Essays on On-Demand Transportation Services: Innovative Technologies and New Business Models

A THESIS
SUBMITTED TO THE GRADUATE SCHOOL
OF THE UNIVERSITY OF MINNESOTA
BY

Xiaotang Yang

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

Advisor: Saif Benjaafar

Auguest, 2023

Acknowledgements

First and foremost, I owe my deepest gratitude to my advisor, Saif Benjaafar. My journey into operations management began a few years ago when I enrolled in a course he instructed. It was through his insightful and captivating lectures that I discovered my passion for OM studies, leading me to embark on this PhD journey. Our every interaction, be it meetings or discussions, was a learning experience brimming with inspiration. Saif's innovative thinking, expansive vision, deep knowledge, and fervent passion have been guiding lights for me. His unwavering support was evident in how he meticulously revised my work, prepared me for presentations, and offered me teaching opportunities. Furthermore, his generosity in sharing resources and information to prepare me for an academic career is deeply appreciated. Simple words fall short of expressing the depth of my gratitude. It has been a great privilege to have been his student, and his guidance, support, and encouragement have been invaluable in shaping my growth.

I am immensely fortunate to have Prof. William Cooper, Prof. Tony Haitao Cui, and Prof. Krishnamurthy Iyer as members of my dissertation committee. Their guidance and support, dating back to my preliminary exams three years ago, have been invaluable. Prof. Cooper meticulously reviewed and revised my research statement and CV prior to my job applications, ensuring every detail was in place. Prof. Iyer generously dedicated an afternoon to thoroughly go through my job talk slides, helping me refine them for my first campus visit. Meanwhile, Prof. Cui imparted valuable insights on behavioral research, broadening my perspective in this domain. Additionally, I would like to convey my heartfelt appreciation to Prof. Mohsen Elhafsi and Prof. Karen Donohue. Prof. Elhafsi consistently provided unwavering support, quickly offering feedback and encouragement whenever needed. Prof. Donohue introduced me to the fascinating world of behavioral

operations, for which I am deeply grateful.

My sincere appreciation also goes to Prof. Shuzhong Zhang, Prof. Nick Arnosti, Prof. Martín Zubeldía, Prof. Kevin Leder, Prof. Saumya Sinha, Prof. Sherwin Doroudi, Prof. Zhaosong Lu, and Prof. Ankur Mani for their invaluable advice, continuous encouragement and support over the past years. I am deeply thankful to Behrooz Pourghannad, Guiyun Feng, Xiaobo Li, Rowan Wang, David Chen, Yimin Yu, Xiang Li and Junfeng Zhu. Their willingness to share their expertise and insights, spanning both research and personal experiences, has enriched my journey. I have gained immeasurably from our interactions. I extend my thanks to the department administrators, with special appreciation to Hongna Byström, Alyssa Benson, and Teresa Nieszner. Their dedication and assistance ensured a seamless progression through my PhD journey.

I'm grateful for the companionship and help of my friends throughout this journey. Sharing an office with Kang Kang and Bingnan Lu was a delight; they transformed our workspace into a place of joy. My time spent with Amy Yuanyuan Ding and Zu You was a source of relaxation and upliftment, and I cherish our moments together. Engaging in the weekly softball and badminton club activities with Jingyuan Wan, Beibei Jia, Chenyu Wu, Xuanming Zhang, Chen Jiang, Chuan He, Hanyang Li, and Xiaonan Li was a highlight, and my gratitude goes to the club organizers, Nuozhou Wang and Sanyou Mei. Lastly, my appreciation extends to my peers and friends at the University of Minnesota: Einar Bjarki Gunnarsson, Jazeem Abdul Jaleel, Jiali Huang, Jing Gao, Xiaobing Shen, Alexander Wickeham, P.A. Nguyen, and Can Yin.

Lastly, I am immensely grateful to my parents, Jian and Fang, for their unconditional love and care. I'm also thankful to my parents in-laws, Ping and Jianxin, for their consistent encouragement. I would like to thank my husband, Zicheng, who brings light and emotional depth to my every day. I can't imagine my life without his unwavering love.

Contents

List of Figures

Chapter 1

Introduction

Technological advances in data communication, portable devices, and electronic payment have led innovative businesses to rapid growth in *on-demand transportation services* enabled by digital platforms. Platforms create value by facilitating communication and matching between supply (vehicles and drivers) and demand (customers). The operation of on-demand transportation services raises unique challenges, including the spatial mismatch between demand and supply and the reliance on independent drivers who act strategically. Breakthroughs in technology (automation and artificial intelligence, among others) and carefully-designed operating policies hold the promise of increasing the productivity and improving the efficiency of these services. In this dissertation, we aim to understand the impact of innovative technologies and new business models in the context of this application on multiple stakeholders, including customers, strategic drivers, and platforms.

In Chapter 2, we analyze the impact of a new technology, *tele-driving*, on transportation-enabled service systems. Tele-driving refers to a novel concept where drivers can remotely operate vehicles (without being physically in the vehicle). By putting the human back "in the loop," tele-driving has emerged recently as a more viable alternative to fully automated vehicles, with ride-hailing (and other on-demand transportation-enabled services) being an important application. Because remote drivers can be operated as a shared resource (any driver can be assigned to any customer regardless of trip origin or destination), it may be possible for such services to deploy fewer drivers than vehicles without significantly

reducing service quality. In this paper, we examine the extent to which this is possible. Using a spatial queueing model that captures the dynamics of both pick up and trip times, we show that the impact of reducing the number of drivers depends crucially on system workload relative to the number of vehicles. In particular, when workload is sufficiently high relative to the number of vehicles, we show that, perhaps surprisingly, reducing the number of drivers relative to the number of vehicles can actually improve service level (e.g., as measured by the amount of demand fulfilled in the case of impatient customers). When workload is sufficiently low relative to the number of vehicles, we show that it is possible to significantly reduce the number of drivers without significantly reducing service level. In systems where customers are patient and willing to queue up for the service, we show that reducing the number of drivers can stabilize a system that would otherwise be unstable. In general, relative to a system where the number of vehicles equals the number of drivers (as in a system with in-vehicle drivers), a system with remote drivers can result in savings in the number of drivers either without significantly degrading performance or while actually improving performance. We discuss how these results can, in part, be explained by the interplay of two counteracting forces: (1) having fewer drivers increasing "service rate" and (2) having fewer drivers reducing the number of "servers," with the relative strength of these forces depending on system workload.

In Chapter 3, motivated by the behavior of drivers on ride-hailing (individual drivers decide whether or not to work based on the offered wage and where to locate themselves in anticipation of future fares), we examine how the introduction of autonomous vehicles impacts the strategic behavior of human drivers and driver welfare. Specifically, we consider a setting where a ride-hailing platform deploys a mixed fleet of conventional vehicles (CVs) and autonomous vehicles (AVs). The CVs are operated by human drivers who make independent decisions about whether to work for the platform and where to position themselves when they become idle. The AVs are under the control of the platform. The platform decides on the wage it pays the drivers, the size of the AV fleet, and how the AVs are positioned spatially when they are idle. The platform can also make decisions on whether to prioritize the AVs or the CVs in assigning vehicles to customer requests. We use a fluid model to characterize the optimal decisions of the platform and contrast those with the optimal decisions in the absence of AVs. We examine the impact of automation on

strategic drivers and the ride-hailing platform. We show that, although the introduction of AVs can displace drivers and depress wages, there are settings where the introduction of AVs leads to higher wages and more drivers being hired. We discuss how these results can, in part, be explained by the interplay of two counteracting effects: (1) the introduction of AVs provides the platform with an additional source of supply and renders human driver substitutable (*displacement effect*); and (2) having access to and control over AVs enables the platform to influence the strategic behavior of CVs, thereby reducing the inefficiency from self-interested behavior (*incentive effect*). The relative strength of these two effects depends on the cost of AVs and the vehicle dispatching policy. Our results uncover a new effect through which the introduction of AVs affects the welfare of human drivers (the incentive effect). Our results have potentially broader applications to other areas where automation is introduced and where workers are strategic.

In Chapter 4, we study competition between two service platforms: an incumbent and a new entrant. The new entrant differentiates itself from the incumbent by occupying a different niche of the market. After the entry of the new platform, the two platforms compete for both workers and customers by deciding on wages to pay workers and prices to charge customers. Workers, who are heterogeneous in their opportunity costs, are sensitive to both the wage they receive and the amount of work they are allocated. Customers, who are heterogeneous in their preferences for the platforms, are sensitive to both the price and the amount of delay they experience. We use an equilibrium model to study price-wage strategies of platforms. We compare equilibrium outcomes before and after the entry of the new platform. We show that competition does not necessarily lead to higher worker welfare and higher consumer surplus. In particular, we show that when the worker pool size is sufficiently large and customer stickiness (the strength of preference of customers for one platform over another) is moderate, both workers and customers are worse off after the new platform enters the market (with workers being busier but earning less, and customers paying more but experiencing more delay). Competition is often viewed as being socially desirable. The results in this paper suggest that some caution is warranted when competition is between service platforms that compete for both workers and customers with workers who multi-home. We distinguish forces that explain our results: multi-homing of workers and stickiness of customers. Such information could be useful for policy makers and

platform managers as they consider the implications of competition between on-demand service platforms on social welfare and profit.

Chapter 2

Human in The Loop Automation: Ride-Hailing with Remote (Tele-) Drivers

2.1 Introduction

It is becoming increasingly apparent that fully autonomous vehicles may take longer than initially expected to become a reality. There is growing consensus that large scale deployment of level 5 autonomy (full autonomy) is unlikely to happen soon (Doll et al. (2020)). Even if the technological challenges were to be surmounted in the near future, the public acceptance of full autonomy may take longer. In particular, concerns about safety continue to be high, along with unease about delegating to machines critical life and death decisions. A new technology, that combines vehicles that are nearly autonomous with tele-operators, has recently emerged with the potential of overcoming both the technological hurdles and the societal concerns of "driverless" vehicles. Tele-operated vehicles may provide the efficiency and flexibility of autonomous vehicles while keeping "humans in the loop" and ultimately responsible for driving decisions.

