0.9 for the normal regime.

Figure 8.1 plots the meeting probabilities of street-hail and e-hail against demand-supply ratio. In particular, Figure 8.1(a) covers a larger range of demand-supply ratio while (b) zooms into a window between 0 to 0.4. We can see the basic models fit well with the data for both modes. Clearly, the meeting probability in e-hail increases much faster with the demand-supply ratio than



(a) normal regime

(b) oversupply

Figure 8.1. Meeting probability vs. demand-supply ratio. Each data point represents the median of meeting probabilities obtained from multiple simulation runs at a given demand-supply ratio, while the solid curves represent model predictions at each demand-supply ratio with all other variables set to be the median values.

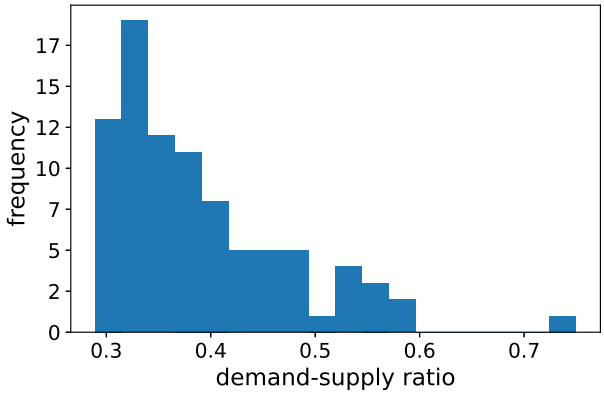


Figure 8.2. Histogram of intersection points of demand-supply ratio between two e-hail models.

in street-hail. It rises over 90% when the ratio approaches 1.0, and quickly converge to 100% after the threshold is crossed. In contrast, the meeting probability in street-hail barely reaches 90% when the demand-supply ratio is as high as 2.5. With an oversupply of vehicles, the street-hail model still produce reasonable predictions on the meeting probability whereas the basic e-hail model does cause a significant error and tend to underestimate the meeting probability. In this case, the adjusted model provides shows a much better performance.

To estimate φ^* in Eq. (8.33), we discretize the values of ρ and v into grids. Then, for each combination of ρ and v, we may construct two curves of m_e with respect to φ, one in normal regime and the other in oversupply regime. The intersection is thus one realization of φ^*. Figure 8.2 plots the histogram of φ^*. It can be seen that its value mostly falls between 0.3 and 0.4. Hence, in the following section, we set $\varphi^* = 0.37$, the median value of all realizations.

8.5. Numerical experiments

In this section, we test the RIVER problem under both street-hail and e-hail modes and for two behavioral assumptions: non-cooperative (NC) routing and cooperative (CO) routing. Two sets of experiments are conducted: the first one is based on a hypothetical network that consists of four zones (Section 8.5.1), and the second is a nine-zone model of Chicago constructed using real data (Section 8.5.2). In all experiments, we set the time interval $\Delta = 0.25$ hr, the discount rate $\gamma = 1$, and the meeting probability functions as those calibrated from the simulation results (the exponents on the demand-supply ratio are set according to the theoretical analysis, i.e., $\beta_{s3} = 1, \beta_{e3} = 2$, $\beta_{\tilde{e}3} = 1$). Additionally, we assume all vehicles are vacant and evenly distributed across zones at the beginning of the analysis horizon.

8.5.1. Simple network

Figure 8.3 shows the four-zone network used in this experiment, along with the travel time τ_{ij} between each pair of zones ($\tau_{ij} = \tau_{ji},\ \forall i, j$). Zone i and zone j are considered neighbors to each other if $\tau_{ij} = 1$. Hence, except for the pair of zone 2 and zone 4, all other five pairs are neighbors. The network is served by a fleet of 200 vehicles (i.e., $M = 200$). The analysis horizon consists of 4 time periods (i.e., $T = 4$). In each zone and time period, 100 passengers arrive in each zone uniformly in time (i.e., $q_i^t A_i = 100,\ \forall t, i$), and their destinations are uniformly distributed in all but the origin zones (i.e., $\alpha_{ij} = 1/3$ if $i \neq j$, otherwise 0). Finally, the trip fare is set to be \$1 for a trip between neighboring zones and \$2 otherwise. Other parameters used in this experiment are

set as follows: road density $\rho = 20$ mile per square mile (mi/sqmi), zone area $A_i = 1$ sqmi $\forall i$, and cruising speed $v = 20$ mile per hour (mph).

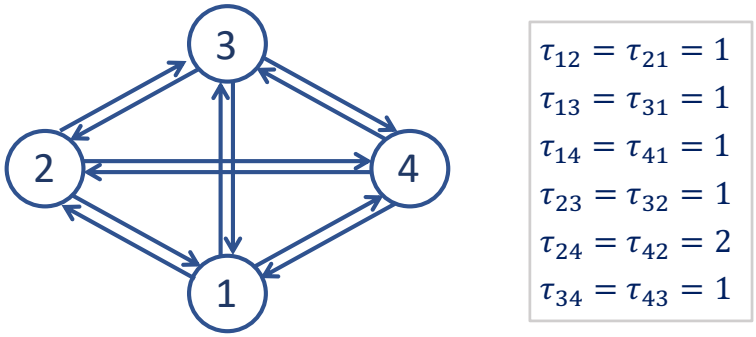
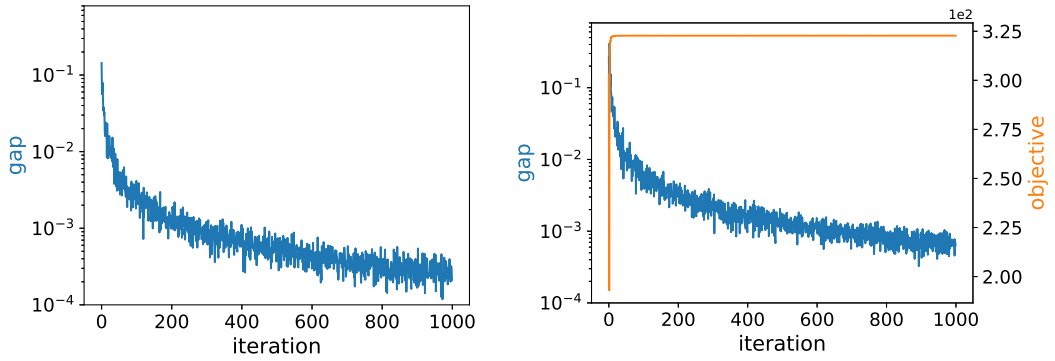


Figure 8.3. Topology of the four-zone network and inter-zonal travel times.



(a) street-hail NC routing

(b) street-hail CO routing

(c) e-hail NC routing

(d) e-hail CO routing

Figure 8.4. Potential and equilibrium gap over iterations.

Figure 8.4 plots the gap values obtained in each iteration when solving the four instances of the RIVER problem, each corresponding to a combination of routing behavioral assumption (i.e., NO/CO routing) and ride-hail model (i.e., street-hail/e-hail). For CO routing, we also plot the objective value. As per Algorithm 8.2, the gap is computed as the total difference between the Q-values associated with positive idle flows and the maximum Q-value at the same state divided by the sum of all absolute Q-values. In other words, it measures how much the system deviates from a Wardrop equilibrium relative to the equilibrium solution. As shown in Figure 8.4, the MSA-based algorithm converges slowly but is capable of producing an approximate equilibrium solution (a gap below 0.001) within a few hundreds of iterations. The only exception is the case of e-hail under CO routing, where the gap only reduces to 0.003 after 1000 iterations. As expected, the objective values for CO routing increase monotonically and quickly stabilize within a few iterations.

Figure 8.5 illustrates vacant vehicular flows $\mathbf{x}$ and vacant vehicle distribution $\mathbf{y}$ at the beginning of each time period at the equilibria of the four RIVER problems. We first note that zones 2 and 4 are evidently more popular destinations for vacant vehicles than the other zones. This preference is developed mainly because trips originated in these two zones yield a higher expected revenue, as the trip fare between zones 2 and 4 is higher than the other trips.

The results suggest CO routing contributes to a more even distribution of vacant vehicles in the network. In particular, it helps mitigate the strong preference for zones 2 and 4. With the exception of the first time interval under e-hail, the maximum inter-zonal difference in the number of vehicles is smaller in CO than NO routing. This result is expected because, in CO routing, drivers anticipate the impact of their relocation decisions on other drivers' expected returns, which tend to discourage excessive congregation of idle drivers in one zone. Another important observation from Figure 8.5 is street-hail has more vehicles in the idle state than e-hail regardless of the assumption about the routing behavior. Take NC routing as an example. Excluding the first period (when the total numbers of vacant vehicles are equal in all scenarios), the average number of vacant vehicles per zone per period is 31 for street-hail and 27 for e-hail. It is reasonable to attribute this 13%

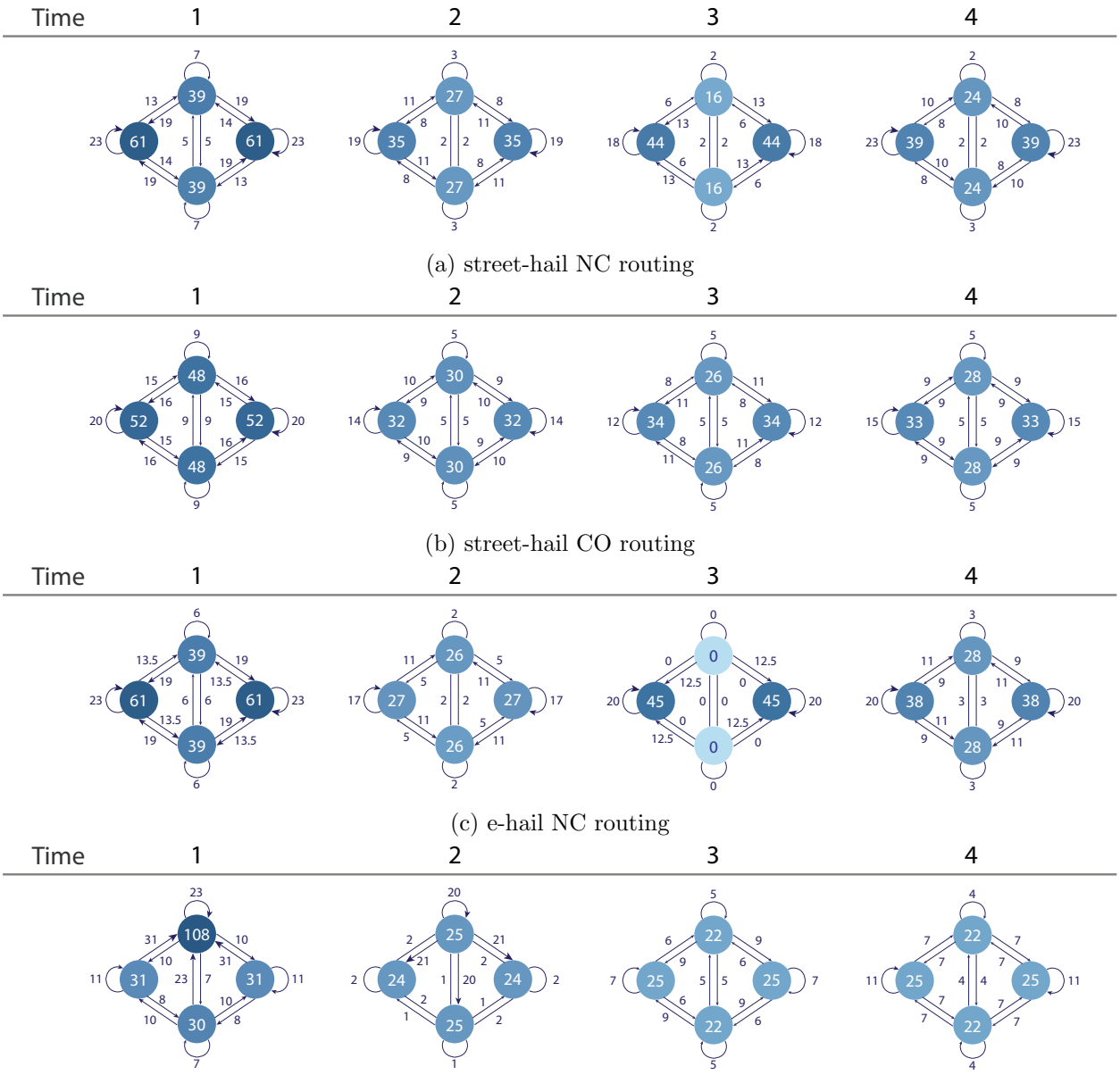


Figure 8.5. Vacant vehicular flow (link labels) and vacant vehicle distribution (node labels) at equilibrium.

drop in the number of vacant vehicles to the advanced matching technology of e-hail. Interestingly, cooperation among drivers appears to magnify this advantage of e-hail. Under CO routing, the average number of vacant vehicles remains the same for street-hail, but drops another 11% to 24 for e-hail.

While e-hail utilizes the fleet more efficiently on average than street-hail, it does not always distribute the vacant vehicle more evenly in the network, especially in the case of NC routing. The worst case is $t = 3$, when all vacant vehicles gather in zones 2 and 4, leaving the other two completely unattended. The reason why e-hail seems more prone to such extreme episodes is not entirely clear, though it is definitely rooted in the different matching mechanisms. Another interesting observation is a major fraction of vehicles are guided to zone 3 at time 0. It seems to help vehicles distribute more evenly in the following time periods, though more investigation into the underlying mechanism is needed.

8.5.2. Chicago network

We next conduct a case study constructed using empirical data collected in the City of Chicago. The same nine communities analyzed in Section 7.6 are selected as the study are; see Figure 8.6. Each community is treated as a zone and communities adjacent to each other are considered neighbors (e.g., Loop has three neighbors, i.e., Near North Side, Near West Side and Near South Side). The analysis horizon is taken as an average weekday from 6:00 to 22:00, divided into 64 quarter-hour periods. All input variables are estimated from the publicly available data collected in the city; see Chapter 10 for more details.

We test two fleet sizes, 3300 and 9000, which are the numbers of e-hail drivers in the city who averaged ten and five rides a day (see details about the data in Section 10.1). Figure 8.7 plots the temporal demand rate $q^t = \sum_i q_i^t$ in an average day, which shows two distinctive peaks around the regular rush hours at 9000 pax/hr. Hence, a fleet of 3300 is relatively small compared to this peak demand level and is used to represent a market with a relatively tight supply.[4] The 9000-vehicle fleet, on the other hand, should be more than adequate even for the peak period.

To better analyze the results, we first define a few aggregate measures of the system performance as follows:

[4]The average trip duration is 0.3 hr and thus the occupied vehicle is around 3000.

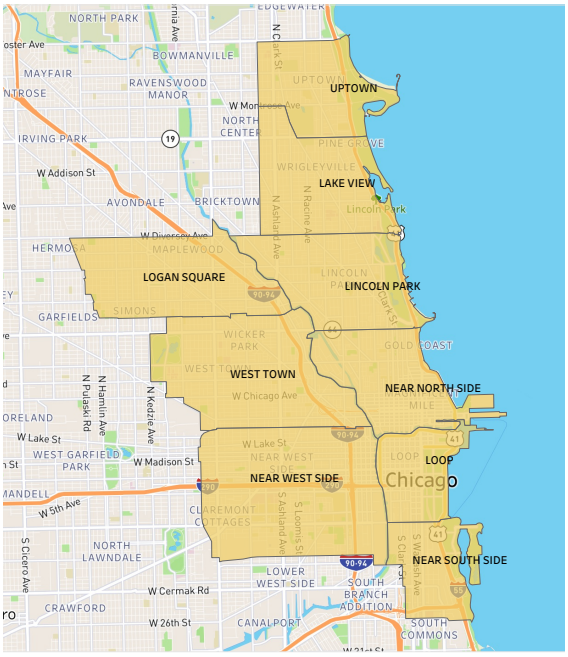


Figure 8.6. Nine communities selected for the Chicago case study.

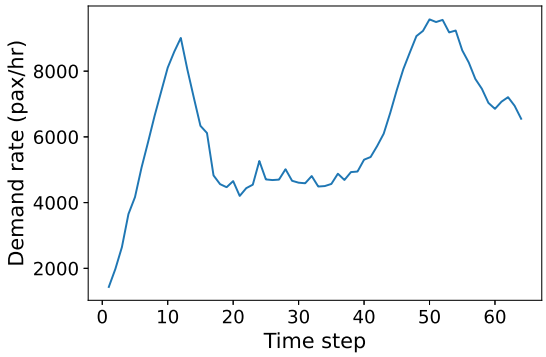


Figure 8.7. Demand rate over time steps.

- Total return:

$$TR = \Phi_c(\pi) = \sum_t \gamma^t \left[\sum_i y_i^t r(y_i^t, q_i^t) \right]. \tag{8.37}$$

- Average return:

$$AR = \frac{1}{M} \Phi_c(\pi). \tag{8.38}$$

- Price of anarchy (PoA):

$$PoA = \frac{\Phi_c(\pi) \text{ under CO routing}}{\Phi_c(\pi). \text{ at NC routing}} \tag{8.39}$$

- Vehicle utilization rate:

$$\mu^t = 1 - \frac{1}{M} \sum_i y_i^t. \tag{8.40}$$

- Fraction of served demand [5]:

$$r^t = \frac{\sum_i m_i^t y_i^t}{\sum_i q_i^t}. \tag{8.41}$$

As mentioned earlier, PoA measures the loss of efficiency due to selfish routing behavior of ride-hail drivers. The concept has been extensively studied in the context of routing game Roughgarden and Tardos [2002], and is closely related to Braess Paradox Braess et al. [2005].