Several pilots are demonstrating the commercial viability of tele-driving. Vay, a German car sharing company, has begun testing a car sharing service where vehicles operated by "tele-drivers" from a central command center (Blanco (2021)). Vay's stated aim is to offer

"an Uber-like service with remote drivers." Halo is a US-based company collaborating with T-Mobile, a telecommunication company, to pilot a "robo-taxi" service with 5G-powered vehicles that can be operated remotely (Pegoraro (2021)). The viability of tele-driving has been demonstrated in other settings, including food delivery using remotely-controlled robots (e.g., Coco Coco (2022) and Starship Starship (2022)), material handling in factories and warehouses (e.g., remotely-operated forklift trucks (PhantomAuto (2022)) and remotely-operated trucks (Sawers (2022)). In these and other applications, tele-driving is being enabled by advances in wireless communication (e.g., 5G technology and AWS wavelength technology to minimize latency), video compression (to enable remote video display), and edge computing (to allow for in-vehicle data processing and controls), among others; see Sawers (2020) for further discussion.

Tele-operated vehicles in the context of ride-hailing (the focus of this paper) offer several advantages (over conventional vehicles with in-vehicle drivers). Perhaps the most important of these is the ability to treat remote drivers as a common resource. While vehicles continue to be constrained by their geographic location, remote drivers can be interchangeably assigned to fulfill trips regardless of a trip origin or destination. Moreover, the fact that each driver is not dedicated to a single vehicle implies that a service could operate with fewer drivers than vehicles, important in settings where drivers are costly or are in short supply.

In this paper, we examine the extent to which this is possible. Specifically, we are interested in investigating the impact of operating with fewer remote drivers than vehicles on system performance, where performance measures of interest include the amount of demand that can be fulfilled and the amount of delay customers experience[1]. The setting we consider is of a service that operates over a continuous region. Requests for transport

[1]There are potentially other benefits to tele-operation: (1) it could reduce inefficiencies associated with drivers acting strategically by driving empty from locations perceived to be low demand to those perceived to be high demand (Ashkrof et al. (2022)); (2) it can eliminate discriminatory behavior on the part of drivers towards travelers based on their race and gender (Ge et al. (2016) and Tang et al. (2021)); (3) it can increase the safety of both drivers and riders, particularly women who have been disproportionately the targets of in-vehicle assault and other criminal behavior (According to Uber, in 2020, "rape incidents made up 0.00002% of total trips. About 91% of the victims of rape were riders and about 7% of the victims were drivers. Women made up 81% of the victims while men comprised about 15%" (O'Brien (2022)); (4) it removes the requirement that drivers own a vehicle (reducing their costs and the associated risks of owning a physical asset); and (5) it could broaden labor participation as drivers may work remotely at locations that are most convenient (Robotics Tomorrow (2022)).

(from an origin to a destination) arise continuously over time. The operator of the service manages a fleet of vehicles and a pool of remote drivers. We consider two settings, one where customers are impatient and one where they are patient. In the case of impatient customers, a customer would leave the system without receiving service if they cannot be immediately matched with a vehicle and a driver. In the case of patient customers, a customer is willing to wait to be matched. The operator of the service decides on which vehicles to match with which customers (remote drivers are assumed to be homogeneous, and any available driver can be matched with any customer-vehicle pair). Once a vehicle-driver pair is assigned to a customer, the driver takes over the control of the vehicle and drives it (remotely) to the location of the customer. Once the customer is picked up, the remote driver drives the vehicle to the customer's requested destination. Upon trip completion, the driver and the vehicle are "uncoupled" and become available to be independently assigned to future customer requests (see Figure 2.1 for a graphical illustration).

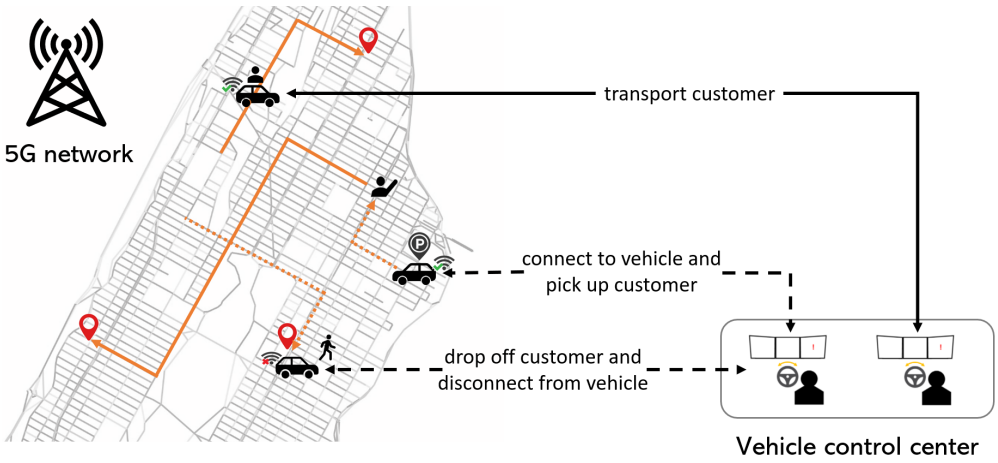


Figure 2.1: An illustration of a ride-hailing service with tele-drivers

Note that the setting we consider is one where the number of vehicles and drivers are determined by the service provider. The service provider is also in control of decisions about matching requests with vehicles and drivers and vehicle repositioning (i.e., we do not allow for drivers to act strategically).

We show how the dynamics of the system can be approximately captured by a multi-server queueing model with state-dependent service time. We use this model to examine the impact of varying the number of drivers using a benchmark case where the number of drivers equals the number of vehicles. The following is a summary of our main findings. We discuss first the case of impatient customers.

- There are three distinct operating regimes in which the effect of varying the number of remote drivers varies: a *supply-limited* regime, an *intermediate* regime, and a *supply-rich* regime, where "supply" refers to the number of vehicles relative to workload. In the supply-limited regime, we show that reducing the number of drivers (relative to the number of vehicles) can actually lead to a higher service level, as measured by the fraction of demand fulfilled. We provide a necessary and sufficient condition for this to occur and show that the improvement in service level can be significant. In general, in this regime, the effect of varying the number of drivers on service level (all else staying the same) is non-monotonic.

- In the intermediate regime, we show that reducing the number of remote drivers can similarly lead to a higher service level. We provide a necessary and sufficient condition for this to occur. However, when it occurs, the improvement can be insignificant (asymptotically, the improvement is negligible).

- In the supply-rich regime, we show that, although reducing the number of drivers always leads to a lower service level, the decrease in service level can be small even when the decrease in the number of drivers is significant (asymptotically, this decrease is at least a half). In general, we provide a well-specified bound on the relationship between service level and the number of drivers when the number of drivers is sufficiently large.

The following is a summary of our main findings for systems with patient customers.

- Here too, there are three distinct regimes. In the supply-limited regime, the system, regardless of the number of drivers is unstable (customer delay is not finite).

- In the intermediate regime, we show that reducing the number of drivers can make a system that is otherwise unstable stable.

- In the supply-rich regime, we show that a result similar to the one obtained for impatient customers holds. That is, it is possible to significantly decrease the number of drivers while only marginally increasing expected customer delay (asymptotically, the number of drivers can be decreased by at least a half).

Lastly, we provide results from numerical experiments using real world taxi data from the city of New York.

An intuitive explanation. The results obtained in this paper can be explained by the interplay of two counteracting forces: (1) having fewer drivers increases "service rate," or equivalently reduces "service time," and (2) having fewer drivers reduces the number of "servers." In particular, we note the following (the statements will be made precise in subsequent sections of the paper).

- *Pick-up time decreases with the number of idle vehicles.* In a ride-hailing system, vehicles can be in one of three modes: (1) idle, (2) on the way to pick up a customer, and (3) en-route transporting a customer (trip time). This means that the effective service time is the sum of pick-up time and trip time. While trip time is determined exogenously by the origin and destination of the trips requested by the customers, pick-up times depend on the number of idle vehicles. In particular, fewer idle vehicles increase the likelihood that a customer request would be matched with a distant vehicle, resulting in a longer pick-up time. This means that, when the system is short on vehicles, pick-up times are likely to be the longest.

- *Having fewer drivers than vehicles reduces pick-up times.* Having fewer drivers than vehicles ensures that there are always idle vehicles, with the fewer the drivers the likelier for there to be more idle vehicles. Consequently, the fewer the drivers the likelier for the pick-up times to be shorter (making overall shorter service times likelier). As we note next, having shorter service times comes however at the cost of having fewer "servers."

- *Having fewer drivers than vehicles reduces the number of servers.* Although reducing the number of drivers shortens pick-up times, it also reduces the number of servers

available to handle customer requests. Hence, unless matched with a sufficient reduction in pick-up times, reducing the number of drivers risks reducing overall service capacity.

Depending on the relative strength of these two forces, different outcomes are possible. Our analysis shows that this crucially depends on the number of vehicles relative to workload. In particular, the effect of shorter pick-up times is strong when the likelihood of long pick-up times is high. This is true when the system is short on vehicles. The effect of shorter pick-up times is weak when the number of vehicles is large. In this case, the number of idle vehicles tends to be large and hence, some reduction in the number of drivers does not have a significant effect.

These effects manifest themselves differently depending on whether customers are impatient or not. In the supply-limited regime, having fewer drivers can improve system throughput (amount of demand fulfilled) when customers are impatient. In the intermediate regime, having fewer drivers can stabilize the system when customers are patient. That is, in the supply-limited (and intermediate) regimes, having fewer drivers paradoxically increases capacity. In the supply-rich regime, there is an opportunity to significantly reduce the number of drivers without significantly reducing the performance (as measured by throughput in the case of impatient customers and delay in the case of patient customers). In the paper, we make all of the above statements precise and provide well-specified conditions under which the different noted outcomes arise.

Although we use ride-hailing as the motivating application for this paper, the models and the analysis have broader applicability to other tele-operated services such as food delivery (e.g., vehicles travelling to pick up a food order from a restaurant and then delivering to a home address) or warehouse operation (e.g., forklifts travelling to pick up products from one area of the warehouse to deliver them to another). The models and analysis can also be adapted to study settings where the remote driver is needed for only one segment of the trip (e.g., the remote driver is needed to deliver the car to a customer but the customer drives on her own to her desired destination), and to settings where the vehicle may not be a car but a small robot that navigates over sidewalks instead.

The rest of the paper is organized as follows. In Section 2.2, we discuss related literature. In Section 2.3 (2.4), we provide analysis for systems with impatient (patient) customers. In

Section 2.5, we provide additional discussion of various aspects of the modeling, analysis, and results. In Section 2.6, we offer concluding comments. Proofs for all the results, unless otherwise stated, are included in the Appendix.

2.2 Related Literature

This paper contributes to the growing body of literature on on-demand mobility, including ride-hailing. General reviews of this literature can be found in Benjaafar and Hu (2020), Hu (2021), and Freund et al. (2019). Papers within this literature that are of particular relevance include those that consider "dimensioning problems" and those that study "spatial mismatches between supply and demand."

System Dimensioning. Besbes et al. (2022) consider the capacity sizing problem for a ride-hailing service modeled as a multi-server queue. They show that the square root safety staffing rule (having buffer service capacity proportional to the square root of the nominal workload) is incapable of balancing the cost of supply capacity and service quality. Instead, buffer capacity must be in the order of a power of $\frac{2}{3}$ of the nominal workload. George and Xia (2011) model a one-way vehicle-sharing system as a closed queueing network and propose algorithms for determining the optimal number of vehicles. For the same problem, Benjaafar et al. (2022b) develop an approximation for the number of vehicles needed to guarantee a specified service level.

Spatial mismatches. There is a growing body of literature that studies the inefficiencies of ride-hailing systems caused by spatial mismatches between demand and supply. The "Wild Goose Chase" phenomenon, as described by Castillo et al. (2021), is the scenario where, under limited supply, drivers are dispatched to pick up customers who are too distant, exacerbating the supply shortage. Feng et al. (2021) consider the case of a ride-hailing platform that operates on a circular road. They devise an approximation approach that models the non-monotonic relationship between expected pick-up time and customer arrival rate.

Additionally, there is a substantial body of research devoted to resolving spatial mismatches. A widely considered strategy is to impose a matching radius, meaning that customers and vehicles are only matched if they are sufficiently close; another strategy is to match a vehicle with the closest customer regardless of the order in which customers

arrived (Castillo et al. (2021), Feng et al. (2021), Wang et al. (2022), and Xu et al. (2020)). Feng et al. (2021) propose a heuristic method for determining the near-optimal matching radius that minimizes customer average waiting time. Wang et al. (2022) develop a fluid model for a system in which a customer may cancel the service if the platform does not allocate a driver promptly. They obtain the fluid-based optimal matching radius in order to maximize system throughput. Xu et al. (2020) model the system as a double-ended queue and show that the smaller the matching radius, the weaker the "Wild Goose Chase" effect. Castillo et al. (2021) demonstrate that, despite the fact that implementing the matching radius can resolve "Wild Goose Chase," it is dominated by a "surge" pricing strategy. In this paper, we show that the wild goose chase phenomenon could be mitigated using remote drivers who are fewer in numbers than vehicles.

Another stream of literature investigates spatial matching strategies. Banerjee et al. (2022b) adopt a closed queueing network model to study the dynamic matching between vehicles and customers. They propose a family of state-dependent policies which can result in an exponential decay of demand-loss probability as system parameters scale to infinity. Kanoria (2021) investigate matching units of supply and demand in a d-dimensional hypercube under various models. They demonstrate how the minimum achievable cost (expected average distance between matched pairs) scales with model parameters and provide a matching algorithm that can achieve a near-optimal cost in the dynamic model. Özkan and Ward (2020) consider a ride-hailing system with time-varying driver and customer arrival rates, taking into consideration both the drivers' and customers' patience. They propose a matching policy based on a continuous linear program and show that it is asymptotically optimal in terms of maximizing the number of (weighted) matches. There is also literature that studies how (temporally or spatially) pooling strategies, such as serving more than one customer at a time or allowing pickups only at designated locations) can be used to improve matching (see for example Hu (2022), Chen and Hu (2022), Cao and Qi (2022) and Santi et al. (2014)).

A stream of literature considers how curtailing demand, either indirectly via pricing and directly via admission control, can be used to address the spatial mismatch between supply and demand; see for example Waserhole and Jost (2016), Banerjee et al. (2022a), Benjaafar and Shen (2023), and Afèche et al. (2022). The literature on admission control

in queueing systems is extensive; see for example Stidham (1985), Altman et al. (2001), Örmeci et al. (2001), Stidham (2002).

Our work contributes to the emerging body of literature on **the impact of autonomous vehicles (AVs) on transportation**. Benjaafar et al. (2023b) investigate how a ride-hailing platform may use autonomous vehicles (through repositioning and vehicle assignment priorities) to achieve socially desirable outcomes. Siddiq and Taylor (2022a), Lian and Van Ryzin (2022), Noh et al. (2021), and Castro and Frazelle (2022) examine the introduction of autonomous vehicles from a market design point of view, including the ownership structure and the competition between platforms. Mirzaeian et al. (2020) study the impact of AVs on highway traffic. They do so using a state-dependent queueing system. They do not consider the dynamics of passenger pickup and the associated spatial features. Our work also contributes to the stream of literature that considers service systems with **state-dependent service rate**; see Delasay et al. (2019) for a recent review. In systems where servers correspond to human operators, as in medical clinics or retail stores, there are empirical studies that provide evidence of servers either speeding up (see Wang and Zhou (2018) and Kc and Terwiesch (2009) for example) or slowing down (see Batt and Terwiesch (2012) and Chan et al. (2017)) when these servers observe longer queues. Analytical models that incorporate state-dependent queues include Mandelbaum (1995), Mandelbaum and Pats (1998), Zhong et al. (2022), Delasay et al. (2016) and Dong et al. (2015).

Thematically, papers that are closest to ours are Hampshire et al. (2020) and Daw et al. (2020). Hampshire et al. (2020) consider a setting with autonomous vehicles where remote drivers intervene only to resolve "edge cases" (i.e., situations that the automated system cannot handle on its own). They analyze the system using a standard multi-server queueing model (i.e., one where the service times are not state-dependent) and estimate the necessary number of remote drivers that would satisfy a target service using either the Elang-B or Erlang-C formula. Daw et al. (2020) consider a related system where the requests for assistance from the remote drivers arrive in batches. Our work is different from Hampshire et al. (2020) and Daw et al. (2020) in that (1) we consider a setting where the human support lasts for the full duration of the service while Hampshire et al. (2020) and Daw et al. (2020) are primarily concerned with remote drivers handling edge cases;

and (2) we account for the spatial feature of the service and how this feature interacts with the number of servers. As a result, the modeling and analysis are different and so are the results.

Finally, there is growing literature in various fields of engineering that addresses the technical requirements for tele-operated driving. Reviews of this literature can be found in Zhang (2020) and the references therein.

2.3 Problem Formulation

We consider a ride-hailing platform that relies on remote drivers who operate vehicles at a distance. Requests for trips arise continuously over time, with each request associated with a pick-up and destination location across a specified service region. Let m and n, where with $m \geq n$, denote the number of vehicles and remote drivers, respectively (when $m = n$, the system with remote drivers, all else being the same, has dynamics similar to those of a system with in-vehicle drivers[2]).

We consider the case where customers are impatient (the case of patient customers is treated in Section 2.4). That is, if there are no idle drivers when a customer makes a trip request, the request cannot be fulfilled and the customer is considered lost. If there are idle drivers (which also implies the presence of idle vehicles since $m \geq n$) at the time a trip request is made, an idle driver is assigned to the vehicle that is nearest to the origin of the customer request. The driver takes over the control of the vehicle and drives it (remotely) to the location of the customer. Once the customer is picked up, the remote driver drives the vehicle to the customer's requested destination. Upon trip completion, the driver parks the vehicle where the service is terminated and the driver and the vehicle are "uncoupled" and become available to be independently assigned to future customer requests[3].

[2]See Section 2.5 for more discussions about the differences between tele-driving and traditional systems, including vehicle cost, labor cost, and the possibility that tele-driving is slower than in-vehicle driving and their impacts on our main results.

[3]We assume that vehicles can always find parking in a nearby location to where a customer is dropped off (e.g., at a metered parking or a parking garage). We expect service providers to make arrangements with the cities in which they operate to allow for such parking, not unlike the arrangements free-floating car-sharing companies have made (see, for example, https://eviecarshare.com/). Alternatively, a service provider may invest in a network of designated hubs at which idle vehicles can be left. In this scenario, tele-drivers may spend more time parking the vehicle, which can be added as a third component to service time. These different assumptions on parking can be easily incorporated into the model, but do not affect

The total service time for a customer who is matched with a vehicle and a driver consists of two components: a "pick-up time" and a "trip time." That is total service time = pick-up time + trip time. The pick-up time corresponds to the time it takes the nearest idle vehicle to the customer to travel from its current location to the location of the customer. The trip time corresponds to the time it takes the vehicle to travel from the customer's current location (the trip origin) to the customer's requested destination (the trip destination). Because the pick-up time depends on the location of the "nearest" vehicle, it varies with the number of idle vehicles. For example, when the number of idle vehicles is large, pick-up time is more likely to be short while the reverse is true (pick-up time is more likely to be long when the number of idle vehicles is small). In other words, pick-up time (and consequently total service time) is dependent on the state of the system.

The system as described above can be viewed as a special case of a multi-server queueing system with state-dependent service times. An exact treatment of this system is difficult in general, as service times are dependent on the demand process, its spatial distribution, and the topology of the road network. To allow for tractability, we ground our analysis in a Markovian approximation of both the demand and service process. In particular, we approximate the demand process by a Poisson process with rate λ where the origin-destination pairs associated with each requested trip are assumed to be uniformly distributed over the service region[4]. The service region is approximated by a continuous area, denoted by $\mathcal{C}$, that is a bounded and convex subset of $\mathbb{R}^2$. We also approximate the distribution of service times by a state-dependent exponential distribution [5] with rate $\mu(m, q)$, where $q \in \{1, \cdots, n\}$ is the the number of customers currently being served (customers waiting to be picked up + customers en route to their destination), and m is the number of vehicles. Note that, for a given m, $\mu(m, q)$ is a function of q defined for $q \in [1, m]$. We assume $\mu(m, q)$ satisfies a few mild conditions stated below which account for the spatial feature of a ride-hailing system.

---

our qualitative results.

[4]The distribution of origin and destination locations can be general as long as Assumption 2.3.1 and 2.3.2 hold. In Appendix A.8, we show that these assumptions hold for typical non-uniform traffic generated by large events such as concerts and sporting events. In Section 2.3.4, to test the robustness of our findings, we provide numerical results where the demand and service process are both data-driven using real world data.

[5]In the case where customers are impatient, this assumption can be relaxed. See footnote 9 for more discussions.

Assumption 2.3.1. *(1) $\mu(m,q)$ is strictly increasing in m and decreasing in q. (2) $\mu(m,q)$ is strictly concave in q. (3) $\lim_{m\to\infty} \mu(m,q) = \frac{1}{s}$, where s is the expected trip time between two uniformly* [6] *drawn locations in $\mathcal{C}$. (4) $\mu(m,m)$ is invariant in m*[7]*.*

In Assumption 2.3.1, Condition (1) specifies that the service rate is increasing in the number of vehicles and decreasing in the number of customers currently being served or, equivalently, is increasing in the number of idle vehicles. Condition (2) states that the decrease in the service rate due to having one fewer idle vehicle is more pronounced when the number of idle vehicles is low. Condition (3) states that the pick-up time approaches zero as the number of vehicles approaches infinity. Condition (4) states that the service rate does not vary with the number of vehicles when there is no idle vehicle present.