8.5.2.1. System performance on a normal day. Table 8.3 reports the total return, average return and PoA for the eight combinations of routing behavior, mode and fleet size. With the smaller vehicle fleet, e-hail generates about 35% more reward than street-hail thanks to its more advanced matching technology. In this case, cooperative routing makes little difference, with PoA hovering slightly above 1 for both street-hail and e-hail. Yet, e-hail in general induces a larger PoA, meaning there is more potential in optimizing their operations. Increasing the fleet size does improve the total return in both service modes, though individual drivers suffer a more than 50% loss in their average return. However, so far, we have fixed the demand in the model. In practice, it is expected the improved level-of-service thanks to a larger vehicle supply would attract more passengers choosing ride-hail, which, in turn, enhance the reward of each driver.

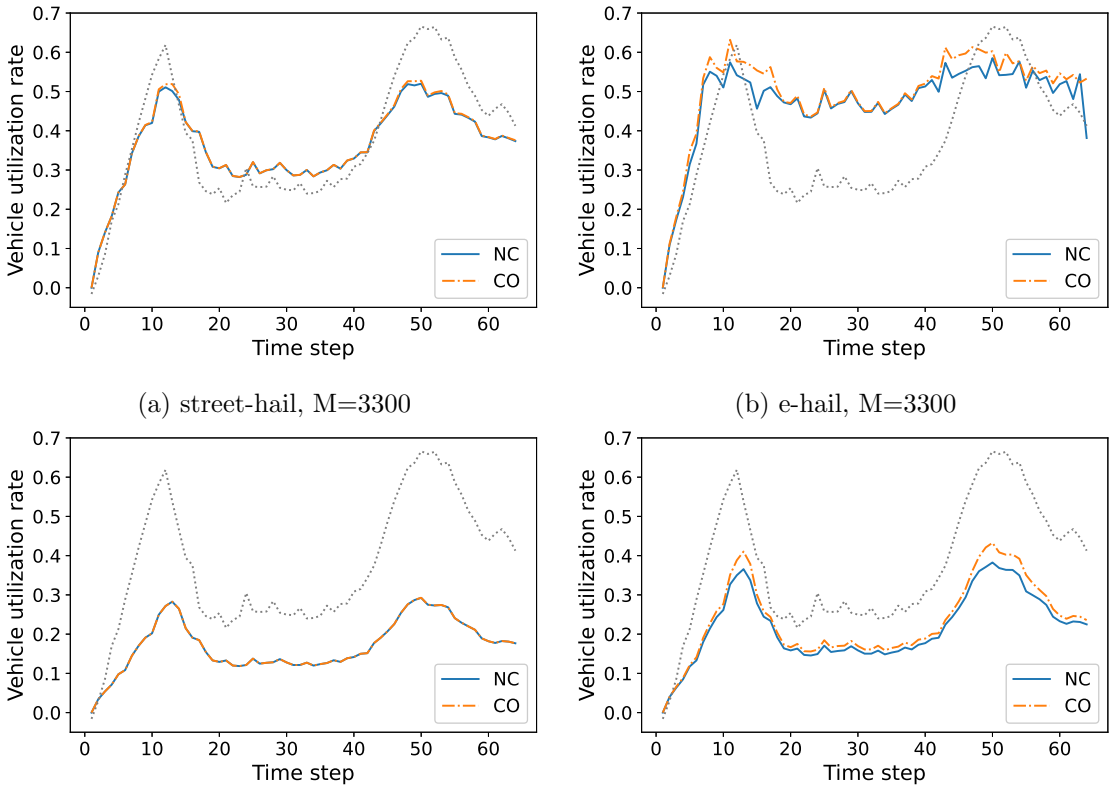
[5]Passengers who fail to find a match in a time interval are assumed to leave the ride-hail market.

Table 8.3. Reward and price of anarchy on a normal day.

Mode	Type	M = 3300			M = 9000		
		TR ($K)	AR ($)	PoA	TR ($K)	AR ($)	PoA
street-hail	NC	644.79	195.39	1.0027	827.92	91.99	1.0003
	CO	646.50	195.91		828.81	92.02	
e-hail	NC	880.08	266.69	1.0281	1053.26	117.03	1.0800
	CO	904.85	274.20		1137.48	126.39	

Figures 8.8 and 8.9 present the vehicle utilization rate and the fraction of served demand overt time. For reference, the temporal demand pattern is plot as a gray dash line. As shown in Figure 8.8, the vehicle occupancy closely follows the demand pattern. Comparing the two service modes, we find e-hail achieves a much higher utilization rate than street-hail, especially during the off-peak period. This is, again, expected and largely due to the more advanced matching technology of e-hail. As the fleet size increases, the utilization rate for both modes drops and the advantage of e-hail diminishes. Like street-hail, the utilization rate of e-hail vehicles drops in the middle of the day.

As illustrated in Figure 8.9, under the smaller vehicle fleet, the fraction of served demand shows an opposite pattern of demand, i.e., the peaks on one curve often meet with the valleys on the other. This finding is expected because, during the peak periods, the supply is insufficient to served the increasing demand and thus the portion of served demand drops. In contrast, during the off-peak periods, almost all passengers are served in e-hail. Due to a larger matching friction, street-hail tends to serve a smaller portion of the demand, but that portion also appears to be less sensitive to the temporal variations than e-hail. The loss of demand during peak periods can be largely lessened by increasing the fleet size. Nevertheless, the NC routing still leaves a small fraction of demand unserved, while this is no longer an issue in cooperative routing (see Figure 8.9(d)).



(a) street-hail, M=3300

(b) e-hail, M=3300

(c) street-hail, M=9000

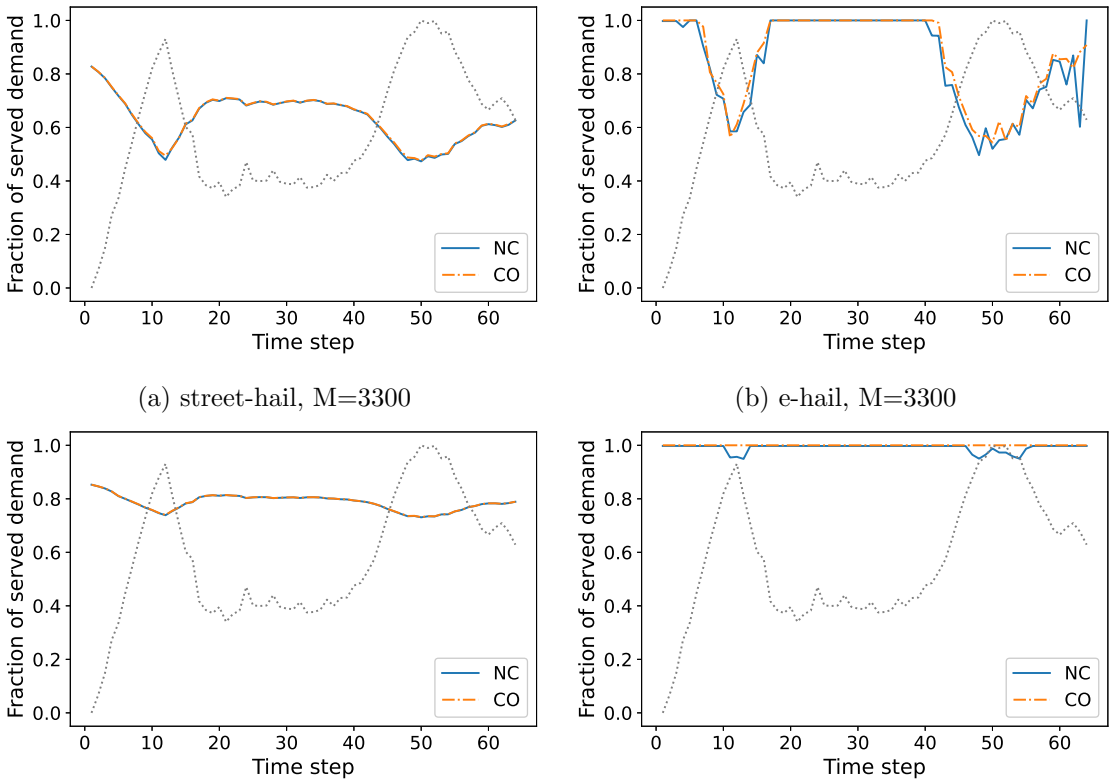
(d) e-hail, M=9000

Figure 8.8. Vehicle utilization rate over time steps.[5]

In each subplot of Figures 8.8 and 8.9, the curves corresponding to NC and CO routing are very close to each other, although the gap for e-hail is slightly larger than that for street-hail. This finding agrees with Table 8.3, indicating again CO routing does not make much of a difference in this case. The largest gap between NC and CO is observed in e-hail during the peak period, when vehicle cooperation improves both utilization rate and serves more passenger demand (see Figures 8.8(d) and 8.9(d)).

8.5.2.2. Impact of events. The last set of experiments aim to investigate the impact of temporal demand peaks and the lack of demand information on the system performance. To this end, we create two hypothetical baseball games in Lake View, where the Chicago baseball stadium is located (see Figure 8.6). Following the regular game schedule, one event is set in the evening (from 18:40

[5] The temporal demand rate is plot as gray dash line to show the demand pattern over time.



(a) street-hail, M=3300

(b) e-hail, M=3300

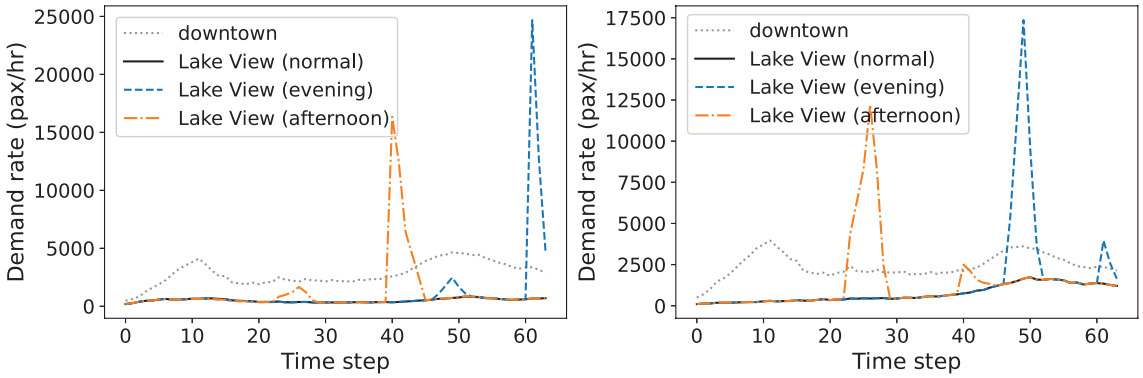
(c) street-hail, M=9000

(d) e-hail, M=9000

Figure 8.9. Fraction of served demand over time steps.[5]

to 21:40) and the other is in the afternoon (from 13:20 to 16:20). We assume the two events both draw 10,000 attendance uniformly distributed among all zones. These people travel to Lake View before the game starts and go back to their home zones right after the event ends. Accordingly, each event generates two considerable demand peaks in addition to regular rush hours, as shown in Figure 8.10. Note that the demand rate in the figure is higher than 10,000 passengers per hour because it is computed based on the 15-minute interval.

With these new defined demand profiles, we study four routing scenarios in e-hail with a large fleet size ($M = 9000$): (i) non-cooperative drivers following their strategies on normal day ("myopic NC"), (ii) cooperative drivers following their strategies on normal day ("myopic CO"), (iii) non-cooperative drivers updating their strategies in response to the updated demand ("NC"), and (iv) cooperative drivers updating their strategies in response to the updated demand ("CO"). In other



(a) aggregated by origin (b) aggregated by destination

Figure 8.10. Demand rate in downtown zones and Lake View.

words, the two myopic driver groups do not have the demand information about the events thus simply follow their original strategies. Table 8.4 reports the total return and PoA under the four routing types, where the PoA is all computed based on the total return under CO. As expected, for both events, the two myopic cases fall behind the other two due to the lack of demand information. The higher total returns observed in the afternoon game imply that the extra demand is better served in this cases compared to the evening game. A possible reason is that the evening game overlaps with the rush hours on a normal day thus fewer vehicles are available to serve the extra demand. For the same reason, the total return of CO routing in the afternoon game is higher than that in the afternoon game as more vehicles are available for relocation in response to the additional demand peaks.

Table 8.4. Reward and price of anarchy on an event day.

Type	evening game		afternoon game	
	TR ($K)	PoA	TR ($K)	PoA
myopic NC	1122.77	1.1268	1182.68	1.1443
myopic CO	1163.29	1.0876	1189.85	1.1374
NC	1181.73	1.0706	1255.24	1.0782
CO	1265.19		1353.38	

The PoA values show a similar pattern as the total returns, Specifically, the largest PoA is observed in the afternoon game under myopic NC routing, which leads to an efficiency loss more than 14%. Interestingly, the results indicate the two events do not cause additional loss of efficiency,

as their corresponding PoA values remain similar to that on a normal day (1.08 as in Table 8.3). In contrast, the lack of demand information does induce large efficiency loss. The result thus emphasizes the importance of demand information in the operations of ride-hail services. Table 8.4 also implies the impact is more severe when the extra demand peaks happen during an off-peak period (the afternoon game in this experiments).

Another intriguing finding from Table 8.3 is about the difference in PoA between NC routing and myopic routing. In the evening game scenario, the PoA scores of NC and myopic CO are very close to each other, while the gap between myopic NC and myopic CO is much larger. On the contrary, in the afternoon game scenario, the two myopic routing types lead to a similar PoA that is significantly below that obtained by NC routing. This is again likely due to the different demand patterns generated by the two events. Specifically, the afternoon game induces a demand pattern that is very different from the normal day. Thus, with such demand information, the NC drivers are able to operate more efficiently than the myopic drivers. In this case, the advantage of cooperation among myopic drivers also diminishes given the incorrect perception of the demand.

8.6. Summary

In this chapter, we model the routing behaviors of idle vehicles in a spatiotemporal ride-hail market, which we refer to as the RIVER problem. At the beginning of each time interval, idle drivers select the next zone for passenger search to maximize their expected return. Such a decision making process is modeled as a Markov decision process (MDP). Collectively, drivers can be seen as playing an MDP congestion game, with congestion being induced by passenger-driver matching in each zone. This matching process is specified for street-hail and e-hail as meeting probability (i.e., the probability of meeting a passenger after searching for one period) functions that are calibrated with simulation data. We define the Wardrop equilibrium for the MDP congestion game and propose to find it by solving a fixed-point problem. The framework is further extended to the case of cooperative routing, in which drivers route cooperatively to maximize the total return. We

show that the according optimization problem can be solved as a Wardrop equilibrium of RIVER problem with an adjusted reward function.

The RIVER problems, both the original and the cooperative versions, are tested in two sets of experiments. One is a based on a hypothetical network and the other is constructed based on a real market. The results show the service mode plays a critical role in shaping the spatial distribution of vacant vehicles and the system performance. Thanks to its more advanced matching technology, e-hail improves the fleet utilization ratio and delivers a higher average return than street-hail. However, it also appears to be more vulnerable to the price of anarchy than street-hail. This disadvantage, however, is insignificant in the case study on real market, with either small or large vehicle fleets. The loss of efficiency for e-hail is well below 5% while that for street-hail is close to zero.

Besides regular demand patterns, we test two counterfactual scenarios. Specifically, we consider a game event, either in the afternoon or evening, happens far from the city and induces demand peaks right before and after it. It is found the system could suffer from a substantial loss of efficiency (up to 9%) if the selfish drivers fail to adjust their search strategies in accordance to the demand pattern. Without demand information, the routing of cooperative drivers is also sub-optimal. However, the loss could be reduced by one-third with myopic cooperative drivers. Even with full information, the non-cooperative drivers still perform worse than normal days.

There are many directions along which the present study can be extended. First, the current model only considers the supply side of the market. Hence, a logical next step is to integrate demand modeling into the market equilibrium. Second, we have revealed in Chapter 7 that the TNC operations have great impact on urban congestion. A future study may formulate the cooperative RIVER problem with the objective to maximize social welfare that accounts for the congestion externality of the ride-hail vehicles. Thirdly, the modeling framework presented in this paper may be extended to support ride-hail service operations, such as designing relocation incentives for drivers.