To account for the trade-off between reduced service capacity and shorter service times, we make the following assumption about the rate at which trips are completed when there are q customers in service, $q\mu(m,q)$.

Assumption 2.3.2. *There exists $\bar{m} > 0$ such if $m \geq \bar{m}$* [8]*, $q\mu(m,q)$ first increases then decreases in q.*

Assumption 2.3.2 states that the rate at which trips are completed when there are q customers in service, $q\mu(m,q)$ is not monotonically increasing in the number of customers being served, or, equivalently, with the number of busy drivers. Noting that $q\mu(m,q)$ is the product of two terms, one increasing, q, and the other decreasing, $\mu(m,q)$. The non-monotonicity implies that the effect of q is stronger when q is small while the effect of $\mu(m,q)$ is stronger when q is large. In Appendix A.8, we show that Assumptions 2.3.1 and 2.3.2 hold for a wide range of service region geometries and are consistent with various approximations of the functional form of $\mu(m,q)$ that have been employed in the literature (see for example Besbes et al. (2022), Kanoria (2021) and Feng et al. (2021)). Note that in a standard multi-server queue where the service rate is not state-dependent (i.e, one where $q\mu(m,q) = q\mu$), $q\mu(m,q)$ is monotonically increasing in q.

[6] Our analysis only requires that $\lim_{m\to\infty} \mu(m,q)$ exists and the limit can depend on the distribution of origin and destination locations.

[7] Our analysis only requires that $\mu(m,m)$ is non-decreasing in m.

[8] In general, our model yields $\bar{m} < 10$ for a wide range of service region geometries and various approximations of the functional form of $\mu(m,q)$ that have been employed in the literature (see for example Besbes et al. (2022), Kanoria (2021) and Feng et al. (2021)).

Using Kendall's notation, the queueing system we consider is an $M/M(q)/l/l$ [9] queueing system (or an Erlang loss system with a state-dependent service rate) where l is the number of servers (for our system with m vehicles and n driver, $l = n$).

Let $\pi_{m,n}(q)$ denote the stationary probability of having q customers in the system, given that the system has m vehicles and n remote drivers. These probabilities can be obtained by solving the underlying Markov chain, a birth and death process with birth rate λ and death rate $q\mu(m,q)$ when the system is in state q. Letting

$$\rho(q) = \frac{\lambda}{q\mu(m,q)}, \tag{2.1}$$

we have

$$\pi_{m,n}(0) = \left[1 + \sum_{i=1}^{n} \prod_{k=1}^{i} \rho(k)\right]^{-1}, \quad \text{and} \tag{2.2}$$

$$\pi_{m,n}(q) = \pi_{m,n}(0) \prod_{k=1}^{q} \rho(k), \quad \text{for} \;\; 1 \leq q \leq n. \tag{2.3}$$

Let $SL(m,n)$ denote the service level for a system with m vehicles and n remote drivers, where service level is defined as the long-run fraction of demand fulfilled. Noting that customers are turned away when all drivers are busy, we have

$$SL(m,n) = 1 - \pi_{m,n}(n).$$

We conclude this section by highlighting two important features of systems with remote drivers (i.e., systems where $n < m$). First, note that for a system with $n < m$, the service rate $\mu(m,q)$ when there are q customers in the system is the same as for one with $n = m$ and there are q customers in the system. That is, even though there are fewer drivers than vehicles, a system with remote drivers can leverage the larger number of vehicles to maintain the same service rate. This suggests that reducing the number of drivers when the drivers are remote may have a lesser impact on performance than in a traditional system

[9] Cheah and Smith (1994) showed the stochastic equivalence between the $M/M(q)/l/l$ system and the $M/G(q)/l/l$ system. Therefore, all of our results for the system with impatient customers hold for general service time distributions.

where reducing the number of drivers is accompanied by a reduction in the number of vehicles.

Second, note that the dynamics of the underlying birth and death process are the same as those for a system with $n = m$, except that the state space is truncated at $q = n$ instead of $q = m$. This means that customers are rejected when the system is in state $q = n$ instead of $q = m$. However, by preventing the system from being in states higher than n, expected service times are upper bounded by $\frac{1}{\mu(m,n)}$ (instead of $\frac{1}{\mu(m,m)}$ when $n = m$). The increased likelihood of shorter service times (due to always having at least $m - n$ idle vehicles) can be particularly valuable when the system is highly loaded and the likelihood of states higher than $q = n$ is high. In other words, by limiting the number of drivers, the likelihood of shorter service time is increased, albeit at the expense of decreasing the number of servers.

In view of these features, it is reasonable to conjecture that (a) when the system is highly loaded, having fewer drivers can actually lead to a better service level (due to the increased likelihood of shorter service times) and (b) when the system is lightly loaded, it may be possible to reduce the number of drivers without significantly reducing service level. The rest of this section provides confirmation for both conjectures.

2.3.1 Preliminaries: A System with An Equal Number of Vehicles and Drivers ($m = n$)

Before we provide additional analysis for systems with remote drivers where $n \leq m$, let us consider, as a benchmark, a system with $n = m$ (all else being the same, this can be viewed as a system with in-vehicle drivers). As we show in the following proposition, there are three distinct regimes of operation, depending on the number of vehicles, characterized by differences in the features of the distribution of the number of customers in the system and the associated system performance. In the next section, we describe how the introduction of remote drivers (i.e., allowing for the possibility of $n < m$) affects these features and the associated system performance. Recall that $\bar{m}$ is introduced in Assumption 2.3.2. Throughout the remainder of this paper, and for ease of presentation, we will restrict ourselves to the case where $m \geq \bar{m}$ and $\lambda > \max\{\frac{1}{s}, \max_{q\in\{1,2,\ldots,\bar{m}\}} q\mu(\bar{m}, q)\}$, which implies

that the system being studied is not too small in size[10] (in the setting we consider, both $\bar{m}$ and the lower bound for λ are very small).

Proposition 2.3.1. *There exist $\gamma_1(\lambda)$ and $\gamma_2(\lambda)$ with $0 < \gamma_1(\lambda) < \gamma_2(\lambda)$ such that the probabilities $\pi_{m,m}(q)$ satisfy the following properties:*

(i) when $m < \gamma_1(\lambda)$, $\pi_{m,m}(q)$ increases in q for $q \leq m$, implying that $\pi_{m,m}(q)$ is unimodal with the mode at m;

(ii) when $\gamma_1(\lambda) < m < \gamma_2(\lambda)$, there exist q_1 and q_2 with $0 < q_1 < q_2$ such that $\pi_{m,m}(q)$ increases in q for $q < q_1$, decreases in q for $q_1 < q < q_2$, and then increases in q for $q > q_2$, implying that $\pi_{m,m}(q)$ is bimodal with one mode at $\lfloor q_1 \rfloor < m$ and the other at m; and

(iii) when $m > \gamma_2(\lambda)$, there exists $q_3 > 0$[11] such that $\pi_{m,m}(q)$ increases in q for $q < q_3$, and decreases in q for $q > q_3$, implying that $\pi_{m,m}(q)$ is unimodal with the mode at $\lfloor q_3 \rfloor < m$.

Proposition 2.3.1 shows that, depending on the number of vehicles m relative to the customer arrival rate λ, the system can be in one of three regimes: (1) $m < \gamma_1(\lambda)$ (we refer to this regime as the *supply-limited* regime), (2) $\gamma_1(\lambda) < m < \gamma_2(\lambda)$ (we refer to this regime as the *intermediate* regime), and (3) $m > \gamma_2(\lambda)$ (we refer to this regime as the *supply-rich* regime).

In the supply-limited regime, the probability $\pi_{m,m}(q)$ is increasing in q. This implies that the more likely states are those where the system is *critically loaded*, with the number of customers in the system being close to the number of vehicles. An arriving customer is most likely to either be rejected or to experience a long pick-up time. In the intermediate regime, the distribution of the number of customers in the system is bimodal, with one mode at $\lfloor q_1 \rfloor < m$ and the other at m. The number of customers "oscillates" between these two modes, suggesting different experiences for customers at different times. Customers who arrive during periods when the system state is around $\lfloor q_1 \rfloor$ are served and experience relatively short pick-up times. Customers who arrive during periods when the number of customers in the system is close to m are either rejected or experience long pick-up times. In the supply-rich regime, the distribution of the number of busy vehicles is unimodal,

[10] Systems where m and λ are small can be analyzed using the same approach, though certain regimes identified in Proposition 2.3.1 may no longer exist. Given that small systems hold less significance, we choose to simplify the presentation and omit their analysis for brevity.

[11] Note that q_1, q_2, and q_3 depend on m and λ. For notational compactness, we do not express this dependence explicitly.

with the mode at $\lfloor q_3 \rfloor < m$. In this regime, customers are more likely to be served and experience relatively short pick-up times.

Remark: Depending on the functional form of $\mu(m, q)$, it may be possible to obtain explicit expressions for $\gamma_1(\lambda)$ and $\gamma_2(\lambda)$ with $0 < \gamma_1(\lambda) < \gamma_2(\lambda)$. For example, in Appendix A.4 (see (A.5)), we show that this is the case when

$$\mu(m, q) = \left[\frac{s}{\sqrt{m - q + 1}} + s \right]^{-1}. \tag{2.4}$$

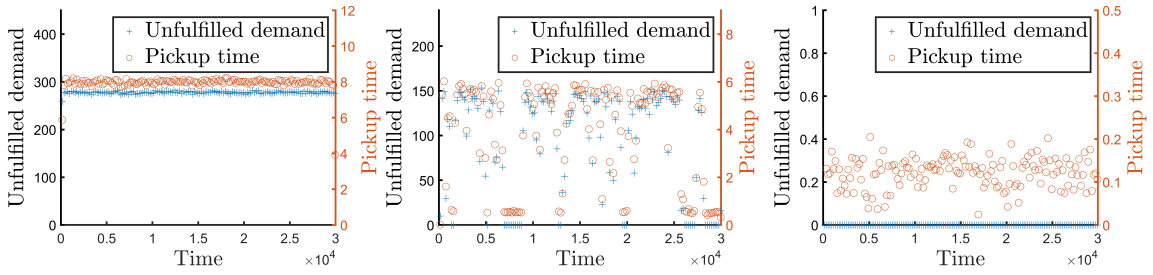
This functional form of service rate is adopted by Besbes et al. (2022) who show that it is asymptotically exact for a setting similar to ours; see also Section 2.3.3 for additional related discussion and details.