8.7. Appendix

8.7.1. Notations

Table 8.5. List of notations

Variable	**Description**	**Unit**
$\mathcal{N}$	set of zone	
$\mathcal{A}_i$	set of zone i's neighbor zones	
$\mathcal{M}$ (M)	set (number) of drivers	
T	number of time periods	
Δ	length of each time period	hr
q_i^t	demand rate from zone i at time t	/hr/sqmi
α_{ij}	trip fraction from zone i to zone j	
p_{ij} ($\bar{p}_i$)	trip fare (average trip fare from zone i)	\$
τ_{ij}	travel time from zone i to zone j	\$
x_{ij}^t	number of relocation vehicles from zone i to zone j at the beginning of time t	
y_i^t	number of vacant vehicle in zone i at the beginning of time i	
m_i^t	meeting probability in zone i at time i	
d_i^t	number of vehicles arriving in zone i by the end of time i	
ρ	road density	mi/sqmi
v	cruising speed	mph
φ ($\varphi^\star$)	demand-supply ratio (that separates normal and oversupply regimes in e-hail)	

Part 4

Data and numerical tool

CHAPTER 9

Agent-based simulation of ride-hail service

The models developed in previous chapters are all purely theoretical. A number of strong assumptions have been imposed to ensure analytical tractability. To validate these theoretical models and also to support our future research on ride-hail service, we develop an agent-based simulation framework based on MATSim (Multi-Agent Transport Simulation). The simulator allows travelers in an urban transportation system to choose among ride-hail services and other modes (e.g., driving and transit). On the other side of the market, each ride-hail driver can freely enter or exit the market according to the earning potential. Additionally, the matching between passengers and vehicles is conducted by specific matching algorithms (e.g., bipartite matching) based on the service mode, instead of dictated by the matching function as per the theoretical models. Further, an interface is developed that allows key control variables (e.g., pricing and matching interval) to be adjusted throughout the simulation. Therefore, the simulator also supports service design through simulation-based optimization (SBO).

In what follows, we first introduce MATSim, the baseline simulation framework, and then outline the ride-hail module. After that, we present a pricing module based on the idea of metamodel-based SBO [Barton and Meckesheimer, 2006] that adaptively changes the pricing scheme to optimize the platform's objective. A series of cases studies are reported that validate the performance of both the simulator and the pricing module.

9.1. Introduction of MATSim

MATSim (Multi-Agent Transport Simulation) was initially designed for activity-based travel demand modeling, but has found a wider range of applications in transportation research Horni

et al. [2009], Bassolas et al. [2019], Hörl et al. [2019]. It has an efficient built-in mesoscopic traffic simulator that moves individual vehicles according to macroscopic traffic flow theory Horni et al. [2016]. We choose to develop the proposed ride-hail simulator as an independent MATSim module for several reasons. First, MATSim is open-sourced and modularized. Hence, we can easily link the ride-hail module to MATSim's core simulator and other add-on modules. Second, MATSim provides "template" passenger and vehicle agents, along with a sophisticated travel demand model. These elements dramatically simplify the efforts needed to represent the demand and supply sides of the ride-hail market. Last but not least, MATSim has a built-in vehicle dispatching and routing mechanism, from which most TNC operations can be derived.

A complete MATSim run consists of five steps shown in Figure 9.1. It starts with loading the *plans* of each person. Each plan consists of a series of activities and travel legs between them. In every iteration, each person selects and executes a single plan. The aggregate travel demands are then loaded onto the transport network in the *mobsim* step (i.e., mobility simulation). The *scoring* step recomputes the score associated with each plan. Based on the scores, a fraction of the population is allowed to change their plans in the *replanning* step. The simulation is based on a co-evolutionary principle: each person repeatedly optimizes their plan while competing with others for the service capacity of different transport systems (e.g., roads, transit). This process continues until no one has incentive to further change their plan, which implies the average plan scores stabilize. Finally, statistics summarizing system performance are computed and written into the output files in the *analyses* step.

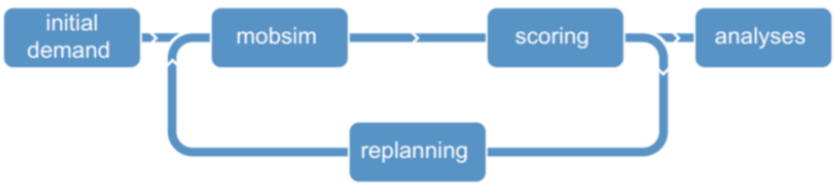


Figure 9.1. The MATSim loop (source: [Horni et al., 2016])

The proposed ride-hail module is integrated into the *mobsim* step. As a new alternative mode, ride-hail offers an additional plan to each person. Thus, it also interacts with the core simulator in the *scoring* and *replanning* steps. Additionally, relevant aggregate measures of ride-hail services (e.g., average passenger waiting time, vehicle utilization rate) need to be computed and stored. Hence, the module also functions in the *analyses* step.

9.2. Ride-hail module

Figure 9.2 illustrates the structure of the ride-hail module. The module consists of three building blocks of a typical ride-hail market: supply, demand, and operation. On the supply side, we define vehicle agents with market entry decisions and cruising behaviors. On the demand side, passengers are defined by their mode choices, which are influenced by, among other things, ride-hail price and level-of-service (LOS). The operation contains trip dispatching and pricing modules. The former collects trip requests from the passengers and matches them with available drivers, while the latter determines trip fare paid by passengers and the compensation rate received by drivers. Each of the components will be discussed in more details in following subsections.

It is worth noting that a module for demand responsive transit (DRT) has developed in MATSim Cich et al. [2017], which shares a number of similarities with ride-hail module proposed here. However, to the best of our knowledge, the current DRT module does not support flexible supply and vehicle cruising. Nor does it embed price optimization.

9.2.1. Passenger demand

Passengers in MATSim adaptively change their daily plan over iterations according to the *score* of each feasible plan. When the daily plan is set to be two dummy activities connected by a single travel leg, the plan score u^{plan} is equivalent to trip utility u^{trav}, i.e.,

$$u^{\text{plan}} = u^{\text{trav}} = \alpha^{\text{mode}} + \beta^{\text{wait}} w + \beta^{\text{trav}} \tau - f, \tag{9.1}$$

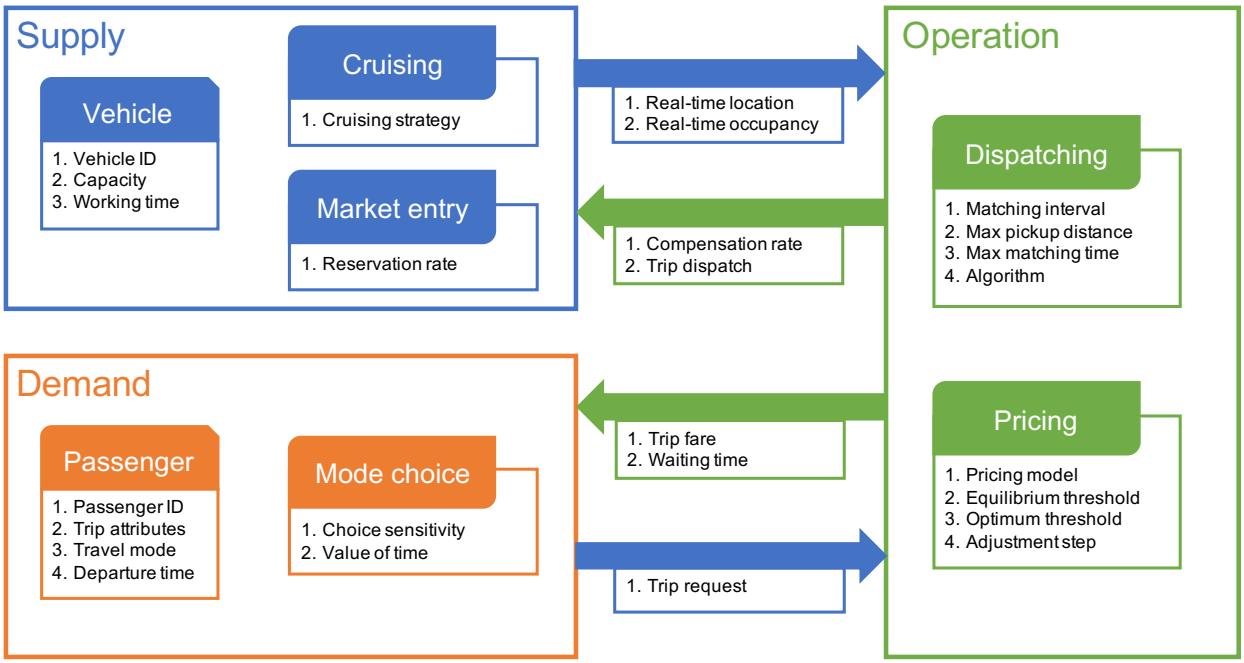


Figure 9.2. Framework of ride-hail module

where α^{mode} is a mode-specific constant, β^{wait} and β^{trav} are the marginal utilities of waiting and travel time, respectively. The waiting time w, travel time τ and the monetary cost f (e.g., trip fare) are updated in the *replannning* step so that passengers are able to reevaluate the plan scores. The default demand model implemented in MATSim is based on multinomial Logit (MNL), which aligns with the setting in most of the numerical experiments in this dissertation (see e.g., Section 4.6).

9.2.2. Vehicle supply

A key difference between TNCs and conventional taxis is the flexibility in the labor supply Hall and Krueger [2018]. Specifically, a taxi driver usually cannot freely decide whether to enter the market whereas a registered TNC driver can. To account for the fact that the operator does not have direct control over its fleet, we implement a flexible supply model in the ride-hail module.

Following the analytical models developed in previous chapters, we assume each driver has a fallback job option and its earning rate, referred to as the reservation rate, follows a given probability distribution. Accordingly, a driver would choose to join the market if the average earning rate of

ride-hail service, denoted by e, is higher than their reservation rate. The current version computes the earning rate using Eq. (4.4) defined in Section 4.1 and, following Section 4.6, we assume the reservation rate follows a uniform distribution.

The earning rate is computed after each simulation run (i.e., *mobsim* step). However, to prevent large jump in vehicle supply between consecutive simulation runs, we do not use the earning rate in the current run e to determine the supply in the next run. Instead, we import the exponential moving average $\bar{e}$ into the supply model, which is computed as

$$\bar{e} = (1 - \alpha_s)\bar{e} + \alpha_s e, \tag{9.2}$$

where α_s denotes the decay rate and has default value $\alpha_s = 0.1$. In other words, drivers choose whether to join market based on a long-term average of historical earning rates.

9.3. Vehicle cruising

The cruising model is another important feature newly developed in the ride-hail module. In practice, vacant ride-hail vehicles usually cruise around to search for passengers or relocate to areas with higher demand. To capture such behaviors, we implement a cruising model in which each vacant vehicle is directed to a cruising destination based on a given cruising strategy. Once the vehicle receives a ride request, it immediately terminates its current cruising trip and sets out to pick up the passenger. If the vehicle reaches the cruising destination without meeting a passenger, a new cruising destination will be generated. In the current version, we only implement a simple random-cruising strategy to all vehicles (i.e., randomly assign cruising destination and if the cruising destination coincides with the vehicle's current location, the vehicle would wait there for a certain time interval). However, the framework has been developed to support various cruising strategies, including the one proposed in Chapter 8.

9.4. Trip dispatching

The trip dispatching module is designed to accommodate different ride-hail modes and dispatching algorithms. The current version supports the following three service modes:

- Street-hail (e.g., street-cruising taxis): Passengers wait on street and hail vacant vehicles when they cruise by.
- E-hail (e.g., UberX): Vehicles are dispatched on demand to serve one passenger at a time.
- Pooling (e.g., UberPool): Vehicles are dispatched on demand to serve multiple passengers with overlapping routes. Pooling passengers are picked up and dropped off in sequence.

The trip dispatching process has three important parameters:

- *Matching interval* decides how often the trip dispatcher is called to process trip requests in batch.
- *Maximum pickup distance* sets a limit on how far away a driver can be summoned to serve a passenger.
- *Maximum matching time* limits how long a passenger is willing to wait for a match before switching to an alternative mode.

We next discuss in detail how the trip dispatcher works in each service mode.

9.4.1. Street-hail

Since passengers and vehicles in street-hail are physically matched on street, the maximum pickup distance, also called hail distance in [Chen et al., 2018], is rather limited (15-50 meters). In addition, as cruising vehicles and waiting passengers keep searching for each other, the matching interval is simply the minimum simulation interval (1 second in MATSim). Thus, the trip dispatcher continuously tracks the location of each idle vehicle and waiting passenger, and creates a match whenever the distance between an idle vehicle and a waiting passenger is smaller than the maximum pickup distance.

9.4.2. E-hail

Although many dispatching strategies may be implemented for e-hail, the current version only implements a basic, perhaps the most commonly used, matching algorithm. Yet, the framework is flexible to integrate more advanced matching algorithms. At the end of each matching interval, the trip dispatcher solves the following bipartite matching problem:

$$\min \quad \sum_{ij} c_{ij} x_{ij} \tag{9.3a}$$

$$s.t. \quad \sum_{j} x_{ij} = 1, \tag{9.3b}$$

$$\sum_{i} x_{ij} \leq 1, \tag{9.3c}$$

$$x_{ij} \in \{0, 1\}. \tag{9.3d}$$

In Problem (9.3), the binary decision variables x_{ij} indicates whether passenger i is assigned to driver j, and c_{ij} represents the pickup cost (e.g., pickup time). Constraint Eq. (9.3b) states every passenger is matched with a vehicle while Eq. (9.3c) indicates each vehicle is at most assigned to one passenger. Thus, the problem seeks to minimize total pickup cost while matching all passengers. Here, it is assumed the number of vehicles is larger than that of passengers. If the opposite applies, Eq. (9.3c) turns to be an equality constraint while Eq. (9.3b) becomes an inequality constraint.

9.4.3. Pooling

In the pooling mode, the trip dispatcher attempts to bundle together multiple passengers whose trips have as much overlap as possible. The current ride-hail module implements a simple pooling strategy that following the matching process described in Chapter 3. At the end of each matching interval, the trip dispatcher will first match passengers in pairs of two and then assign these pooling pairs to vacant vehicles. The procedure is summarized in Algorithms 9.1. To prevent unreasonably long detours in pooling trips, we introduce another parameter named maximum detour time $\bar{\Delta}$. If

the en-detour of any passenger in a feasible pooling pair is larger than $\bar{\Delta}$, we assign an arbitrary large value M to this pair, which prevents it from being selected in the first matching problem. Finally, the passengers who do not find a pooling pair will be matched with remaining vacant vehicles in the same way as e-hail trip dispatching.

Algorithm 9.1 Two-step bipartite matching

Inputs: Set of unmatched trip requests $\mathcal{R}$; set of available vehicles $\mathcal{V}$; maximum detour time $\bar{\Delta}$; arbitrary large value M.
Outputs: Assignments for pooling trips $\mathcal{A}_{\text{pool}} = \{(i,j,v) : i,j \in \mathcal{R}, v \in \mathcal{V}\}$ and solo trips $\mathcal{A}_{\text{solo}} = \{(i,v) : i \in \mathcal{R}, v \in \mathcal{V}\}$.
Initialize $\mathcal{A}_{\text{pool}} = \emptyset$ and $\mathcal{A}_{\text{solo}} = \emptyset$.
if $|\mathcal{R}| > 0$ and $|\mathcal{V}| > 0$ **then**
    for each pair of requests $i, j \in \mathcal{R}$ **do**
        Compute the minimum total trip distance starting with picking up i, $\underline{\tau}^i$, and detours for i and j, Δ_i^i and Δ_j^i.
        if $\Delta_i^i > \bar{\Delta}$ or $\Delta_j^i > \bar{\Delta}$ **then**
            Set $\underline{\tau}^i = M$.
        end if
        Compute the minimum total trip distance starting with picking up j, $\underline{\tau}^j$, and detours for i and j, Δ_i^j and Δ_j^j.
        if $\Delta_i^j > \bar{\Delta}$ or $\Delta_j^j > \bar{\Delta}$ **then**
            Set $\underline{\tau}^j = M$.
        end if
        Compute the average minimum total trip distance $\underline{\tau}_{ij} = \begin{cases} (\underline{\tau}^i + \underline{\tau}^j)/2, & \underline{\tau}^i < M \text{ and } \underline{\tau}^j < M \\ M, & \text{otherwise} \end{cases}$
    end for
    Match pooling passengers into pairs by solving Problem (9.3) where the cost is given by $\underline{\tau}_{ij}$.
    Break loops in assignments and keep pairs with the smaller total trip distance.
    Add pooling pairs into set $\mathcal{R}_{\text{pool}} = \{(i,j) : i,j \in \mathcal{R}\}$ and other requests into set $\mathcal{R}_{\text{solo}} = \{i : i \in \mathcal{R}\}$
    for each pooling request $k = (i,j) \in \mathcal{R}_{\text{pool}}$ and each vehicle $v \in \mathcal{V}$ **do**
        Compute pickup cost $c_{kv} = \arg\min_{c \in \{c_{iv}, c_{jv}\}} \left\{c_{iv} + \underline{\tau}^i, c_{jv} + \underline{\tau}^j\right\}$, where c_{iv} (c_{jv}) are the cost to pick up i.
    end for
    Match pooling requests with vehicles by solving Problem (9.3) where the cost is given by c_{kv}.
    Add assignments results to $\mathcal{A}_{\text{pool}}$ and removed matched vehicles from $\mathcal{V}$.
    Break remaining pooling trips into single requests and add them into $\mathcal{R}_{\text{solo}}$.
    for each solo request $i \in \mathcal{R}_{\text{solo}}$ and each vehicle $v \in \mathcal{V}$ **do**
        Compute pickup cost c_{iv}.
    end for
    Match solo requests with vehicles by solving Problem (9.3) where the cost is given by c_{iv}.
    Add assignments results to $\mathcal{A}_{\text{solo}}$.
end if

Again, the developed framework is compatible with other matching algorithms for pooling trips.

9.5. Pricing module

The pricing module controls and optimizes the pricing strategy of the service. The simulation can be executed with or without the price module. In the absence of pricing, a preset pricing strategy is used. When the pricing module is activated, an iterative process is introduced to adjust the pricing strategy according to the "current" market equilibrium reached in simulation (see Figure 9.3). After a complete simulation run, the module checks whether an equilibrium is reached by computing the *market share* as the fraction of passengers choosing ride-hail service and the *market entry* as the fraction of drivers entering the market. Both values should stabilize at equilibrium. Once an equilibrium is reached, the optimizer is called to update the pricing strategy according to the gradient computed by the surrogate model. The key advantage of this method is its ability to anticipate the effect of any small price change on the market equilibrium. Specifically, the gradient gives the direction along which the objective (e.g., the platform profit) could be improved in the fastest way. After each update, it returns back to simulation to equilibriate the market again under the new pricing strategy. If the gradient is below certain threshold, the current pricing strategy is considered close enough to the optimum and thus the entire simulation terminates. This threshold is also a predefined parameter.

In the current version, the monopoly pricing model developed in Chapter 5 is implemented as the surrogate model and the pricing objective is set to be maximizing the platform profit. However, it is easy to switch to other objective (e.g., social welfare) or pricing models (e.g., duopoly pricing discussed in Chapter 6). It is compatible with other demand and supply models, so long as they are continuously differentiable. Moreover, the optimal pricing problem can be replaced by other service design problems (e.g., optimizing the matching interval and driver incentives for relocation), while additional modeling efforts are needed to carefully account for the response of demand and supply.

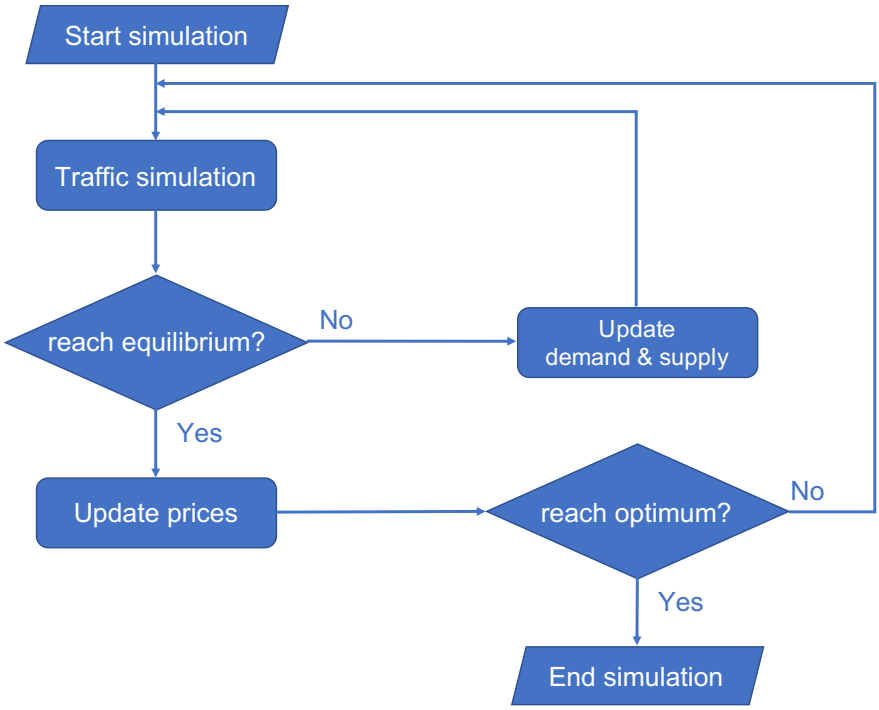


Figure 9.3. Flow chart of pricing module

9.6. Case studies

We conduct a series of case studies to test the functionalities of the ride-hail module and to validate the pricing module. Through these experiments, we aims to answer the following questions:

(1) How does the service quality vary among different service modes under the same demand and supply? Whether the surrogate model could correctly recognize these variations?

(2) How effective is the pricing module to improve the operator's profitability?

(3) Whether the pricing module could be embedded with other mobility services?

In this section, we first introduce the study area and then move to the experiment results.

9.6.1. Study area and baseline model

The Wayne County in Michigan, U.S. (see Figure 9.4) is selected for the case studies since a well-defined road network and calibrated travel demand model are readily available. As the primary objective here is to demonstrate the capability of the ride-hail module, we only consider a one-hour

time window in the peak period (see Figure 9.5) and sample only 1% of the population. The population sample is further filtered by income level and availability of automobiles to obtain a group of homogeneous travelers. This leads to a sample of about 3,000 travelers.

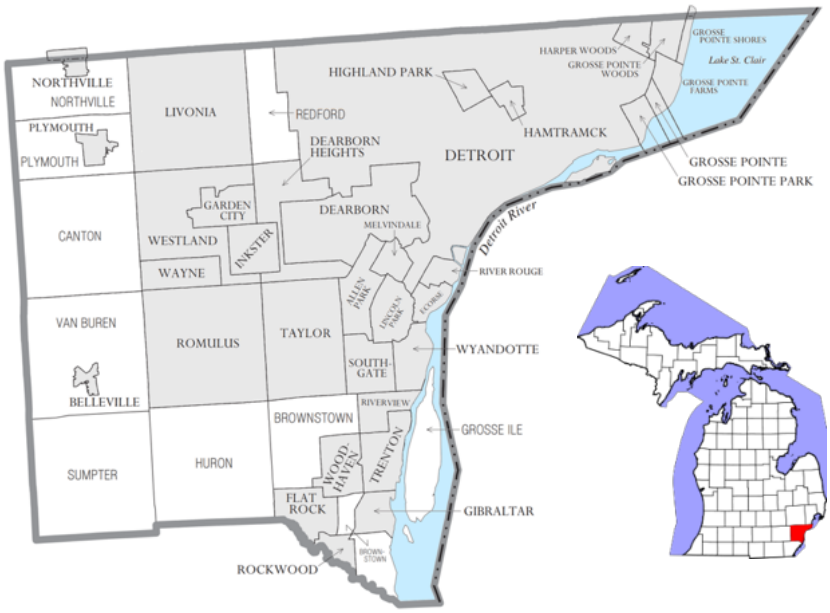


Figure 9.4. Wayne County (Source: Wikipedia).

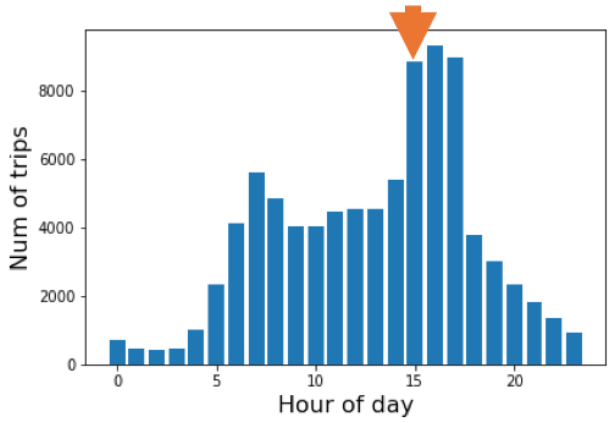


Figure 9.5. Hourly travel demand and selected operation time (pointed by the orange arrow).

We first run a baseline model without ride-hail service. The default travel modes are car, public transit (PT), bike and walk. As shown in Figure 9.6, a majority of the population choose driving while the share of PT is the lowest (below 1%).

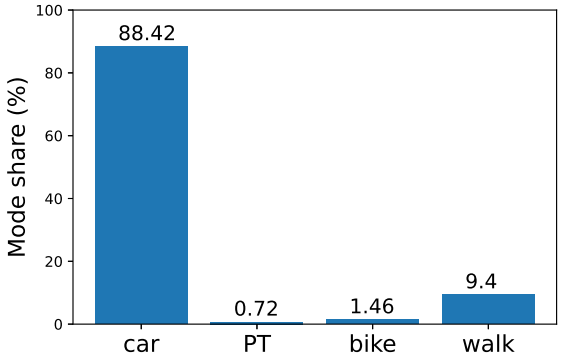


Figure 9.6. Mode share of baseline model.

9.6.2. Service mode design

The first set of experiments are conducted to compare the service quality of the three implemented service modes (i.e., street-hail, e-hail and pooling). As discussed above, we adopt the pricing model developed in Chapter 5, which is based on the matching model presented in Chapter 3. Specifically, a simplified version of the cumulative distribution function (CDF) of street-hail wait time w_s and e-hail wait time w_e is given as follows:

$$\text{street-hail: } F_s(t) = 1 - \exp(-k_s t), \tag{9.4}$$

$$\text{e-hail: } F_s(t) = 1 - \exp(-k_e t^2), \tag{9.5}$$

where k_s and k_e are parameters that are related to the supply-demand relationship, network topology and matching efficiency.

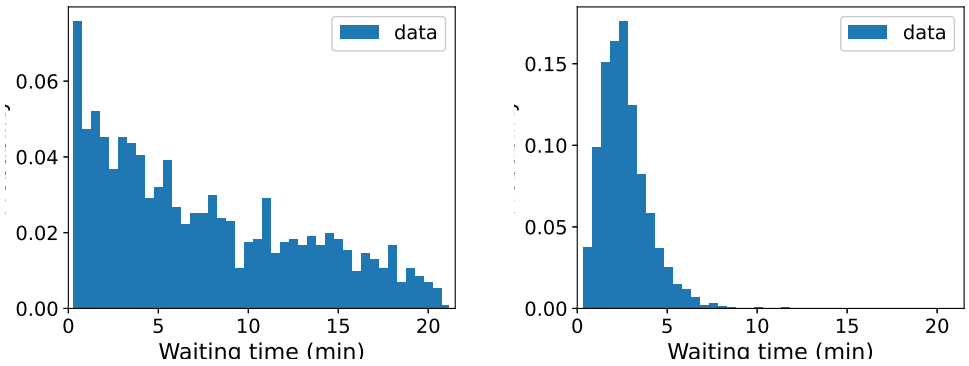
As per Eqs. (9.4) and (9.5), the street-hail wait time follows an exponential distribution with a constant failure rate, whereas that for e-hail follows a Weibull distribution with an increasing

failure rate. This implies, without everything else equal, e-hail passengers are more likely to find a vehicle as wait time increase. In contrast, for street-hail passengers, the likelihood of finding a vehicle at each moment does not change over time. As a result, the distribution of street-hail wait time has a much longer tail than that of e-hail. This difference well explains why e-hail greatly outperforms street-hail in low-density areas. As derived in Chapter 3, the two parts of pooling wait time share the similar physics in matching with e-hail. Hence, they have the same functional form of CDF but with different parameters k_{p1} and k_{p2}, respectively. However, due to the existence of pickup detour, the total wait time for pooling should be more dispersed.

To ensure the pricing module work correctly, we need to first validate the above matching models. We are also interested in comparing the LOS of different modes. To these ends, we fix both demand and supply and use a sufficiently large vehicle fleet (i.e., 1000 vehicles) to ensure most passengers can be served. Following Chapter 3, we assume each pooling trip is shared by two passengers and adopt Algorithm 9.1 for trip dispatching.

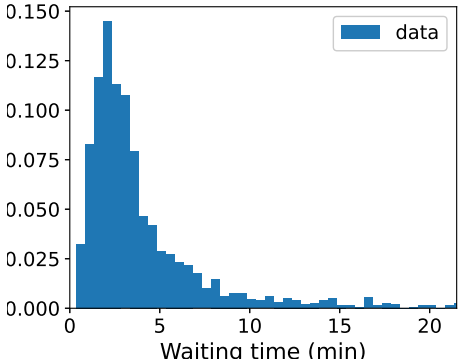
The properties of wait time distributions discussed above are validated by the histograms of simulated passenger waiting times shown in Figure 9.7. While very few e-hail passengers experience a waiting time longer than 10 minutes, a large fraction of street-hail passengers do so. Besides, a few pooling passengers suffer from long waits (more than 15 minute), whereas the other enjoy a quite similar LOS as e-hail.

We proceed to calibrate the waiting time distribution using maximum likelihood estimation (MLE) and plot the cumulative distribution function (CDF) of both simulated and calibrated passenger waiting time in Figure 9.8. The result indicates the model well predicts the waiting time distribution by correctly capturing the matching mechanism in different ride-hail services.



(a) Street-hail

(b) E-hail



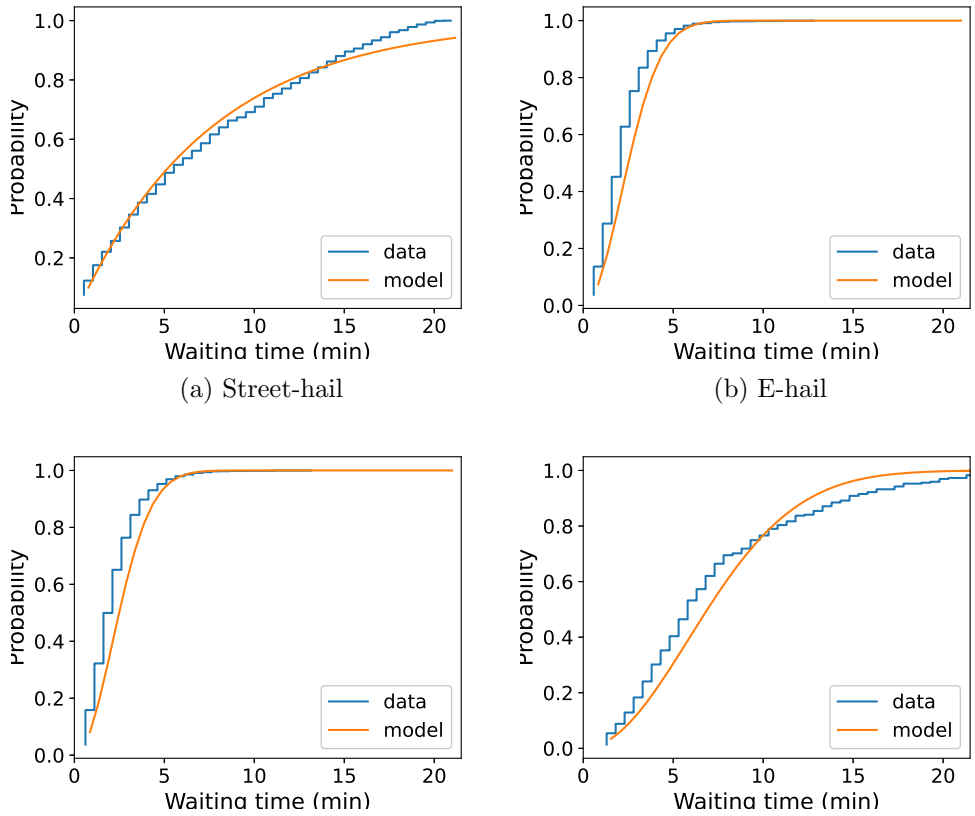
(c) Pooling

Figure 9.7. Histograms of passenger waiting times in different service modes.

Table 9.1. Main calibration results of waiting time distribution.

	street-hail	e-hail	pooling	
			first pickup time	pickup detour
MLE k	2.24 e-03	3.07 e-05	3.11 e-05	4.03 e-06
Mean (min)	7.44 (+0)	3.01 (+0.33)	2.99 (+0.35)	8.29 (+0.51)
90th percentile (min)	17.13 (+1.10)	4.56 (-0.02)	4.53 (-0.09)	12.58 (-2.72)

- Values in the parentheses are differences between model and data. A positive value means it is overestimated by the model.



(a) Street-hail (b) E-hail

(c) Pooling, first pickup time (d) Pooling, pickup detour

Figure 9.8. Simulated and calibrated CDF of passenger waiting time in different service modes.

Table 9.1 reports the main calibration results along with the differences from the simulation data. We find the model well fits the simulation data with the gap of mean estimates less than one minute in all cases. The calibration results also confirm e-hail is more reliable than street-hail with a shorter average waiting time and much fewer long waits. Moreover, the first pickup time of pooling follows a distribution close to e-hail waiting time, while the long waits are largely due to the pickup detour. The average pickup detour is around 8 minutes and around 10% passengers have a pickup detour longer than 12.5 minutes.

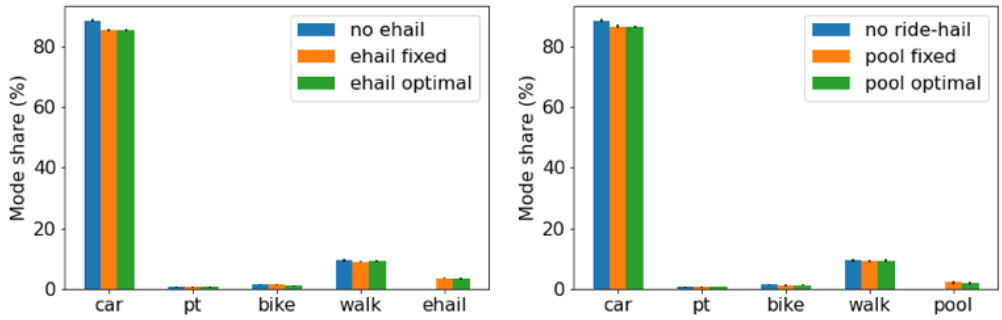
9.6.3. Price optimization

We proceed to test to what extent the pricing module could improve the operator's profit. We run simulations for e-hail and pooling modes separately, with and without price optimization. Here, the fixed pricing strategy is set to be \$1/km for the trip fare rate and \$20/hr for the compensation rate. Since the decision variable of the optimal pricing problem is the trip fare per ride (see Eq. (5.1)), we divide the gradient $\partial R/\partial f$ by the average trip distance based on the simulation results to obtain the moving direction of the trip fare rate. For both e-hail and pooling trips, passengers are rejected if they cannot be matched with vehicles after 5 min. The maximum detour for pooling trips is set to be 15 min.