To illustrate the temporal dynamics in each of the regimes (and to validate the intuition above), we provide simulation results in Figure 2.2 for an example system where the customer arrival rate is $\lambda = 500$ per minute, and the state-dependent service rate is given by (2.4) with $s = 10$. We simulate the corresponding $M/M(q)/l/l$ queueing system [12] starting in state 0 (i.e., starting with an empty system with no customers in service). The simulation generates sample paths of customer demand not fulfilled (averaged every 200 minutes) and pick-up time (averaged every 200 minutes) over a time period of $30,000$ minutes. Panels (a), (b), and (c) illustrate respectively the temporal dynamics for a supply-limited regime ($m = 4000$), an intermediate regime ($m = 5600$), and a supply-rich regime ($m = 12000$). As we can see, in the supply-limited regime (panel (a)), both pick-up time and unfulfilled demand are, over time, consistently high. In the intermediate regime (panel (b)), pick-up time and unfulfilled demand oscillate between high and low values [13]. In the supply-rich regime (panel (c)), both pick-up times and unfulfilled demand are negligible.

Besbes et al. (2022) make related observations for a similar system (see Proposition 1 in Besbes et al. (2022)). Our results are different in that: (1) we focus on finite systems while Besbes et al. (2022) analyze systems in heavy traffic, (2) we consider systems with impatient

[12] We also simulate systems in which vehicles travel at a constant speed between locations using both the L_1 and L_2 distance metrics in service regions with different geometries. These simulations yield the same outcomes.

[13] A similar phenomenon was observed empirically by Castillo et al. (2021) using data from the ride-hailing service Uber, with pick-up times and fulfilled demand varying significantly even when the number of vehicles stays the same.



(a) $m = 4000$. (b) $m = 5600$. (c) $m = 12000$.

Figure 2.2: The impact of vehicle supply capacity on pick-up times and unfulfilled demand.

customers while they consider systems with patient customers, and (3) we characterize different regimes by vehicle capacity while they use scaling parameters that capture the rate under which the nominal workload relative to the server capacity approaches 1.

2.3.2 The System with Fewer (Remote) Drivers than Vehicles ($n \leq m$)

In this section, we provide analysis for the case with remote drivers (i.e., $n \leq m$) and examine how the introduction of remote drivers, where drivers may be fewer than vehicles, impacts system dynamics and performance. We begin by describing how the results in Proposition 2.3.1 (the features of the distribution of the number of customers) are affected. Recall that $\gamma_1(\lambda)$, $\gamma_2(\lambda)$, q_1, q_2, and q_3 are introduced in Proposition 2.3.1.

Proposition 2.3.2. *The probabilities $\pi_{m,n}(q)$ satisfy the following properties:*

(i) When $m < \gamma_1(\lambda)$, $\pi_{m,n}(q)$ increases in q for $q \leq n$, implying that $\pi_{m,n}(q)$ for $q \in \{0, \cdots, n\}$ is unimodal with the mode at n.

(ii) When $\gamma_1(\lambda) < m < \gamma_2(\lambda)$, depending on n, we have:

(ii.i) if $n < q_1$, $\pi_{m,n}(q)$ increases in q for $q \leq n$, implying that $\pi_{m,n}(q)$ is unimodal with the mode at n;

(ii.ii) if $q_1 < n < q_2$, $\pi_{m,n}(q)$ increases in q for $q < q_1$ and decreases in q for $q_1 < q \leq n$, implying that $\pi_{m,n}(q)$ is unimodal with the mode at $\lfloor q_1 \rfloor$;

(ii.iii) if $n > q_2$, $\pi_{m,n}(q)$ increases in q for $q < q_1$, decreases in q for $q_1 < q < q_2$, and then increases in q for $q_2 < q \leq n$, implying that $\pi_{m,n}(q)$ is bimodal with one mode at $\lfloor q_1 \rfloor$

and the other at n.

(iii) When $m > \gamma_2(\lambda)$, depending on n, we have:

(iii.i) if $n < q_3$, $\pi_{m,n}(q)$ increases in q for $q \leq n$, implying that $\pi_{m,n}(q)$ is unimodal with the mode at n;

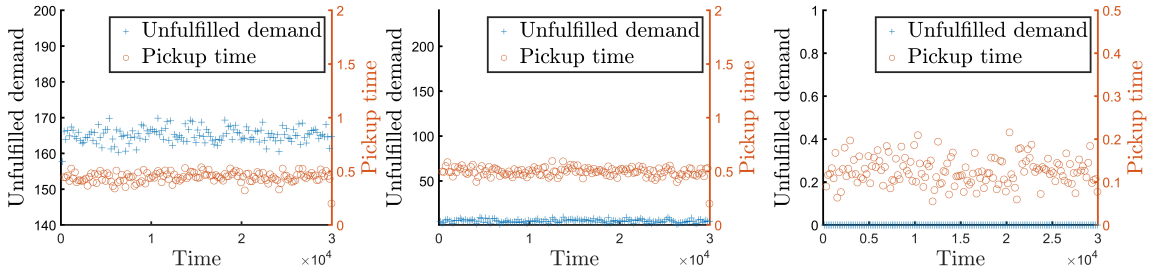
(iii.ii) if $n > q_3$, $\pi_{m,n}(q)$ increases in q for $q < q_3$ and decreases in q for $q_3 < q \leq n$, implying that $\pi_{m,n}(q)$ is unimodal with the mode at $\lfloor q_3 \rfloor$.

Proposition 2.3.2 shows that, depending on the number of vehicles, the system can again be in one of three similarly defined regimes (i.e., the thresholds on vehicle capacity are the same) as those defined in Proposition 2.3.1 for a system with an equal number of vehicles and drivers. We again refer to these three regimes as a supply-limited regime, an intermediate regime, and a supply-rich regime. Although the regimes are similarly defined, there are important differences in the features of the distribution of the number of customers (or equivalently busy vehicles) in each regime.

In the supply-limited regime, the distribution of the number of customers in the system is again unimodal. However, the mode is now at n. The more likely states are those where the number of customers in the system is close to the number of drivers. However, pick-up times in those states are now shorter because the number of vehicles available is always greater than or equal to $m - n$ (which is strictly positive when $n < m$). Because pick-up times are shorter, this suggests that enough driver capacity may be freed up (as long as the number of drivers is not too small) to allow for more demand to be fulfilled.

In the intermediate regime, we observe three sub-regimes with respect to n. If the ratio of driver capacity to vehicle capacity is low, the distribution of the number of customers in the system is unimodal, with the mode at n. If the ratio of driver capacity to vehicle capacity is moderate, the distribution of the number of customers in the system is unimodal, with the mode at $\lfloor q_1 \rfloor < n$. If the ratio of driver capacity to vehicle capacity is high, the distribution of the number of customers in the system is bimodal, with one mode at $\lfloor q_1 \rfloor$ and the other at n. Therefore, compared to Proposition 2.3.1, having remote drivers that are fewer than vehicles can change the number of modes.

In the supply-rich regime, we observe two sub-regimes with respect to n. If the ratio of driver capacity to vehicle capacity is low, the distribution of the number of customers in the system is unimodal, with the mode at n. If the ratio of driver capacity to vehicle



(a) $m = 4000$, $n = 3500$. (b) $m = 5600$, $n = 5250$. (c) $m = 12000$, $n = 8000$.

Figure 2.3: The impact of vehicle supply capacity and remote driver capacity on pick-up times and unfulfilled demand.

capacity is high, the distribution of the number of customers in the system is unimodal, with the mode at $\lfloor q_3 \rfloor$. In the supply-rich regime, having fewer drivers can shift the mode of the distribution from $\lfloor q_3 \rfloor$ to n. However, as long as $n > \lfloor q_3 \rfloor$, the more likely states are those around $\lfloor q_3 \rfloor$. Because the service rate only depends on the number of idle vehicles in the system (as long as there are some idle drivers), this suggests that it may be possible to reduce the number of drivers without significantly impacting system performance.

To illustrate the temporal dynamics in each of the regimes (and to validate the intuition above), we provide simulation results in Figure 2.3 using the same parameters and procedure as those used for the results shown in Figure 2.2 for a system with an equal number of vehicles and drivers. Panels (a),(b) and (c) illustrate respectively the temporal dynamics for a supply-limited regime ($m = 4000$ and $n = 3500$), an intermediate regime ($m = 5600$ and $n = 5250$), and a supply-rich regime ($m = 12000$ and $n = 8000$). As we can see, in the supply-limited regime, the system with remote drivers loses less demand and the pick-up time is shorter (panel (a) in Figure 2.3) than a system with an equal number of vehicles and drivers (panel (a) in Figure 2.2). In the intermediate regime, the system with remote drivers eliminates the oscillation patterns (panel (b) in Figure 2.2) and has consistently lower pick-up time and unfulfilled demand (panel (b) in Figure 2.3). In the supply-rich regime, the system with remote drivers maintains a similar pick-up time and level of unfulfilled demand (panel (c) in Figure 2.3) as the system with an equal number of vehicles and drivers (panel (c) in Figure 2.2).

We next present a set of results (Theorems 2.3.1.A–2.3.1.C) regarding service level in a system with remote drivers. Recall that we use $SL(m, n)$ to denote the service level in a system with m vehicles and n remote drivers.

Theorem 2.3.1.A. *For $m < \gamma_1(\lambda)$, if*

$$\lambda SL(m, m) > m\mu(m, m), \tag{2.5}$$

then there exists a threshold $\tilde{n}_1 \leq m-1$ such that $SL(m, n)$ is increasing in n when $n \leq \tilde{n}_1$ and is decreasing in n when $n > \tilde{n}_1$. Otherwise, $SL(m, n)$ is non-decreasing in n.

Theorem 2.3.1.A shows that, under the well-specified condition (2.5) (more about this condition later), service level may not be monotonic in the number of drivers. More importantly, Theorem 2.3.1.A shows that conditions exist under which reducing the number of drivers improves service level (i.e., having fewer drivers can improve service level). The fact that the service level may not be monotonic in the number of drivers n implies that there is an optimal number of drivers under which the service level is maximized (this optimal number of drivers can be obtained via a simple line search).