We consider flexible demand and supply in this case study. Specifically, travelers could choose from ride-hail (e-hail/pooling), driving, PT, walking and biking. Given the small market share of modes other than driving in the baseline model, we assume the potential vehicle supply is 100 vehicles and their reservation rate follows a uniform distribution with an average of \$10/hr. We run ten simulations for each scenario with different random seeds. The maximum number of runs per simulation is set to be 500. The threshold for equilibrium is 1% for both market share and market entry. The threshold for price optimization is 0.05 for the sum of absolute gradients of trip fare and compensation rate. Finally, the moving step size $\alpha = 0.01$.

Figure 9.9 illustrates the mode share in four tested scenarios (e-hail/pool without/with price optimization), along with the baseline model. The bars show mean values across the ten simulation runs while the standard deviations are plotted as the error bars. The very small standard deviations indicate the market evolves to roughly the same equilibrium in each scenario. The results show ride-hail services mainly substitute driving, but the overall mode share is quite low (below 4%). The share of pooling is about half the share of e-hail, only slightly higher than biking. Besides, the price optimization shows little impact on the mode share.

Table 9.2 reports and compares main statistics in the four scenarios. The values superscripted with * means they are with 5% significant level in the two-sample t-test. First, these results confirm



(a) E-hail (b) Pooling

Figure 9.9. Mode share with ride-hail service.*

* The average values are plotted as bars and the standard deviations are plotted as error bars (i.e., the black tick on top of each bar).

the finding that the price optimization makes little differences in mode share, as the change is significant for neither e-hail (see Column (2)-(1)) nor pooling (see Column (4)-(3)). However, e-hail is indeed more attractive to passengers compared to pooling (with a 49% increase in mode share with price optimization).

Price optimization also has a negligible effect on market entry, as the shifts are all within 5% and not statistically significant. Yet, e-hail attracts 22% more drivers than pooling thanks to the higher demand for the service. According to supply model, the market entry is determined by the earning rate (see Eq. (4.3)), which is, in turn, a function of the compensation rate and vehicle utilization rate (see Eq. (4.4)). As shown in Table 9.2, neither of these two variables is affected much by price optimization. Hence, it is hardly surprising that vehicle supply does not change much under price optimization.

Table 9.2. Main statistics of ride-hail services.

	(1) e-hail fixed	(2) e-hail opt.	(2) - (1)	(3) pooling fixed	(4) pooling opt.	(4) - (3)	(4) - (2)
Mode share (%)	3.66	3.46	(-6%)	1.66	1.75	(+5%)	(-49%)*
Market entry (%)	31.08	30.79	(-1%)	22.7	23.95	(+4%)	(-22%)*
Trip fare rate ($/km)	1	1.89	(+89%)*	1	1.48	(+48%)*	(-22%)*
Comp. rate ($/hr)	20	20.35	(+2%)	20	20.04	(+0%)	(-2%)
Trip distance (km)	7.27	7.75	(+7)*	7.2	7.4	(+2%)	(-5%)
Waiting time (min)	9.58	9.79	(+2%)	8.47	8.1	(-2%)	(-17%)*
Veh. occupancy (%)	30.97	31.64	(+4%)*	22.89	23.85	(+4%)	(-25%)*
Profit ($/hr)	269.48	721.34	(+168%)*	159.82	291.44	(+82%)*	(-60%)*

- * means the value is with 5% significant level in the two-sample t-test.

- "opt." stands for "optimized", "comp." stands for "compensation" and "veh." stands for "vehicle"

The demand side, however, is more intriguing. On the one hand, the trip fare does increase considerably after optimization: 89% for e-hail and 48% for pooling. On the other hand, the LOS measured by the average passenger waiting time does not improve for either e-hail or pooling. In fact, it even drops a little for e-hail, though the change is not statistically significant. How could the operator get away with such a dramatic price hike without matching it with a better service? A closer look into the demand model calibrated for Wayne County reveals that travelers are shown to be insensitive to the monetary cost and have a strong propensity for driving (a large constant in the utility function α^{mode}). Evidently, the price optimizer discovers it can raise trip fare substantially at the expense of a rather modest loss of ridership. Moreover, it does not have to attract more drivers with a better compensation, because the pressure to raise the LOS is low. The result is a

significant gain in profit for the operator, though it is not necessarily a desirable outcome for the society.

We note that findings from this case study should not be seen as managerial insights to the real practice for several reasons. First, and most importantly, we are using only 1% sample of the population. As the LOS of ride-hail largely depends on the market scale (i.e., the density of waiting passengers and vacant vehicles Zhang et al. [2019a]), we should not use the LOS under 1% demand and supply to estimate that under a full sample of demand and supply. Second, we have implemented the basic trip dispatching algorithm and adopted simple behavioral models for drivers. To better replicate the real practice, more sophisticated modeling is definitely in need. Last but not least, parameters use in the model (e.g., value of time, reservation rate) need to be recalibrated when implementing the simulation in another area of interest. As discussed above, the price optimizer tends to increase trip fare because travelers in the Wayne County are insensitive to monetary cost. This may not be the case in other places and thus a different pricing strategy may apply. Nevertheless, results reported here demonstrate our model is capable of recognizing the main trade-off in the pricing problem in given scenarios and producing a pricing policy that greatly raises the profit compared to the baseline policy.

9.6.4. Extension to DRT pricing

The door-to-door demand responsive transit (DRT) service shares a similar service form and cost structure with ride-hail, though it typically uses larger vehicles (e.g., minivans) owned by the operator. In this section, we extend the pricing module through the same metamodel framework and integrate it with a MATSim DRT model designed for Wayne County Kagho and Hensle [2021].

As in the previous experiments, we simulate the DRT service with and without price optimization. The default fixed trip fare for DRT is set to be $2.5 per ride. The setting of passenger demand is also the same as before, while the fleet size is fixed at 100 since drivers no longer make the market entry choice. The waiting time function for pooling is adopted here to approximate that

in a DRT service, given their similarity in the passenger-vehicle matching. Since the DRT operator own its fleet, the objective of the optimal pricing problem becomes maximizing $R = fQ - cN$, where the cost per vehicle is set to be a constant \$10/hr, and the only decision variable is the trip fare. Finally, the vehicle capacity is set at 4, which is common in practice.

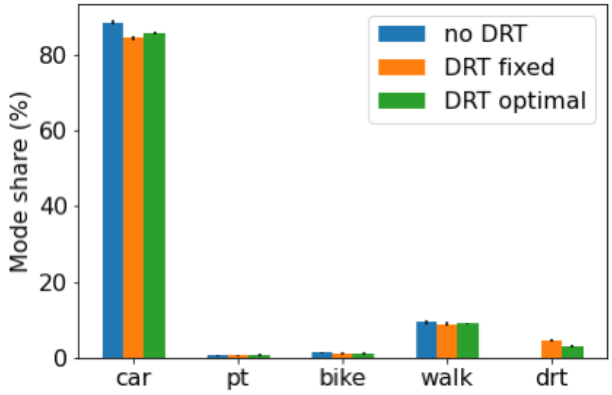


Figure 9.10. Mode share with DRT service. (The error bars are plotted in the same way as Figure 9.9)

As shown in Figure 9.10, DRT has a similarly low mode share as e-hail (around 4%) and mainly substitutes driving. Unlike e-hail, price optimization does compromise DRT's mode share. A better view of the difference is provided by Table 9.3. We find the mode share drops 30% as the price increases almost fourfold. Despite the loss in ridership, the operator's profit grows by 21%. It is interesting to note that the optimized price for DRT is quite close to the that in pooling (see Table 9.2), which indicates the pricing module perceives similar characteristics in the two services. One possible reason is we use the same waiting time function in the surrogate pricing model in both scenarios.

The service quality improves modestly with a slightly lower rejection rate. We note that the rejection rate is quite high because a different rejection mechanism is adopted in the DRT model. Since vehicles cannot freely exit, the market is severely oversupplied with an occupancy rate around 5% (meaning vehicles spend the rest 95% time in the idle state). With such a low utilization rate, the operator is doomed to lose money, and profit-maximization is not nearly enough to eliminate

the very large deficit. The redundant vehicle supply, however, leads to a much shorter passenger waiting time compared to those provided by the ride-hail services.

Table 9.3. Main statistics of DRT services.

	(1) Fixed	(2) Optimized	(2) - (1)
Mode share (%)	4.61	3.23	(-30%)*
Trip fare rate ($/km)	0.35	1.48	(+326%)*
Rejection rate (%)	77.19	74.19	(-4%)*
Waiting time (min)	4.14	3.97	(-4%)
Vehicle occupancy (%)	5.51	3.45	(-37%)*
Profit ($/hr)	-923.7	-733.58	(+21%)*

- * means the value is with 5% significant level in the two-sample t-test.
- The trip fare rate under fixed trip fare is an approximate because all trips in this scenario are charged for $2.5.
- The computation of mode share includes passengers who choose DRT but get rejected.

9.7. Summary

In this chapter, we develop a ride-hail module in MATSim and embed an optimal pricing model into the simulator. The ride-hail module explicitly characterizes the market entry decision and cruising behaviors of drivers and the trip dispatching and pricing of the service operator, on top of the mode choice of passengers readily available in MATSim. The module is is also flexible enough to accommodate a variety of service modes, trip dispatching algorithms and vehicle cruising strategies. The embedded price optimizer, developed based on a theoretical surrogate of the ride-hail market, aims to maximize profit while anticipating the change in market equilibrium achieved through the simulation. A set of case studies are constructed based on a travel demand model of Wayne County, MI. Results from these studies are summarized as follows:

(1) The surrogate model correctly predicts the distribution of passenger waiting time obtained by the ride-hail simulator for different ride-hail service modes, including street-hail, e-hail and pooling.

(2) The pricing optimizer can recognize the price elasticity of market equilibrium through the surrogate model embedded in the simulator, following the idea of metamodel-based SBO. As a result, it is effective in discovering pricing strategies that help boost the operator's profitability in all test cases.

(3) The pricing optimizer can also be implemented into DRT, an independent module in MATSim, and show satisfactory performance in enhancing the service profitability. The exercise suggests the surrogate pricing model may be generalized to other mobility service and design contexts.

The simulation can be further improved in a number of directions. First, the current optimal pricing model only produces a uniform and static pricing strategy. The next step is to consider dynamic and spatial pricing, which requires modeling more realistic behaviors of passengers and drivers. Secondly, the ride-hail module is capable of accommodating other design problems, such as vacant vehicle rebalancing and trip dispatching. Thus, developing compatible surrogate model for these problems and integrating them into a simulation framework will be another focus of the future work. A future study could also consider a mixed-mode service, in which the operator uses the same fleet to serve different type of trips (e.g., UberX and UberPool). Finally, the methodology developed herein can also be used to analyze the broader impact of emerging ride-hail services on urban transport systems. Such applications are readily available since MATSim is equipped with a full-blown activity-based model and traffic simulation.

CHAPTER 10

Data and parameter estimation

10.1. Data description

Most numerical experiments in this disseration are constructed based on a hisotrical TNC dataset collected in the City of Chicago.[1] The driver data are provided in a separate table.[2] We select the trip records in September 2019 with both pickup and dropoff locations inside the city, which amounts 7.66 milion rows. Further, we select nine of 77 communities in Chicago as the sutdy area, which cover over 70% of total trips. In addition, we only consider trips starting between 6 AM–9 PM from Monday to Thursday, during which the demand patterns are relatively stable.

To estimate total travel demand, we also download the ridership data from Chicago Transit Authority.[3] Additionally, we query traffic speed data[4] during the same time period to support the analysis in Section 7.6.

10.2. Parameter estimation

A number of parameters are estimated using the data described in the previous section. On the demand side, the total passenger demand D_0 is estimated as the sum of ride-hail and transit ridership. The OD demand pattern is set according to the pattern revealed from the ride-hail trips. For the analysis of the aggregate model (i.e., Chapters 4, 5 and 6), the ride-hail trip durations τ_s and τ_p are directly estimated from the TNC data. Since the traffic speeds are endogenously

[1] Available at `https://data.cityofchicago.org/Transportation/Transportation-Network-Providers-Trips/m6dm-c72p`.

[2] Availabel at `https://data.cityofchicago.org/Transportation/Transportation-Network-Providers-Drivers/j6wf-834c`.

[3] Available at `https://www.transitchicago.com/ridership/`

[4] Available at `https://data.cityofchicago.org/Transportation/Chicago-Traffic-Tracker-Congestion-Estimates-by-Se/n4j6-wkkf`

determined in the two-node model (Chapter 7), we instead estimate the trip distance d_{ij}^s and d_{ij}^p between the identified CBD and peripheral area, as well as the additional travel time ε_s and ε_p. As for the network model (Chapter 8), the trip duration between each pair of zones is first estimated from the TNC data and then rounded to be multiples of Δ (the length of each time interval). To estimate the transit trip duration, we query the travel time estimate through Google MAP API between each pair of cencus zone and then computed the average weighted by the travel demand. On the supply side, the potential supply S_0 is estimated using the TNC driver data. Finally, the trip fare used in the numerical experiments are also estimated from the TNC data, weighted by the travel demand in the case of aggregate and two-node models.

The other key parameters used in the numerical experiments are estimated as follows:

- The transit fare f_t and relative disutility of transit ζ are set accordingly to Schwieterman [2019].
- The vale of time ν is estimated based on the value of business trips reported in US Bureau of Labor Statistics[5], adjusted to 2019 US dollar value.
- The matching efficiency k is taken from Zhang et al. [2019a], who calibrate the e-hail matching model using TNC data collected in Shenzhen, China.
- The pooling efficiency b is is set to be the ratio between the average matching time (taken as 15 s) and the average total wait time (take as 5 min).
- The congestion cost per vehicle c_0 is approximated based on Erhardt et al. [2019], which estimate the launch of TNC services in San Francisco has caused about 26,000 extra vehicle delayed hours (VDH) per day in 2016 compared to 2010. Castiglione et al. [2016] report that over 6500 TNC vehicles operate in San Francisco during peak hour on a typical weekday in 2016. If we assume the average number of TNC vehicles in operation is 5000 over the day, then roughly about 0.22 hr of VDH per hour can be attributed to each vehicle. According to Cookson and Pishue [2017], an average American driver lost 99

[5]Available at `https://www.transportation.gov/office-policy/transportation-policy/guidance-value-time`

hours to traffic congestion, translated to a monetary value of \$1377. Thus, we estimate the congestion cost per TNC vehicle is \$2.9/hr.

- The average reservation rate e_0 used in the analysis of monopoly market is set based on the mean hourly wage of for the occupational group "Transportation and material moving" in the Chicago area[6]. That of duopoly market is set to be slightly lower to adjust for the different supply models. In the two-node model, the minimum reservation rate is set to be the minimum wage rate in the U.S., while the maximum is set to have the mean value coincide with e_0 discussed above under uniform distribution.
- The free-flow speed v_f and jam density ρ_{jam} are taken from Mahmassani et al. [2013], who conduct dynamic traffic simulations on a Chicago downtown road network. Specifically, the latter is transformed from lane density to space density according to the road density estimated in Mohareb et al. [2016].
- The aggregate road capacity C_p is estimated according to Greenshield's model, i.e., $C_p = v_f \rho_{\text{jam}}/4$.
- The background traffic is back-calculated from the traffic model (Eqs. (7.2a) and (7.2b)) using the observed TNC demand flow and traffic speeds. Accordingly, the traffic speed without TNC vehicles $\boldsymbol{\tau}_0$ are derived from the calibrated traffic model by setting ride-hail vehicle flows as zero.

Finally, to determine the parameter κ in Eq. (3.18), we tested a range of values between 1 and 6. We found $\hat{A}(d,l)$ tends to overestimate (underestimate) $A(d,l)$ when $\kappa = 6$ (1). Also, a value between 2 and 4 delivers similar approximation quality. Importantly, within this range, the performance of the equilibrium model seems insensitive to the choice of κ. Based on the above tests, $\kappa = 4$ is finally selected in numerical experiments.

[6]Available at `https://www.bls.gov/regions/midwest/news-release/occupationalemploymentandwages_chicago.htm`.

CHAPTER 11

Conclusions

11.1. Research summary

This dissertation studies the pooling service in the context of a ride-hail market following a bottom-up approach. A physical matching model is first proposed to describe the interaction between passengers and vehicles in the matching process of a pooling trip (Chapter 3). This enables us to identify the main factors that affect the expected passenger wait time for pooling and the key differences from other ride-hail modes. The base model is further extended to account for the matching in a market with multiple platforms (Chapter 4) and multiple zones (Chapter 7). It is also used to derive the likelihood that an idle driver successfully pick up a passenger after searching in a zone for a certain time period, which plays a critical role in drivers' search decision (Chapter 8).

The main insights drawn from the matching model are summarized as follows. Different from traditional street-hail taxis, the passenger wait time for regular e-hail trips is subject to passenger competition. Since passengers could be virtually matched with any vehicle at a distance, they are competing for the same pool of vacant vehicles. Then, pooling essentially mitigates such a competition because these passengers now require fewer vehicles. This phenomenon is captured by the effective waiting passenger density defined in the matching. In an extreme case, if all passengers are will to share their trips with another passenger, the passenger competition would be cut by half. Accordingly, passengers choosing to ride alone (i.e., solo passengers) also benefit from the reduced passenger competition.

In this dissertation, we assume each pooling trip is shared by two passengers. Then, the pooling wait time can be divided into two parts. The first is the pickup time for the passenger closer to the matched vehicle and the second is the detour to pick up the other passenger. It is shown that

the pooling passengers' competing power for vehicles is expanded. Suppose the vehicle is matched to the closest waiting passenger. Then, a pooling passenger could be matched with a vehicle very far away as long as their pooling peer is close enough to the vehicle. Such an advantage reflects in the first part of the pooling wait time and increases with the distance between the two pooling passengers. However, a larger distance between the two passengers also results in a longer pickup detour. This leads to a unique trade-off in pooling service and makes the pooling wait time not monotonically increase with the waiting passenger density, which is observed in regular e-hail and solo trips.

Another interesting finding from the matching model regards the different supply modes in a multi-platform market. In this dissertation, we study two supply modes, i.e., single-homing and multi-homing. Under single-homing, each driver only joins one platform, whereas, under multi-homing, drivers join all platforms if they choose to enter the market. Intuitively, multi-homing should help improve the level-of-service (LOS) because it expands the vehicle supply. However, the matching model tells a different story. Although passengers get access to a large pool of vehicles, they are also facing more intense competition against passengers from other platforms. As a result, multi-homing does not benefit the LOS at all.

The matching model serves as an important building block of the analysis of market equilibrium. In particular, it dictates the passenger wait time that links the vehicle supply (in terms of vacant vehicle time) to the passenger demand. Chapter 4 establishes the equilibrium model for an aggregate ride-hail market with either single or multiple platforms. The model assumes passengers choose among different ride-hail services and transit based on their generalized cost, which includes trip fare, wait time and in-vehicle travel time. On the supply side, drivers decide whether to joint market and which platform to join in the case of single-homing based on the average earning rate. Collectively, the equilibrium is formulated as a system of equations and shown to be equivalent to a fixed point. Accordingly, the equilibrium is proved to always exist under mild conditions and can be easily solved through fixed point iterations. In Chapter 7, the equilibrium model is further extended

to study a stylized two-node market. To characterize the congestion effect, a simple traffic model is proposed so that the traffic speed is endogenously determined by the ride-hail service operations. Again, the equilibrium can be reduced to a fixed point system and easily satisfies the conditions for existence.

The equilibrium models developed for both aggregate market (Chapter 4) and spatial market (Chapter 7) build a foundation for the upper-level research on the operations management of ride-hail service. This dissertation has been focused on the platform's optimal pricing strategies and the impact of regulations. Chapter 5 initiates the discussion by assuming a single ride-hail platform monopolizes the market. In particular, a gradient-based algorithm is developed to solve the optimal pricing problem with complex and nonconvex equilibrium constraints. Our analysis of the optimum conditions of the pricing problem reveals that serving a mixture of solo and pooling trips could enhance the platform's market power and thus boost its profit. This finding is confirmed by the numerical results. Besides, a few interesting findings are also derived from the numerical experiments. Although the minimum wage has been advocated in both academia and industry, we found it only improves social welfare in the short-term by largely sacrificing the platform profit. In a long run, the platform would tend to limit its driver pool. Consequently, the policy may reduce social welfare to even below the unregulated state. Furthermore, the policy pushes the platform to serve more solo trips and discourage pooling because solo trips more effectively consume the extra vehicle capacity induced by the higher wage rate. Another policy tested in this chapter is a congestion tax on each solo trip. It is found the policy considerably encourages pooling but fails to improve social welfare. However, a joint policy of minimum wage and congestion tax could improve both the pooling ratio and social welfare.

Chapter 6 continues to investigate the inter-platform competition. We consider two platforms competing for both passengers and vehicles by setting their trip fare and compensation rate. The pricing game ends up with a duopoly equilibrium characterized by a Nash equilibrium (NE) in the unregulated case and a generalized Nash equilibrium (GNE) in the regulated one. We show the

existence of these duopoly equilibrium states and propose algorithms to solve them. The difference between single-homing and multi-homing is then discussed in the context of a pricing game. It is found multi-homing is detrimental to the market because platforms would be less incentive to increase trip fare or compensation rate to improve LOS. This phenomenon is first shown analytically and then demonstrated in numerical experiments. Consequently, multi-homing leads to a market failure, where passengers, drivers and platforms are all worse off. In this case, the minimum wage policy becomes a remedy. By maintaining a sufficient vehicle supply, the policy prevents the LOS from being deteriorated by the pricing game. In the numerical experiments, we also discuss the case of asymmetric platforms. Interestingly, we found a platform dedicated to pooling service is more likely to survive in a multi-homing market. Since multi-homing drivers make market entry decisions based on the average earning rate of the entire market rather than that of a single platform, the platform only serving pooling trips can operate at a fairly low cost. Particularly, the numerical results show such a platform could achieve a comparable profit as that of a platform serving both solo and pooling trips, even though its market share is much lower than the other.

Chapter 7 is devoted to examining the congestion effect of ride-hail vehicles and evaluating different congestion mitigation policies targeted at ride-hail service operations. Specifically, three policies are considered: (i) a trip-based fee charged on each solo trip starting or ending in CBD, (ii) a cordon-based fee charged on each vehicle entering CBD with one or no passenger, and (iii) a cruising cap that requires a minimum occupancy rate for all ride-hail vehicles in CBD. The numerical results show the trip-based fee delivers the best performance. By effectively promoting pooling, it manages to reduce traffic congestion without affecting the LOS. The cordon-based fee demonstrates similar behaviors but trails behind in nearly all metrics. In contrast, the cruising cap achieves the largest congestion relief at a great cost to passengers, drivers and the platform. To maintain a high vehicle occupancy in CBD, it even pushes the platform to serve more solo trips in CBD, which substantially worsens the traffic there.

As the analysis of ride-hail service moves from a static aggregate market to a dynamic spatial market, one key challenge is to model the vehicle movements driven by their own interest. Chapter 8 tackles this vehicle routing problem from a game-theoretic perspective. In brief, the routing decisions of individual drivers lead to vehicle supply in each zone, and, in turn, the demand-supply relationship affects the expected payoff of each driver. Therefore, each driver has to choose strategies based on the behaviors of all other drivers. We model individual driver's routing behavior as a Markov decision process (MDP). Accordingly, the collective routing behaviors are formulated as an MDP congestion game and Wardrop equilibrium is introduced to describe the routing strategy at equilibrium. We show that the equilibrium is equivalent to a fixed point and prove its existence via the fixed point theorem. The model is further extended to account for the cooperative routing, where drivers collaborate with each other to maximize the total reward. It is proved that the optimal routing strategy can be derived by solving the Wardrop equilibrium of a non-cooperative game with an adjusted reward function. It thus implies that vehicle coordination can be done in a distributed manner, which is expected to save a lot of communication and computation costs. To solve the equilibrium, an iterative algorithm is developed where the routing strategy is updated in each iteration according to the successive average manner.

A key component of the routing model is the meeting probability, which dictates the likelihood of picking up a passenger after search in a zone for a time period. We specify its functional forms for street-hail and e-hail and then fit them with simulation data. The regression results indicate that the analytical model correctly captures the main factors of the meeting probability. Two sets of numerical experiments are then constructed using the calibrated meeting probability functions. One is conducted on a hypothetical network to demonstrate the solution algorithm and to compare the routing strategy in different service modes. The other is based on a real market to evaluate the system performance and the loss of efficiency due to selfish routing in real practice. It is found the cooperative routing does not deliver a significant improvement in system efficiency. However, it does outperform non-cooperative routing when real demand information is not available to the

drivers. This finding is implied from a counterfactual study of special events that induce short-term demand peaks outside the high-demand area and during an off-peak period.

Although most findings presented in this dissertation are derived from analytical models that rely on strong assumptions. Chapter 9 provides an option to validate these models and further embed them into a simulation-based service design tool. Specifically, an agent-based simulation of ride-hail service is developed based on MATSim, a widely used transportation simulation framework. The simulation supports a number of features that are not available in currently open-sourced simulation tools. Importantly, the simulation provides an interface to have some control variables adjusted over the simulation process. A pricing module is then developed using this interface to update the pricing strategy based on the pricing model developed in Chapter 5. The performance of this pricing module is demonstrated through numerical experiments.

11.2. Future directions

This dissertation gives one of the first attempts to explicitly model the matching process of pooling and integrate it into the economic analysis of ride-hail market. Hence, this work could be extended in a number of directions.

First and foremost, the matching model is definitely not perfect yet. It is worthwhile to examine the physical meaning of parameters defined in the model (e.g., the matching efficiency k and b) and connect them to the detailed matching process in real practice. Besides, it is also interesting to adjust the model to accommodate special market conditions, such as WGC. Given the difference in the matching mechanism, it is also worth examining the economies of scale in the matching of pooling.

Another important direction is to complete the network equilibrium model. The vehicle routing model presented in this dissertation only considers the supply side. To study the operational problems in a spatial market (e.g., spatial pricing), we need to complement the current model with a demand model. Accordingly, the definition of equilibrium as well as the solution algorithm needs

to be updated. Besides, the current model is only specified for street-hail and e-hail. An immediate next step is thus to the meeting probability in a market serving both solo and pooling trips.

In this dissertation, ride-hail and transit are considered as alternatives of each other. However, a number of studies have proposed an integrated mobility system, where the fixed-route transit system takes in charge of the consolidated travel demand and the ride-hail platforms provide first-/last-mile trips. Despite its prominent capability of reducing VMT and increasing transit ridership, a number of questions arise from the different stakeholders' perspectives. For instance, why would the ride-hail platforms collaborate with the transit agency? Why would passengers choose multi-modal trips? Would there still be sufficient vehicle supply given that a large fraction of demand is directed to transit? These are all interesting research problems. After all, pooling is expected to have greater potential in such a multi-modal mobility system because it becomes easier to find someone to share part of your trip.

Last but not least, this dissertation provides a comprehensive framework to study an array of service operation and regulation problems. On the demand side, we could study spatial and temporal surge pricing. On the supply, one interesting research topic is the relocation incentives. From the perspective of a regulator, a number of policies could be explored, including, but not limited to, trip-based versus time-based congestion fee, commission cap versus minimum wage, displaying trip destination to drivers prior to they accepting the order. It is also interesting to develop practical tools based on the analytical models proposed in this work to support the decision-making process in real practice.

Bibliography

Muhammad Adnan, Francisco C Pereira, Carlos Miguel Lima Azevedo, Kakali Basak, Milan Lovric, Sebastián Raveau, Yi Zhu, Joseph Ferreira, Christopher Zegras, and Moshe Ben-Akiva. Simmobility: A multi-scale integrated agent-based simulation platform. In *95th Annual Meeting of the Transportation Research Board Forthcoming in Transportation Research Record*, 2016.

Philipp Afeche, Zhe Liu, and Costis Maglaras. Ride-hailing networks with strategic drivers: The impact of platform control capabilities on performance. *Columbia Business School Research Paper*, (18-19):18–19, 2018.

Fatima Afifah and Zhaomiao Guo. Spatial pricing of ride-sourcing services in a congestedtransportation network. Available at arXiv:2006.00164 (Accessed: 2020-07-07), 2020.

Charalambos Aliprantis and Kim C. Border. *Infinite Dimensional Analysis: A Hitchhiker's Guide*. Springer, 2006.

Javier Alonso-Mora, Samitha Samaranayake, Alex Wallar, Emilio Frazzoli, and Daniela Rus. On-demand high-capacity ride-sharing via dynamic trip-vehicle assignment. *Proceedings of the National Academy of Sciences*, 114(3):462–467, 2017.

Chokri Aloui and Khaïreddine Jebsi. Optimal Pricing of a Duopoly Platform with Two-Sided Congestion Effect. *Journal of Research in Industrial Organization*, 2011:1–10, 2011.

Joshua D Angrist, Sydnee Caldwell, and Jonathan V Hall. Uber vs. taxi: A driver's eye view. Technical report, National Bureau of Economic Research, 2017.

Mark Armstrong. Competition in two-sided markets. *The RAND Journal of Economics*, 37(3): 668–691, 2006.

Richard Arnott. Taxi travel should be subsidized. *Journal of Urban Economics*, 40(3):316–333, 1996.

Kenneth J Arrow and Gerard Debreu. Existence of an equilibrium for a competitive economy. *Econometrica: Journal of the Econometric Society*, pages 265–290, 1954.

Arash Asadpour, Ilan Lobel, and Garrett van Ryzin. Minimum earnings regulation and the stability of marketplaces. Available at SSRN 3502607 (Accessed: 2020-10-25), 2019.

Mohammad Asghari and Cyrus Shahabi. Adapt-pricing: a dynamic and predictive technique for pricing to maximize revenue in ridesharing platforms. In *Proceedings of the 26th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems*, pages 189–198. ACM, 2018.

Eduardo M Azevedo and E Glen Weyl. Matching markets in the digital age. *Science*, 352(6289): 1056–1057, 2016.

Jiaru Bai and Christopher S Tang. Can two competing on-demand service platforms be both profitable? Available at SSRN 3282395 (Accessed: 2019-10-23), 2018.

Siddhartha Banerjee, Carlos Riquelme, and Ramesh Johari. Pricing in ride-share platforms: A queueing-theoretic approach. Available at: SSRN 2568258 (Accessed: 2018-11-12), 2015.

Siddhartha Banerjee, Daniel Freund, and Thodoris Lykouris. Pricing and optimization in shared vehicle systems: An approximation framework. 2017.

Russell R Barton and Martin Meckesheimer. Metamodel-based simulation optimization. In *Handbooks in operations research and management science*, chapter 18, pages 535–574. Elsevier, 2006.

Aleix Bassolas, José J Ramasco, Ricardo Herranz, and Oliva G Cantú-Ros. Mobile phone records to feed activity-based travel demand models: Matsim for studying a cordon toll policy in barcelona. *Transportation Research Part A: Policy and Practice*, 121:56–74, 2019.

Atılım Günes Baydin, Barak A Pearlmutter, Alexey Andreyevich Radul, and Jeffrey Mark Siskind. Automatic differentiation in machine learning: a survey. *The Journal of Machine Learning Research*, 18(1):5595–5637, 2017.

Martin Beckmann, Charles B McGuire, and Christopher B Winsten. Studies in the economics of transportation. Technical report, 1956.

Michael E Beesley and Steven Glaister. Information for regulating: the case of taxis. *The Economic Journal*, 93(371):594–615, 1983.

Paul Belleflamme and Eric Toulemonde. Negative intra-group externalities in two-sided markets. *International Economic Review*, 50(1):245–272, 2009.

M. Ben-Akiva and S. R. Lerman. *Discrete Choice Analysis: Theory and Application to Travel Demand*. Cambridge, MA: MIT Press, 1985.

Caio Vitor Beojone and Nikolas Geroliminis. On the inefficiency of ride-sourcing services towards urban congestion. *Transportation research part C: emerging technologies*, 124:102890, 2021.

Fernando Bernstein, Gregory DeCroix, and N Bora Keskin. Competition between two-sided platforms under demand and supply congestion effects, 2019. Available at SSRN 3250224 (Accessed: 2020-03-05).

Omar Besbes, Francisco Castro, and Ilan Lobel. Spatial capacity planning. 2018. Available at SSRN 3292651 (Accessed: 2019-10-23).

Omar Besbes, Francisco Castro, and Ilan Lobel. Surge pricing and its spatial supply response. *Management Science*, 67(3):1350–1367, 2021.

Kostas Bimpikis, Ozan Candogan, and Daniela Saban. Spatial pricing in ride-sharing networks. *Operations Research*, 2019.

Julius R Blum. Multidimensional stochastic approximation methods. *The Annals of Mathematical Statistics*, pages 737–744, 1954.

Enrico Böhme, Christopher Müller, and Johann Wolfgang Goethe-university. Comparing monopoly and duopoly on a two-sided market without product differentiation comparing. Available at urlhttps://mpra.ub.uni-muenchen.de/26938/ (Accessed: 2019-11-11), 2010.

Francis P Boscoe, Kevin A Henry, and Michael S Zdeb. A nationwide comparison of driving distance versus straight-line distance to hospitals. *The Professional Geographer*, 64(2):188–196, 2012.

Dietrich Braess, Anna Nagurney, and Tina Wakolbinger. On a paradox of traffic planning. *Transportation science*, 39(4):446–450, 2005.

Anton Braverman, Jim G Dai, Xin Liu, and Lei Ying. Empty-car routing in ridesharing systems. *Operations Research*, 2019.