Condition (2.5) can be elucidated as follows. The right-hand side of the inequality represents the rate at which trips are completed when all drivers are occupied while the left-hand side represents the rate at which trip requests are fulfilled or, equivalently, the average trip completion rate over time. This condition suggests that when the trip completion rate under full driver occupancy is lower than the average rate, the platform can benefit from reducing the number of drivers. Condition (2.5) captures the trade-off between reduced service capacity and shorter service times. Note that in a standard multi-server queue where the service rate is not state-dependent (i.e., one where $q\mu(m, q) = q\mu$), the inequality does not hold. This implies that Assumption 2.3.2 (which states that $q\mu(m, q)$ first increases then decreases in q) is pivotal to the result that reducing the number of drivers relative to the number of vehicles can actually improve service level.

In Figure 2.4, we consider an example system where the state-dependent service rate is given by (2.4) (i.e., $\mu(m, q) = \left[\frac{s}{\sqrt{m-q+1}} + s\right]^{-1}$). In this case, condition (2.5) reduces to one that depends only on the number of vehicles m and the system workload λs. Figure 2.4 illustrates the range of parameter values under which the condition holds. The results

suggest that (2.5) holds in the supply-limited regime as long as m is not too small.

To illustrate the efficiencies gained by operating a system with remote drivers, we provide in Figure 2.5 numerical results for an example system where $\mu(m,q) = \left[\frac{s}{\sqrt{m-q+1}} + s\right]^{-1}$, $\lambda = 10$ and $s = 10$. We vary the vehicle capacity (m) and, in each case, choose the number of drivers that maximizes the service level. Panel (a) shows both the percentage and absolute (relative to the benchmark system where $n = m$) improvement in service level as the number of vehicles is varied. Panel (b) shows both the percentage and absolute reduction in the number of drivers as the number of vehicles is varied. In Section 2.3.3, we provide sharper analytical results by considering the asymptotic regime of very large demand.

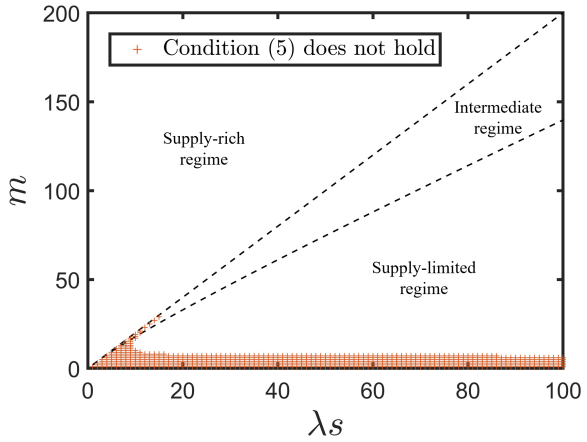


Figure 2.4: An illustration of the parameter range for Condition (2.5) to hold in the supply-limited regime and the intermediate regime

Theorem 2.3.1.B provides an analogous result to Theorem 2.3.1.A for the intermediate regime.

Theorem 2.3.1.B. *For $\gamma_1(\lambda) < m < \gamma_2(\lambda)$, if (2.5) holds, then there exists a threshold $\tilde{n}_2 \leq m-1$ such that $SL(m,n)$ is increasing in n when $n \leq \tilde{n}_2$ and is decreasing in n when $n > \tilde{n}_2$. Otherwise, $SL(m,n)$ is non-decreasing in n.*

Theorem 2.3.1.B shows that, for the intermediate regime, condition (2.5) is again sufficient and necessary for service level to be non-monotonic in the number of drivers and for

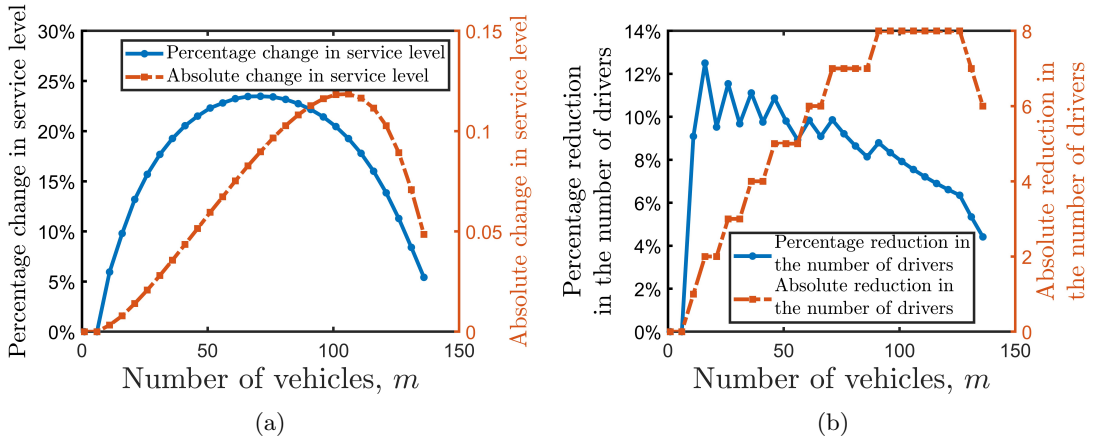


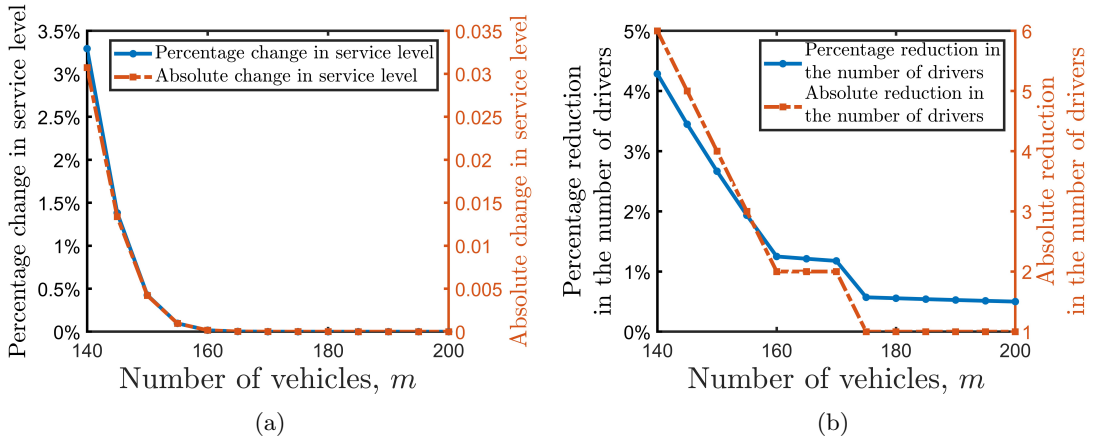
Figure 2.5: The impact of using remote drivers on system performance in the supply-limited regime.

reducing the number of drivers to improve service level. Figure 2.4 shows that condition (2.5) holds (when $\mu(m,q) = \left[\frac{s}{\sqrt{m-q+1}} + s\right]^{-1}$) in the intermediate regime as long as either m or λs is not too small. Figure 2.6 provides numerical results (for the same example considered for Figure 2.5) illustrating gains that can be obtained by using remote drivers (relative to a system with $n = m$). The results show that these gains are less significant compared to those observed in the supply-limited regime. In Section 2.3.3, we provide analytical support for these observations.

For the supply-rich regime, we have the following result.

Theorem 2.3.1.C. *For $m > \gamma_2(\lambda)$, $SL(m,n)$ is increasing in n. Moreover, for any $n > q_3$, where q_3 is introduced in Proposition 2.3.1, $SL(m,n)$ is lower bounded by $1 - \frac{1}{n-\lfloor q_3 \rfloor}$.*

The first part of Theorem 2.3.1.C shows that, in the supply-rich regime, reducing the number of drivers always reduces the service level. The second part of the theorem shows that, when the number of drivers is sufficiently large (i.e., $n > q_3$), the service level is lower-bounded by a function that is increasing concave in n and has the form $1 - \frac{1}{n-\lfloor q_3 \rfloor}$. This suggests that the number of drivers may have a diminishing effect and that there may be an opportunity to reduce the number of drivers without significantly reducing the service level. This observation is supported by the numerical results in Figure 2.7. In the



(a) (b)

Figure 2.6: The impact of using remote drivers in the intermediate regime

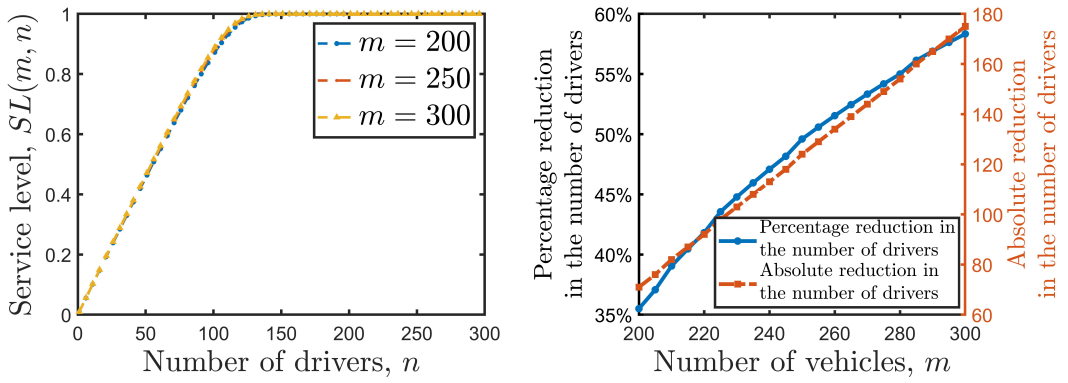
next section, we provide analytical support for these observations.

2.3.3 Asymptotic Analysis

In this section, we consider the asymptotic regime where the demand approaches infinity and the number of vehicles scales proportionally. Such a setting, corresponding to a *large market*, may be of particular relevance to a service that operates in a large dense urban area. In what follows, we show that, in such a setting, it is possible to provide a sharper characterization of the impact of using remote drivers on service level.

To carry out this analysis, we rely on the asymptotically-optimal approximation of service rate proposed by Besbes et al. (2022) referenced in Section 2.3.1: $\frac{1}{\mu(m,q)} = \frac{s}{\sqrt{m-q+1}} + s$, where s corresponds the expected travel time between two uniformly drawn locations in the service region $\mathcal{C}$. The term $\frac{s}{\sqrt{m-q+1}}$ captures the pick-up time and the term s captures trip time. Note that this approximation of service rate satisfies Assumption 2.3.1 and 2.3.2. Moreover, as mentioned earlier, it is asymptotically exact (up to a geometry-dependent coefficient) when the origin-destination pairs associated with each requested trip is uniformly distributed over $\mathcal{C}$ (see Lemma 1 in Besbes et al. (2022) for validation and detailed discussions).

We provide results for each of the supply regimes considered in the previous sections.



(a) (b)

Figure 2.7: The impact of using remote drivers in the supply-rich regime (panel (b) shows both the maximum percentage and maximum absolute reductions in the number of drivers that guarantee a reduction in service level of less than 0.01 (relative to a system with $n = m$)).

Note that because we consider proportional scaling of demand and number of servers such that $m_\lambda = \lfloor \alpha\lambda s \rfloor$ for $\alpha > 0$, the conditions that correspond to different supply regimes have now much simpler forms. In particular, because $\lim_{\lambda\to\infty} \frac{\lambda s}{\gamma_1(\lambda)} = 1$ and $\lim_{\lambda\to\infty} \frac{\lambda s}{\gamma_2(\lambda)} = \frac{1}{2}$, the system is in the supply-limited regime if $\alpha \in (0,1)$, is in the intermediate regime if $\alpha \in (1,2)$, and is in the supply-rich regime if $\alpha > 2$ (See Appendix A.4 for detailed analysis).

We consider first the supply-limited regime.

Theorem 2.3.2.A. *Let* $m_\lambda = \lfloor \alpha\lambda s \rfloor$ *for* $\alpha \in (0,1)$, *then* $\liminf_{\lambda\to\infty} \left[\max_{n\in\{1,\ldots,m_\lambda\}} \frac{SL(m_\lambda,n)}{SL(m_\lambda,m_\lambda)} \right] \geq 2-\alpha$.

Theorem 2.3.2.A shows that, in the supply-limited regime, a system with remote drivers (when the number of drivers is chosen to maximize service level) improves service level (relative to a system with an equal number of drivers and vehicles) asymptotically by at least a factor of $2-\alpha$ ($\alpha \in (0,1)$). The bound on the improvement decreases in α and vanishes as α approaches 1, which is consistent with the result we obtain below for the intermediate regime. In Appendix A.4.4, we provide numerical results comparing the asymptotic bound and the service level ratio derived from the simulation of a finite system.

Theorem 2.3.2.B. *Let* $m_\lambda = \lfloor \alpha\lambda s \rfloor$ *for* $\alpha \in (1,2)$, *then* $\lim_{\lambda\to\infty}\left[\max_{n\in\{1,\ldots,m_\lambda\}} \frac{SL(m_\lambda,n)}{SL(m_\lambda,m_\lambda)}\right] = 1.$

Theorem 2.3.2.B shows that, in the intermediate regime, the relative difference between a system with remote drivers (with the number of drivers chosen optimally to maximize service level) and a system with an equal number of drivers and vehicles vanishes asymptotically. The result in Theorem 2.3.2.B can be explained by the fact that, as λ approaches infinity, the more likely states (in a system with an equal number of drivers and vehicles) are those where the system is non-critically loaded (states around $\lfloor q_1 \rfloor$). In fact, in this case $\pi_{m_\lambda,m_\lambda}(\lfloor q_1 \rfloor)$ dominates (with a higher order) $\pi_{m_\lambda,m_\lambda}(m_\lambda)$.

Lastly, we consider the supply-rich regime in Theorem 2.3.2.C.

Theorem 2.3.2.C. *Let* $m_\lambda = \lfloor \alpha\lambda s \rfloor$ *for* $\alpha > 2$ *and* $n^*_\lambda = \min\{n : SL(m_\lambda, m_\lambda) - SL(m_\lambda, n) < \epsilon\}$ *for* $\epsilon > 0$. *Then,* $\limsup_{\lambda\to\infty} \frac{n^*_\lambda}{m_\lambda} \leq \frac{1}{\alpha}$ [14].

Theorem 2.3.2.C shows that, under the supply-rich regime, a system with remote drivers can reduce the number of drivers (relative to a system with an equal number of drivers and vehicles) asymptotically by at least a factor of $\frac{1}{\alpha}$ (ϵ can be arbitrarily small). Since $\alpha > 2$ in this case, this means that the number of drivers can be reduced by at least half. In Appendix A.4.4, we provide numerical results comparing the asymptotic bound and the actual driver-to-vehicle ratio derived from the simulation of a finite system.

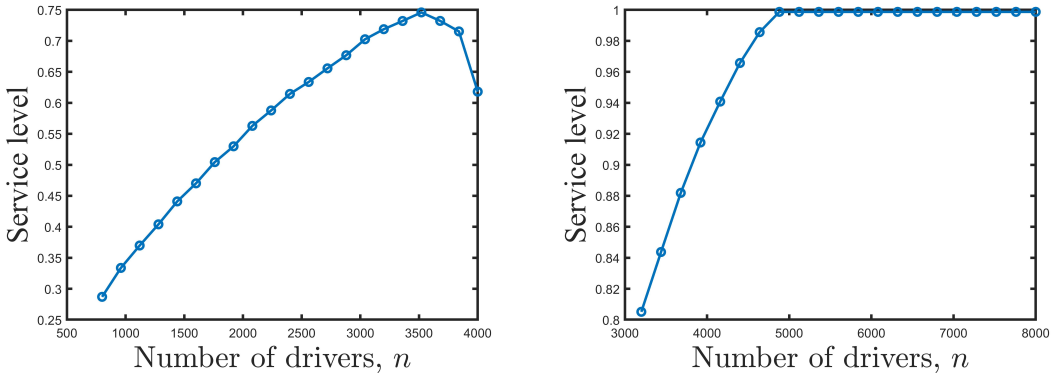
2.3.4 Numerical Results using Data from New York City

In this section, we provide numerical results where the demand process, pick-up times, and trip times are determined based on real world data. The data we use is from the New York City Taxi & Limousine Commission (TLC) [15]. The data contains GPS coordinates for the pick-up and drop-off locations of all yellow cab trips over multiple years and the associated pick-up and drop-off times (note that the pick-up time in the TLC data refers to the time when the customer is picked up by the taxi, which is different from the pick-up time defined in this paper). In estimating pick-up times, we rely on the city's road network and prevailing travel speeds for the times considered. The data set we use covers

[14] Note that the bound does not depend on ϵ as $\lim_{\lambda\to\infty}[SL(m_\lambda, m_\lambda) - SL(m_\lambda, \lfloor(1+\delta)q_3\rfloor)] \to 0$ for any $\delta > 0$ when $\alpha > 2$. See the Proof of Theorem 2.3.2.C in Appendix A.4.3 for details.

[15] `https://www1.nyc.gov/site/tlc/about/data.page`

the periods of June 2015 to August 2015. More information on the data and the numerical procedure can be found in Appendix A.7.1–A.7.2.



(a) Supply-limited regime

(b) Supply-rich regime

Figure 2.8: Results from numerical experiments based on TLC data.

Figure 2.8 provides representative results (based on data from 06/16/2015). In particular, results for two scenarios for the number of vehicles are shown, $m = 4,000$ (panel (a)) and $8,000$ (panel (b)), corresponding respectively to the supply-limited and supply-rich regimes. The results are consistent with those in Section 2.3.2. In the supply-limited regime, a system where the number of remote drivers is appropriately selected can improve the service level by 20.7%. In the supply-rich regime, a system with remote drivers can reduce the number of drivers by 42% (relative to the number of vehicles) while maintaining roughly the same service level. Additional numerical results can be found in Figure A.6 in Appendix A.7.3. A more extensive set of results is also available upon request.

2.4 Systems with Patient Customers

In this section, we consider the case where customers are patient. That is, we consider the setting where customers are willing to wait until both a vehicle and driver are available to serve them. We assume that customers are served on a first-come first-served (FCFS) basis and are matched with the closest idle vehicle, a so-called *first dispatch policy* (see for example Castillo et al. (2021)).

The total amount of time a customer spends in the system has now three components: (1) time waiting for a vehicle and driver to be dispatched, (2) time it takes a vehicle to pick up a customer once it is dispatched, and (3) time it takes a vehicle to complete the trip (from the pick-up location to the requested destination) with the customer onboard. We refer to the first component as *waiting time*, and continue to refer to the second and third as pick-up time and trip time, respectively. Hence, the customer's total time in the system is now the sum: waiting time + pick-up time + trip time.

As with the case of impatient customers, we approximate the dynamics of the system by those of a multi-server queueing system with state-dependent service time but now with an infinite waiting room, namely an $M/M(q)/n$ queue. We assume the state-dependent service rate function, $\mu(m, q)$, satisfies Assumption 2.3.1 and 2.3.2.

Recall that we define $\rho(q) = \frac{\lambda}{q\mu(m,q)}$. A condition for the system to be stable (a queueing system is said to be stable if its long run average over time of the number of customers in the system exists and is finite) is given by the following lemma.

Lemma 2.4.1. *The system is stable if and only if* $\rho(n) < 1$.

Let $\pi_{m,n}(q)$ denote the stationary probability of having q customers in the system given that the system has m vehicles and n remote drivers. These probabilities can be obtained by solving the underlying Markov chain, a birth and death process with birth rate λ and death rate $\min\{q, n\}\mu(m, \min\{q, n\})$ when the system is in state q. Given that the stability condition in Lemma 2.4.1 is satisfied,

$$\pi_{m,n}(0) = \left[1 + \sum_{i=1}^{n} \prod_{k=1}^{i} \rho(k) + \frac{\rho(n)}{1-\rho(n)} \prod_{k=1}^{n} \rho(k)\right]^{-1}, \tag{2.6}$$

$$\pi_{m,n}(q) = \pi_{m,n}(0) \prod_{k=1}^{q} \rho(k), \quad \text{for} \quad q \in \{1, \cdots n\}, \quad \text{and} \tag{2.7}$$

$$\pi_{m,n}(q) = [\rho(n)]^{q-n}\, \pi_{m,n}(0) \prod_{k=1}^{n} \rho(k), \quad \text{for} \quad q > n. \tag{2.8}$$

Let $W(m, n)$ refer to the expected delay a customer experiences (i.e., the expected time a customer waits before her trip begins) in a system with m vehicles and n drivers. Then,

by virtue of Little's law,

$$W(m,n) = \frac{\sum_{q=0}^{\infty} q\pi_{m,n}(q)}{\lambda} - s, \tag{2.9}$$

where s is the expected travel time between two uniformly drawn locations in the service region $\mathcal{C}$.

Before we proceed with the analysis of a system with remote drivers, let us first consider the benchmark case of a system with an equal number of drivers and vehicles (i.e., a system with $n = m$). When $n = m$, the stability condition in Lemma 2.4.1 can be restated per the following corollary.

Corollary 2.4.1. *A system with $m = n$ is stable if and only if $m > \gamma_2(\lambda)$, where $\gamma_2(\lambda)$ is defined in Proposition 2.3.1.*

Corollary 2.4.1 shows that, when $m = n$, the system is stable only in the supply-rich regime. The following proposition contrasts this result with the result for a system with remote drivers ($n \leq m$). Recall that q_1, q_2, and q_3 are introduced in Proposition 2.3.1.