L E J. Brouwer. Über abbildung von mannigfaltigkeiten. *Mathematische Annalen*, 71(1):97–115, 1911.

Nicholas Buchholz. Spatial equilibrium, search frictions and efficient regulation in the taxi industry. Available at `https://scholar.princeton.edu/sites/default/files/nbuchholz/files/taxi_draft.pdf` (Accessed: 2020-03-04), 2019.

Gerard P Cachon, Kaitlin M Daniels, and Ruben Lobel. The role of surge pricing on a service platform with self-scheduling capacity. *Manufacturing & Service Operations Management*, 19(3): 368–384, 2017.

Robert D Cairns and Catherine Liston-Heyes. Competition and regulation in the taxi industry. *Journal of Public Economics*, 59(1):1–15, 1996.

Dan Calderone and S Shankar Sastry. Markov decision process routing games. In *Proceedings of the 8th International Conference on Cyber-Physical Systems*, pages 273–279. ACM, 2017.

Dan Calderone and S Shankar. Infinite-horizon average-cost markov decision process routing games. In *2017 IEEE 20th International Conference on Intelligent Transportation Systems (ITSC)*, pages 1–6. IEEE, 2017.

Joe Castiglione, Tilly Chang, Drew Cooper, Jeff Hobson, Warren Logan, Eric Young, Billy Charlton, Christo Wilson, Alan Mislove, Le Chen, et al. TNCs today: a profile of San Francisco transportation network company activity. Technical report, San Francisco County Transportation Authority, 2016.

Juan Castillo, Daniel T Knoepfle, and E Glen Weyl. Surge pricing solves the wild goose chase. Available at SSRN 2890666 (Accessed: 2018-5-3), 2018.

Nelson D Chan and Susan A Shaheen. Ridesharing in north america: Past, present, and future. *Transport Reviews*, 32(1):93–112, 2012.

Chuqiao Chen, Fugen Yao, Dong Mo, Jiangtao Zhu, and Xiqun Michael Chen. Spatial-temporal pricing for ride-sourcing platform with reinforcement learning. *Transportation Research Part C: Emerging Technologies*, 130:103272, 2021.

Hongyu Chen, Kenan Zhang, Xiaobo Liu, and Yu Marco Nie. A physical model of street ride-hail. Available at SSRN 3318557 (Accessed: 2019-1-18), 2018.

M Keith Chen and Michael Sheldon. Dynamic pricing in a labor market: Surge pricing and flexible work on the uber platform. *Working Paper, University of California, Los Angeles*, 2017.

Xiqun Michael Chen, Chenfeng Xiong, Xiang He, Zheng Zhu, and Lei Zhang. Time-of-day vehicle mileage fees for congestion mitigation and revenue generation: A simulation-based optimization method and its real-world application. *Transportation Research Part C: Emerging Technologies*, 63:71–95, 2016.

Xiqun Michael Chen, Hongyu Zheng, Jintao Ke, and Hai Yang. Dynamic optimization strategies for on-demand ride services platform: Surge pricing, commission rate, and incentives. *Transportation Research Part B: Methodological*, 138:23–45, 2020.

Qixiu Cheng, Shuaian Wang, Zhiyuan Liu, and Yu Yuan. Surrogate-based simulation optimization approach for day-to-day dynamics model calibration with real data. *Transportation Research Part C: Emerging Technologies*, 105:422–438, 2019.

Linsen Chong and Carolina Osorio. A simulation-based optimization algorithm for dynamic large-scale urban transportation problems. *Transportation Science*, 52(3):637–656, 2018.

Glenn Cich, Luk Knapen, Michał Maciejewski, Tom Bellemans, Davy Janssens, et al. Modeling demand responsive transport using sarl and matsim. *Procedia Computer Science*, 109:1074–1079, 2017.

Blerim Cici, Athina Markopoulou, Enrique Frias-Martinez, and Nikolaos Laoutaris. Assessing the potential of ride-sharing using mobile and social data: a tale of four cities. In *Proceedings of*

the 2014 ACM International Joint Conference on Pervasive and Ubiquitous Computing, pages 201–211, 2014.

Charles W Cobb and Paul H Douglas. A theory of production. *American Economic Review*, 18: 139–165, 1928.

Graham Cookson and Bob Pishue. Global traffic scorecard. Technical report, INRIX, 2017.

Judd Cramer and Alan B Krueger. Disruptive change in the taxi business: The case of uber. *American Economic Review*, 106(5):177–82, 2016.

Gerard De Jong, Andrew Daly, Marits Pieters, and Toon Van der Hoorn. The logsum as an evaluation measure: Review of the literature and new results. *Transportation Research Part A: Policy and Practice*, 41(9):874–889, 2007.

André de Palma and Robin Lindsey. Traffic congestion pricing methodologies and technologies. *Transportation Research Part C: Emerging Technologies*, 19(6):1377–1399, 2011.

Arthur S De Vany. Capacity utilization under alternative regulatory restraints: an analysis of taxi markets. *The Journal of Political Economy*, pages 83–94, 1975.

Xuan Di and Xuegang Jeff Ban. A unified equilibrium framework of new shared mobility systems. *Transportation Research Part B: Methodological*, 129:50–78, 2019.

Mi Diao, Hui Kong, and Jinhua Zhao. Impacts of transportation network companies on urban mobility. *Nature Sustainability*, pages 1–7, 2021.

Shadi Djavadian and Joseph YJ Chow. An agent-based day-to-day adjustment process for modeling 'mobility as a service'with a two-sided flexible transport market. *Transportation research part B: methodological*, 104:36–57, 2017.

Tingting Dong, Zhengtian Xu, Qi Luo, Yafeng Yin, Jian Wang, and Jieping Ye. Optimal contract design for ride-sourcing services under dual sourcing. *Transportation Research Part B: Methodological*, 146:289–313, 2021.

George W Douglas. Price regulation and optimal service standards: The taxicab industry. *Journal of Transport Economics and Policy*, pages 116–127, 1972.

Stephen Edelstein. Lyft will sue New York City to block driver minimum wage law, 2019a. Available at https://www.thedrive.com/news/26247/lyft-will-sue-new-york-city-to-block-driver-minimum-wage-law (Accessed: 2019-02-03).

Stephen Edelstein. Uber and Lyft have officially stopped hiring drivers in NYC, 2019b. Available at https://www.thedrive.com/news/27756/uber-and-lyft-have-officially-stopped-hiring-drivers-in-nyc-will-resume-in-2020 (Accessed: 2019-05-02).

Gregory D Erhardt, Sneha Roy, Drew Cooper, Bhargava Sana, Mei Chen, and Joe Castiglione. Do transportation network companies decrease or increase congestion? *Science advances*, 5(5): eaau2670, 2019.

Alvaro Estandia, Maximilian Schiffer, Federico Rossi, Justin Luke, Emre Can Kara, Ram Rajagopal, and Marco Pavone. On the interaction between autonomous mobility on demand systems and power distribution networks—an optimal power flow approach. *IEEE Transactions on Control of Network Systems*, 2021.

Francisco Facchinei and Christian Kanzow. Penalty methods for the solution of generalized nash equilibrium problems. *SIAM Journal on Optimization*, 20(5):2228–2253, 2010.

DB Fairthorne. The distance between pairs of points in towns of simple geometric shape. In *Proceedings 2nd International Symposium on the Theory of Traffic Flow, Paris, 1964*. OECD, 1964.

Guiyun Feng, Guangwen Kong, and Zizhuo Wang. We are on the way: Analysis of on-demand ride-hailing systems. *Manufacturing & Service Operations Management*, 2020.

Guillaume R Frechette, Alessandro Lizzeri, and Tobias Salz. Frictions in a competitive, regulated market: Evidence from taxis. *American Economic Review*, 109(8):2954–92, 2019.

Masabumi Furuhata, Maged Dessouky, Fernando Ordóñez, Marc-Etienne Brunet, Xiaoqing Wang, and Sven Koenig. Ridesharing: The state-of-the-art and future directions. *Transportation Research Part B: Methodological*, 57:28–46, 2013.

Nikhil Garg and Hamid Nazerzadeh. Driver surge pricing. *Management Science*, 2021.

Yong Ge, Hui Xiong, Alexander Tuzhilin, Keli Xiao, Marco Gruteser, and Michael Pazzani. An energy-efficient mobile recommender system. In *Proceedings of the 16th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 899–908, 2010.

Nikolas Geroliminis and Carlos F Daganzo. Existence of urban-scale macroscopic fundamental diagrams: Some experimental findings. *Transportation Research Part B: Methodological*, 42(9): 759–770, 2008.

BD Greenshields, JR Bibbins, WS Channing, and HH Miller. A study of traffic capacity. In *Highway research board proceedings*, volume 1935. National Research Council (USA), Highway Research Board, 1935.

Harish Guda and Upender Subramanian. Your uber is arriving: Managing on-demand workers through surge pricing, forecast communication, and worker incentives. *Management Science*, 65 (5):1995–2014, 2019.

Itai Gurvich, Martin Lariviere, and Antonio Moreno. Operations in the on-demand economy: Staffing services with self-scheduling capacity. In *Sharing Economy*, pages 249–278. Springer, 2019.

Clayton Guse. Judge throws out NYC's strict regulations on Uber and Lyft, calls them 'arbitrary', 2019. Available at `https://www.nydailynews.com/new-york/ny-judge-lawsuit-uber-lyft-cruising-cap-20191223-q4f4g3xat5gdlkdtm5zlezbtbe-story.html` (Accessed: 2020-01-04.

Jonathan V Hall and Alan B Krueger. An analysis of the labor market for uber's driver-partners in the united states. *ILR Review*, 71(3):705–732, 2018.

Patrick T Harker. Generalized nash games and quasi-variational inequalities. *European journal of Operational research*, 54(1):81–94, 1991.

Patrick T Harker and Jong-Shi Pang. Existence of optimal solutions to mathematical programs with equilibrium constraints. *Operations research letters*, 7(2):61–64, 1988.

Patrick T Harker and Jong-Shi Pang. Finite-dimensional variational inequality and nonlinear complementarity problems: a survey of theory, algorithms and applications. *Mathematical programming*, 48(1-3):161–220, 1990.

Fang He and Zuo-Jun Max Shen. Modeling taxi services with smartphone-based e-hailing applications. *Transportation Research Part C: Emerging Technologies*, 58:93–106, 2015.

Fang He, Xiaolei Wang, Xi Lin, and Xindi Tang. Pricing and penalty/compensation strategies of a taxi-hailing platform. *Transportation Research Part C: Emerging Technologies*, 86:263–279, 2018.

Xiang He, Xiqun Chen, Chenfeng Xiong, Zheng Zhu, and Lei Zhang. Optimal time-varying pricing for toll roads under multiple objectives: a simulation-based optimization approach. *Transportation Science*, 51(2):412–426, 2017.

John Holler, Risto Vuorio, Zhiwei Qin, Xiaocheng Tang, Yan Jiao, Tiancheng Jin, Satinder Singh, Chenxi Wang, and Jieping Ye. Deep reinforcement learning for multi-driver vehicle dispatching and repositioning problem. In *2019 IEEE International Conference on Data Mining (ICDM)*, pages 1090–1095. IEEE, 2019.

Sebastian Hörl, Claudio Ruch, Felix Becker, Emilio Frazzoli, and Kay W Axhausen. Fleet operational policies for automated mobility: A simulation assessment for zurich. *Transportation Research Part C: Emerging Technologies*, 102:20–31, 2019.

Andreas Horni, Darren M Scott, Michael Balmer, and Kay W Axhausen. Location choice modeling for shopping and leisure activities with matsim: combining microsimulation and time geography. *Transportation Research Record*, 2135(1):87–95, 2009.

Andreas Horni, Kai Nagel, and Kay W Axhausen. *The multi-agent transport simulation MATSim*. Ubiquity Press, 2016.

Joel L Horowitz. The stability of stochastic equilibrium in a two-link transportation network. *Transportation Research Part B: Methodological*, 18(1):13–28, 1984.

Bin Hu, Ming Hu, and Han Zhu. Surge pricing and two-sided temporal responses in ride-hailing. Available at SSRN 3278023 (Accessed: 2019-08-04), 2018.

Ming Hu and Yun Zhou. Dynamic type matching. *Manufacturing & Service Operations Management*, 2021.

Ren-Hung Hwang, Yu-Ling Hsueh, and Yu-Ting Chen. An effective taxi recommender system based on a spatio-temporal factor analysis model. *Information Sciences*, 314:28–40, 2015.

Michael F Hyland and Hani S Mahmassani. Sharing is caring: Dynamic autonomous vehicle fleet operations under demand surges. In *97th Annual Meeting of the Transportation Research Board*, 2018.

Tatsuro Ichiishi. *Game Theory for Economic Analysis*. Academic Press, 1983.

Ramon Iglesias, Federico Rossi, Rick Zhang, and Marco Pavone. A bcmp network approach to modeling and controlling autonomous mobility-on-demand systems. *The International Journal of Robotics Research*, 38(2-3):357–374, 2019.

Jagan Jacob and Ricky Roet-Green. Ride solo or pool: Designing price-service menus for a ride-sharing platform. *European Journal of Operational Research*, 2021.

Meera Joshi, Nicholas Cowan, Olivia Limone, Kelly McGuinness, and Rohan Rao. E-hail regulation in global cities. Technical report, Rudin Center for Transportation Policy and Management, NYU Wagner School of Public Service, 2019.

Grace Orowo Kagho and David Hensle. Demand responsive transit simulation of wayne county, mi. In *100th Annual Meeting of the Transportation Research Board (TRB 2021)*, 2021.

Jintao Ke, Hai Yang, Xinwei Li, Hai Wang, and Jieping Ye. Pricing and equilibrium in on-demand ride-pooling markets. *Transportation Research Part B: Methodological*, 139:411–431, 2020a.

Jintao Ke, Hai Yang, and Zhengfei Zheng. On ride-pooling and traffic congestion. *Transportation Research Part B: Methodological*, 142:213–231, 2020b.

Jintao Ke, Zhengfei Zheng, Hai Yang, and Jieping Ye. Data-driven analysis on matching probability, routing distance and detour distance in ride-pooling services. *Transportation Research Part C:*

Emerging Technologies, 124:102922, 2021.

Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. Available at arXiv:1412.6980 (Accessed: 2020-10-26), 2014.

Serbjeet Kohli and Andrew Daly. The use of logsums in welfare estimation: application in prism. In *Proceedings of the European Transport Conference, Strasbourg*, 2006.

SG Krantz and HR Parks. *The implicit function theorem: history, theory and applications*. Springer Science and Business Media, 2012.

Ricardo Lagos. An alternative approach to search frictions. *Journal of Political Economy*, 108(5): 851–873, 2000.

Ricardo Lagos. An analysis of the market for taxicab rides in new york city. *International Economic Review*, 44(2):423–434, 2003.

Richard C Larson and Amedeo R Odoni. *Urban operations research*. Number Monograph. 1981.

Jean-Michel Lasry and Pierre-Louis Lions. Mean field games. *Japanese journal of mathematics*, 2 (1):229–260, 2007.

Chao Lei, Zhoutong Jiang, and Yanfeng Ouyang. Path-based dynamic pricing for vehicle allocation in ridesharing systems with fully compliant drivers. *Transportation Research Part B: Methodological*, 2019.

A. P. Lerner. The concept of monopoly and the measurement of monopoly power. *The Review of Economic Studies*, 1(3):157–175, 1934.

Bin Li, Daqing Zhang, Lin Sun, Chao Chen, Shijian Li, Guande Qi, and Qiang Yang. Hunting or waiting? discovering passenger-finding strategies from a large-scale real-world taxi dataset. In *2011 IEEE International Conference on Pervasive Computing and Communications Workshops (PERCOM Workshops)*, pages 63–68. IEEE, 2011.

Sarah HQ Li, Daniel Calderone, Lillian Ratliff, and Behçet Açıkmeşe. Sensitivity analysis for markov decision process congestion games. In *2019 IEEE 58th Conference on Decision and Control (CDC)*, pages 1301–1306. IEEE, 2019a.

Sarah HQ Li, Yue Yu, Daniel Calderone, Lillian Ratliff, and Behçet Açrkmeşe. Tolling for constraint satisfaction in markov decision process congestion games. In *2019 American Control Conference (ACC)*, pages 1238–1243. IEEE, 2019b.

Sen Li, Hai Yang, Kameshwar Poolla, and Pravin Varaiya. Sptial pricing in ride-sourcing markets under a congestion charge. Availabel at arXiv:2010.09260 (Accessed 2020-10-26), 2020.

Sen Li, Kameshwar Poolla, and Pravin Varaiya. Impact of congestion charge and minimum wage on tncs: A case study for san francisco. *Transportation Research Part A: Policy and Practice*, 148:237–261, 2021.

Shen Li, Hamidreza Tavafoghi, Kamewhar Poolla, and Pravin Varaiya. Regulating tncs: Should uber and lyft set their own rules? *Transportation Research Part B: Methodological*, 129, 2019c.

Kaixiang Lin, Renyu Zhao, Zhe Xu, and Jiayu Zhou. 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, 2018.

John DC Little. A proof for the queuing formula: L= λ w. *Operations research*, 9(3):383–387, 1961.

Liang Liu, Clio Andris, and Carlo Ratti. Uncovering cabdrivers' behavior patterns from their digital traces. *Computers, Environment and Urban Systems*, 34(6):541–548, 2010.

Ilan Lobel and Sebastien Martin. Detours in shared rides. Available at SSRN 3711072 (Accessed: 2020-11-11), 2020.

Hani S Mahmassani, Meead Saberi, and Ali Zockaie. Urban network gridlock: Theory, characteristics, and dynamics. *Procedia-Social and Behavioral Sciences*, 80:79–98, 2013.

John Miller, Yu Nie, and Xiaobo Liu. Hyperpath truck routing in an online freight exchange platform. *Transportation Science*, 54(6):1676–1696, 2020.

Eugene Mohareb, Sybil Derrible, and Farideddin Peiravian. Intersections of jane jacobs' conditions for diversity and low-carbon urban systems: A look at four global cities. *Journal of Urban Planning and Development*, 142(2):05015004, 2016.

Bat-hen Nahmias-Biran, Jimi B Oke, Nishant Kumar, Kakali Basak, Andrea Araldo, Ravi Seshadri, Arun Akkinepally, Carlos Lima Azevedo, and Moshe Ben-Akiva. From traditional to automated mobility on demand: a comprehensive framework for modeling on-demand services in simmobility. *Transportation Research Record*, 2673(12):15–29, 2019.

Linglin Ni, Chuqiao Chen, Xiaokun Cara Wang, and Xiqun Michael Chen. Modeling network equilibrium of competitive ride-sourcing market with heterogeneous transportation network companies. *Transportation Research Part C: Emerging Technologies*, 130:103277, 2021.

Yu Marco Nie. How can the taxi industry survive the tide of ridesourcing? evidence from shenzhen, china. *Transportation Research Part C: Emerging Technologies*, 79:242–256, 2017.

Afshin Nikzad. Thickness and competition in ride-sharing markets. Available at SSRN 3065672 (Accessed: 2019-10-15), 2017.

Mehdi Nourinejad and Mohsen Ramezani. Ride-sourcing modeling and pricing in non-equilibrium two-sided markets. *Transportation Research Part B: Methodological*, 2019.

Mehdi Nourinejad and Matthew J Roorda. Agent based model for dynamic ridesharing. *Transportation Research Part C: Emerging Technologies*, 64:117–132, 2016.

NYC Taxi and Limousine Commission. Improving efficiency and managing growth in new york's for-hire vehicle sector. Technical report, 2019.

Carolina Osorio and Michel Bierlaire. A simulation-based optimization framework for urban transportation problems. *Operations Research*, 61(6):1333–1345, 2013.

Elinor Ostrom, Roy Gardner, James Walker, James M Walker, and Jimmy Walker. *Rules, games, and common-pool resources*. University of Michigan Press, 1994.

Erhun Özkan and Amy R Ward. Dynamic matching for real-time ride sharing. *Stochastic Systems*, 2020.

James A. Parrott and Michael Reich. An earning standard for new york city app-based drivers: Economic analysis and policy assessment, 2018.

Michael Patriksson. Sensitivity analysis of traffic equilibria. *Transportation Science*, 38(3):258–281, 2004.

Marco Pavone, Stephen L Smith, Emilio Frazzoli, and Daniela Rus. Robotic load balancing for mobility-on-demand systems. *The International Journal of Robotics Research*, 31(7):839–854, 2012.

Helen KRF Pinto, Michael F Hyland, Hani S Mahmassani, and I Ömer Verbas. Joint design of multimodal transit networks and shared autonomous mobility fleets. *Transportation Research Part C: Emerging Technologies*, 113:2–20, 2020.

Warren B Powell, Patrick Jaillet, and Amedeo Odoni. Stochastic and dynamic networks and routing. *Handbooks in operations research and management science*, 8:141–295, 1995.

Guoyang Qin, Qi Luo, Yafeng Yin, Jian Sun, and Jieping Ye. Optimizing matching time intervals for ride-hailing services using reinforcement learning. *Transportation Research Part C: Emerging Technologies*, 129:103239, 2021.

Meng Qu, Hengshu Zhu, Junming Liu, Guannan Liu, and Hui Xiong. 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, 2014.

Mohsen Ramezani and Mehdi Nourinejad. Dynamic modeling and control of taxi services in large-scale urban networks: A macroscopic approach. *Transportation Research Part C: Emerging Technologies*, 94:203–219, 2018.

Lisa Rayle, Susan A Shaheen, Nelson Chan, Danielle Dai, and Robert Cervero. App-based, on-demand ride services: comparing taxi and ridesourcing trips and user characteristics in san francisco. Technical report, Citeseer, 2014.

Herbert Robbins and Sutton Monro. A stochastic approximation method. *The annals of mathematical statistics*, pages 400–407, 1951.

Jean-Charles Rochet and Jean Tirole. Platform competition in two-sided markets. *Journal of the european economic association*, 1(4):990–1029, 2003.

Jean-Charles Rochet and Jean Tirole. Two-sided markets: a progress report. *The RAND journal of economics*, 37(3):645–667, 2006.

Tim Roughgarden and Éva Tardos. How bad is selfish routing? *Journal of the ACM (JACM)*, 49 (2):236–259, 2002.

Claudio Ruch, Sebastian Hörl, and Emilio Frazzoli. Amodeus, a simulation-based testbed for autonomous mobility-on-demand systems. In *2018 21st International Conference on Intelligent Transportation Systems (ITSC)*, pages 3639–3644. IEEE, 2018.

Marc Rysman. The economics of two-sided markets. *Journal of economic perspectives*, 23(3): 125–43, 2009.

Mauro Salazar, Federico Rossi, Maximilian Schiffer, Christopher H Onder, and Marco Pavone. On the interaction between autonomous mobility-on-demand and public transportation systems. In *2018 21st International Conference on Intelligent Transportation Systems (ITSC)*, pages 2262–2269. IEEE, 2018.

Rabih Salhab, Jerome Le Ny, and Roland P Malhamé. A dynamic ride-sourcing game with many drivers. In *2017 55th Annual Allerton Conference on Communication, Control, and Computing (Allerton)*, pages 770–775. IEEE, 2017.

Paolo Santi, Giovanni Resta, Michael Szell, Stanislav Sobolevsky, Steven H Strogatz, and Carlo Ratti. Quantifying the benefits of vehicle pooling with shareability networks. *Proceedings of the National Academy of Sciences*, 111(37):13290–13294, 2014.

Bruce Schaller. Empty seats, full streets: Fixing manhattan's traffic problem. Technical report, Schaller Consulting, 2017a.

Bruce Schaller. Unsustainable? the growth of app-based ride services and traffic, travel and the future of new york city, 2017b.

Bruce Schaller. Making congestion pricing work for traffic and transit in New York City. Technical report, 2018.

Joseph P Schwieterman. Uber economics: evaluating the monetary and travel time trade-offs of transportation network companies and transit service in chicago, illinois. *Transportation Research Record*, 2673(4):295–304, 2019.

Matthew H Shapiro. Density of Demand and the Benefit of Uber. Availabel at `http://www.shapiromh.com.`, 2018.

Yosef Sheffi. *Urban transportation networks*, volume 6. Prentice-Hall, Englewood Cliffs, NJ, 1985.

Yosef Sheffi and Warren B Powell. An algorithm for the equilibrium assignment problem with random link times. *Networks*, 12(2):191–207, 1982.

Yu Shen, Hongmou Zhang, and Jinhua Zhao. Integrating shared autonomous vehicle in public transportation system: A supply-side simulation of the first-mile service in singapore. *Transportation Research Part A: Policy and Practice*, 113:125–136, 2018.

Zhenyu Shou and Xuan Di. Reward design for driver repositioning using multi-agent reinforcement learning. *Transportation research part C: emerging technologies*, 119:102738, 2020.

Zhenyu Shou, Xuan Di, Jieping Ye, Hongtu Zhu, Zhang Hua, and Robert Hampshire. Optimal passenger-seeking policies on e-hailing platforms using markov decision process and imitation learning. 111:91–113, 2020. Transportation Research Part C: Emerging Technologies.

Kenneth A Small and Harvey S Rosen. Applied welfare economics with discrete choice models. *Econometrica: Journal of the Econometric Society*, pages 105–130, 1981.

Michael J Smith. The stability of a dynamic model of traffic assignment—an application of a method of lyapunov. *Transportation science*, 18(3):245–252, 1984.

Steven H Strogatz. *Nonlinear dynamics and chaos: with applications to physics, biology, chemistry, and engineering*. CRC Press, 2018.

Hao Sun, Hai Wang, and Zhixi Wan. Model and analysis of labor supply for ride-sharing platforms in the presence of sample self-selection and endogeneity. *Transportation Research Part B: Methodological*, 125:76–93, 2019.

Remi Tachet, Oleguer Sagarra, Paolo Santi, Giovanni Resta, Michael Szell, Steven H Strogatz, and Carlo Ratti. Scaling law of urban ride sharing. *Scientific reports*, 7(1):1–6, 2017.

Roger L Tobin and Terry L Friesz. Sensitivity analysis for equilibrium network flow. *Transportation Science*, 22(4):242–250, 1988.

Yongxin Tong, Libin Wang, Zimu Zhou, Lei Chen, Bowen Du, and Jieping Ye. Dynamic pricing in spatial crowdsourcing: A matching-based approach. In *Proceedings of the 2018 International Conference on Management of Data*, pages 773–788. ACM, 2018.

Catherine Tucker and Juanjuan Zhang. Growing two-sided networks by advertising the user base: A field experiment. *Marketing Science*, 29(5):805–814, 2010.

Junji Urata, Zhengtian Xu, Jintao Ke, Yafeng Yin, Guojun Wu, Hai Yang, and Jieping Ye. Learning ride-sourcing drivers' customer-searching behavior: A dynamic discrete choice approach. *Transportation research part C: emerging technologies*, 130:103293, 2021.

Unknown US Bureau of Public Roads. *Traffic Assignment Manual for Application with a Large, High Speed Computer*. US Department of Commerce, 1964.

Daniel Vignon, Yafeng Yin, and Jintao Ke. Regulating ridesourcing services with product differentiation and congestion externality. *Transportation Research Part C: Emerging Technologies*, 127, 2021.

Heinrich Von Stackelberg. *Market structure and equilibrium*. Springer Science & Business Media, 2010.

Guangju Wang, Hailun Zhang, and Jiheng Zhang. On-demand ride-matching in a spatial model with abandonment and cancellation. 2019.

Xiaolei Wang, Fang He, Hai Yang, and H Oliver Gao. Pricing strategies for a taxi-hailing platform. *Transportation Research Part E: Logistics and Transportation Review*, 93:212–231, 2016.

John Glen Wardrop. Some theoretical aspects of road traffic research. *Proceedings of the institution of civil engineers*, 1(3):325–362, 1952.

Ariel Waserhole and Vincent Jost. Vehicle sharing system pricing regulation: A fluid approximation. 2012. Availabel at `https://hal.archives-ouvertes.fr/hal-00727041/document` (Accessed: 2019-08-08).

Huw CWL Williams. On the formation of travel demand models and economic evaluation measures of user benefit. *Environment and planning A*, 9(3):285–344, 1977.

Salomón Wollenstein-Betech, Ioannis Ch Paschalidis, and Christos G Cassandras. Joint pricing and rebalancing of autonomous mobility-on-demand systems. Available at arXiv:2003.13614 (Accessed: 2020-04-05), 2020.

Salomón Wollenstein-Betech, Mauro Salazar, Arian Houshmand, Marco Pavone, Ioannis Ch Paschalidis, and Christos G Cassandras. Routing and rebalancing intermodal autonomous mobility-on-demand systems in mixed traffic. *IEEE Transactions on Intelligent Transportation Systems*, 20 (XX), 2021.

Shining Wu, Shihong Xiao, and Saif Benjaafar. Two-sided competition between on-demand service platforms, 2020. Available at SSRN 3525971 (Accessed: 2020-03-04).

Zhe Xu, Zhixin Li, Qingwen Guan, Dingshui Zhang, Qiang Li, Junxiao Nan, Chunyang Liu, Wei Bian, and Jieping Ye. Large-scale order dispatch in on-demand ride-hailing platforms: A learning and planning approach. In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pages 905–913. ACM, 2018.

Zhengtian Xu, Zhibin Chen, and Yafeng Yin. Equilibrium analysis of urban traffic networks with ride-sourcing services. Available at SSRN 3422294 (Accessed: 2019-07-25), 2019.

Zhengtian Xu, Yafeng Yin, and Jieping Ye. On the supply curve of ride-hailing systems. *Transportation Research Part B: Methodological*, 132:29–43, 2020.

Zhengtian Xu, Yafeng Yin, Xiuli Chao, Hongtu Zhu, and Jieping Ye. A generalized fluid model of ride-hailing systems. *Transportation Research Part B: Methodological*, 150:587–605, 2021.

Chiwei Yan, Helin Zhu, Nikita Korolko, and Dawn Woodard. Dynamic pricing and matching in ride-hailing platforms. *Naval Research Logistics (NRL)*, 2019.

Hai Yang. Heuristic algorithms for the bilevel origin-destination matrix estimation problem. *Transportation Research Part B: Methodological*, 29(4):231–242, 1995.

Hai Yang and SC Wong. A network model of urban taxi services. *Transportation Research Part B: Methodological*, 32(4):235–246, 1998.

Hai Yang and Teng Yang. Equilibrium properties of taxi markets with search frictions. *Transportation Research Part B: Methodological*, 45(4):696–713, 2011.

Hai Yang, Xiaoning Zhang, and Qiang Meng. Stackelberg games and multiple equilibrium behaviors on networks. *Transportation Research Part B: Methodological*, 41(8):841–861, 2007.

Hai Yang, Cowina WY Leung, SC Wong, and Michael GH Bell. Equilibria of bilateral taxi–customer searching and meeting on networks. *Transportation Research Part B: Methodological*, 44(8):1067–1083, 2010.

Hai Yang, Jintao Ke, and Jieping Ye. A universal distribution law of network detour ratios. *Transportation Research Part C: Emerging Technologies*, 96:22–37, 2018.

Hai Yang, Xiaoran Qin, Jintao Ke, and Jieping Ye. Optimizing matching time interval and matching radius in on-demand ride-sourcing markets. *Transportation Research Part B: Methodological*, 131: 84–105, 2020a.

Hai Yang, Chaoyi Shao, Hai Wang, and Jieping Ye. Integrated reward scheme and surge pricing in a ridesourcing market. *Transportation Research Part B: Methodological*, 134:126–142, 2020b.

Yongjian Yang, Xintao Wang, Yuanbo Xu, and Qiuyang Huang. Multiagent reinforcement learning-based taxi predispatching model to balance taxi supply and demand. *Journal of Advanced Transportation*, 2020, 2020c.

Jiayi Joey Yu, Christopher S Tang, Zuo-Jun Max Shen, and Xiqun Michael Chen. A balancing act of regulating on-demand ride services. *Management Science*, 2019a.

Xinlian Yu, Song Gao, Xianbiao Hu, and Hyoshin Park. A markov decision process approach to vacant taxi routing with e-hailing. *Transportation Research Part B: Methodological*, 121:114–134, 2019b.

Jing Yuan, Yu Zheng, Liuhang Zhang, XIng Xie, and Guangzhong Sun. Where to find my next passenger. In *Proceedings of the 13th international conference on Ubiquitous computing*, pages 109–118, 2011.

Liteng Zha, Yafeng Yin, and Hai Yang. Economic analysis of ride-sourcing markets. *Transportation Research Part C: Emerging Technologies*, 71:249–266, 2016.

Liteng Zha, Yafeng Yin, and Yuchuan Du. Surge pricing and labor supply in the ride-sourcing market. *Transportation Research Part B: Methodological*, 117(PB):708–722, 2018a.

Liteng Zha, Yafeng Yin, and Zhengtian Xu. Geometric matching and spatial pricing in ride-sourcing markets. *Transportation Research Part C: Emerging Technologies*, 92:58–75, 2018b.

Daqing Zhang, Lin Sun, Bin Li, Chao Chen, Gang Pan, Shijian Li, and Zhaohui Wu. Understanding taxi service strategies from taxi gps traces. *IEEE Transactions on Intelligent Transportation Systems*, 16(1):123–135, 2014.

Kenan Zhang and Yu Marco Nie. Mitigating the impact of selfish routing: An optimal-ratio control scheme (orcs) inspired by autonomous driving. *Transportation Research Part C: Emerging Technologies*, 87:75–90, 2018.

Kenan Zhang, Hongyu Chen, Song Yao, Linli Xu, Jiaoju Ge, Xiaobo Liu, and Yu Marco Nie. An efficiency paradox of uberization. Availabel at: SSRN 3462912 (Accessed: 2019-10-15), 2019a.

Kenan Zhang, Ying Chen, and Yu Marco Nie. Hunting image: Taxi search strategy recognition using sparse subspace clustering. *Transportation Research Part C: Emerging Technologies*, 109: 250–266, 2019b.

Rick Zhang and Marco Pavone. Control of robotic mobility-on-demand systems: a queueing-theoretical perspective. *The International Journal of Robotics Research*, 35(1-3):186–203, 2016.

Jing Zhou, William HK Lam, and Benjamin G Heydecker. The generalized nash equilibrium model for oligopolistic transit market with elastic demand. *Transportation Research Part B: Methodological*, 39(6):519–544, 2005.