Proposition 2.4.1. *For a system with $n \leq m$, the following holds:*

(i) when $m < \gamma_1(\lambda)$, the system is unstable for any n;

(ii) when $\gamma_1(\lambda) < m < \gamma_2(\lambda)$, the system is stable if and only if $q_1 < n < q_2$; and

(iii) when $m > \gamma_2(\lambda)$, the system is stable if and only if $n > q_3$.

Proposition 2.4.1 shows that in the supply-limited regime, a system with remote drivers is never stable regardless of the number of drivers hired. A system in the supply-rich regime is stable if and only if the number of drivers is sufficiently high (above q_3). A system in the intermediate regime is stable if and only if the number of drivers is relatively moderate ($q_1 < n < q_2$). This result, perhaps surprisingly, indicates that a system that would otherwise be unstable becomes stable by reducing the number of drivers, made possible by the introduction of remote drivers. This is the case in the intermediate regime. The result can be explained again as follows. Reducing the number of drivers increases the number of idle vehicles which shortens pick-up times. When the supply of vehicles is relatively limited, this can lead to a net increase in service capacity. Obviously, when the number of vehicles is sufficiently small (the supply-limited regime), the benefit of reducing pick-up times is not sufficient to overcome the lack of drivers, leading the system to be unstable.

Next, we provide a result analogous to the result in Theorem 2.3.1.A – 2.3.1.C, showing that the impact of the number of drivers on expected delay can be non-monotonic. Define $\tilde{q}$ as the unique solution (between 0 and m) to $\frac{d\rho(q)}{dq} = 0$.

Proposition 2.4.2. *The following holds:*

(i) when $\gamma_1(\lambda) < m < \gamma_2(\lambda)$, $W(m,n)$ decreases in n if $q_1 < n < \tilde{q}$ and increases in n if $\tilde{q} < n < q_2$; and

(ii) when $m > \gamma_2(\lambda)$, $W(m,n)$ decreases in n if $q_3 < n < \tilde{q}$ and increases in n if $\tilde{q} < n \leq m$.

Finally, we provide asymptotic results for the large market setting. For simplicity, we abuse notation and let $q^* = q_1$ when the system is in the intermediate regime, and $q^* = q_3$ when the system is in the supply-rich regime. We show that, when the system is stable, the number of customers (or, equivalently, the number of busy vehicles) is concentrated around q^*.

Proposition 2.4.3. *Let $m_\lambda = \lfloor \alpha\lambda s \rfloor$ for $\alpha > 1$, $n_\lambda = \lfloor \beta m_\lambda \rfloor$ for $\beta \in (0,1]$ such that the stability condition in Lemma 2.4.1 is satisfied. Then, we have*

$$\lim_{\lambda\to\infty} \left| W(m_\lambda, n_\lambda) - \left(\frac{q^*}{\lambda} - s \right) \right| = 0.$$

Moreover, For any $\gamma \in (0,1)$, we have

$$\lim_{\lambda\to\infty} \sum_{q\in((1-\gamma)q^*,(1+\gamma)q^*)} \pi_{m,n}(q) = 1.$$

Because the number of customers is concentrated around q^*, we expect that the number of drivers not to be significantly greater than q^* when the number of vehicles is large. We confirm this intuition when $\mu(m,q)$ is approximated by (2.4) and quantify the savings in the number of drivers in the proposition below.

Proposition 2.4.4. *Let $m_\lambda = \lfloor \alpha\lambda s \rfloor$ for $\alpha > 2$ and define*

$$n^*_\lambda = \min\{n : |W(m_\lambda, n) - W(m_\lambda, m_\lambda)| \leq \epsilon\}$$

for any $\epsilon > 0$. Then, $\limsup_{\lambda\to\infty} \frac{n^*_\lambda}{m_\lambda} \leq \frac{1}{\alpha}$.

Proposition 2.4.4 shows, analogously to the case with impatient customers, that the number of drivers can be reduced by at least a half while continuing to maintain a similar customer delay achieved by a system having as many drivers as vehicles.

We conclude this section by noting that one may also consider the case where customers are *imperfectly* patient. That is customers are willing to wait but only up to a threshold (typically customer-dependent) after which they abandon and leave without receiving service. In Appendix A.5.5, we provide numerical results for a system where customers who are not immediately matched with a vehicle and driver abandon after an exponentially-distributed amount of time. As shown in Figure A.2 in the appendix, reducing the number of drivers can improve service level and reduce customer delay when the supply of vehicles is limited. When the supply of vehicles is ample, it is possible to significantly reduce the number of drivers without significantly affecting service level and customer delay.

2.5 Discussion

In this section, we provide additional discussion of aspects of our modeling, analysis, and results.

Comparing Systems with Remote Drivers and Systems with in-Vehicle Drivers. The analysis in the previous sections showed that a system with remote drivers can operate effectively with fewer drivers than vehicles (in the supply-limited regime, performance actually improves with a reduction in the number for drivers; in the supply-rich regime, the number of drivers can be significantly reduced without significantly reducing service level). It is tempting to use these results as a basis to argue that a shift from a *traditional* system with in-vehicle drivers to one where the vehicles are remotely operated will yield service level improvements and cost savings. However, one must be cautious in making such comparisons. For example, while savings on labor cost may be possible, vehicles that are remotely operated may be more expensive (though such costs are likely to decrease over time). Similarly, vehicles that are remotely operated may need to travel at lower speeds, especially early on in the deployment of the technology [16]. In Figure 2.9, we

[16] According to Zhang (2020) and the references therein, with low latency (less than 170 milliseconds), tele-driving can exhibit performance similar to in-vehicle driving. Under these conditions, tele-drivers are able to adjust their driving behavior to compensate for latency. However, with high latency (more than

provide service level comparisons between a system with in-vehicle drivers and a system with remote drivers where the service rate is scaled down by a factor $\zeta \in (0, 1]$. As we can see, depending on the value of ζ, the system with remote drivers may or may not achieve a higher service level (in Appendix A.6.1, we provide a characterization of the value of ζ for which different outcomes are possible). In Appendix A.6.2, we provide numerical results using realistic parameter values for driver and vehicle costs. We find, for the parameter values considered, that a shift from a system with in-vehicle drivers to one with remote drivers leads to substantial increases in profit in both the supply-limited and supply-rich regimes even when the costs of vehicles and drivers are significantly higher for the system with remote drivers.

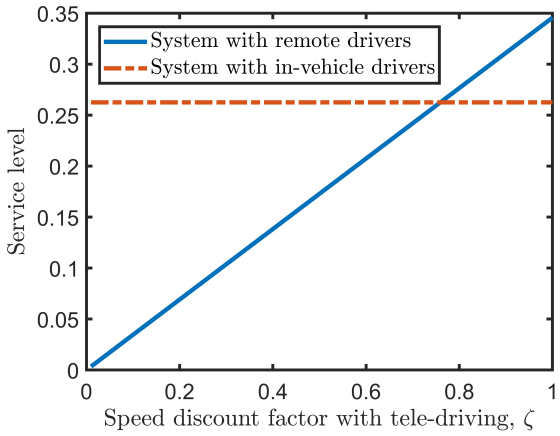


Figure 2.9: Service level comparisons between systems with in vehicle drivers and systems with remote drivers. (Results are shown for $\lambda = 20$ and $m = 100$. For the system with in-vehicle drivers, $s = 10$. For the system with remote drivers, the service rate is scaled down by $\zeta \in (0, 1]$. Given each ζ, we select the number of drivers n^* that maximizes the service level.)

Systems where a Remote Driver Oversees Multiple Vehicles. In this paper, we consider settings where a driver is fully dedicated to a single vehicle when that vehicle is in operation. With further developments in automation technologies, it may become possible for a single driver to oversee the operation of more than one vehicle at a time.

300 milliseconds), tele-drivers tend to adopt a "move-and-wait" approach, which can significantly impact vehicle speed.

Our models and analysis can be easily adapted to a setting where each driver can operate up to k vehicles at a time. In that case the system with n remote drivers corresponds to a multi-server queueing system with kn servers, with the potential reduction in the number of drivers relative to that of vehicles now be more significant.

Moreover, once the automation technology is even more mature, remote drivers may only be needed to intervene in so-called "edge" cases (e.g., scenarios that the automation technology cannot resolve on its own). In this case, a driver need not be assigned to a specific vehicle when it is in operation. Instead, drivers operate as a bank of servers who handle requests for assistance from any of the vehicles currently on the road. Because the rate of such requests is likely to be far smaller in most applications than the rate at which trips occur and because the time it takes to resolve requests for assistance is likely to be much shorter than trip time, the reduction in the number of drivers relative to that of vehicles can be quite significant. Hampshire et al. (2020) consider such a setting and show that for a realistically-parameterized system, the ratio of drivers to vehicles is approximately 1 to 15000. We should note that a standard multi-server queueing model may be adequate in this case as the spatial feature due to the dynamics of pick up time is less prominent.

Mitigating the Wild Goose Chase. The results from the previous sections show that, when the system is in the supply-limited or intermediate regime, having fewer drivers than vehicles can improve performance (e.g., increase the amount of demand fulfilled or reduce customer delay). As explained, this effect is driven by the fact that when the number of available vehicles is low, it might be better to forego demand because fulfilling it would likely involve having vehicles travel long distances to pick up customers (the so-called wild goose chase phenomenon). A system with remote drivers mitigates this by having fewer vehicles than drivers which ensures that the number of available vehicles is always sufficiently high. The benefits of rejecting demand when service capacity is low is a strategy that can also be used when the number of drivers is equal to the number of vehicles. In that case, demand is rejected when the number of available vehicles is sufficiently small (i.e., available vehicles are *strategically idled* when their number is sufficiently small). The dynamics of a system with remote drivers with m vehicles and n drivers can then be replicated by rejecting demand whenever the number of available vehicles is less than or equal to $m-n$. However,

an important difference is that in the case where the number of vehicles equals the number of drivers, when $m-n$ vehicles are idled so are $m-n$ drivers. Depending on how drivers are compensated, idling both vehicle and drivers could carry a higher cost than idling vehicles alone.

Another approach to mitigating the wild goose chase is to apply a *matching radius* in deciding whether or not to assign a vehicle to a customer request. Specifically, a customer-vehicle pair is only matched if the pick-up distance is smaller than the matching radius (see Castillo et al. (2021), Feng et al. (2021), Xu et al. (2020) and Wang et al. (2022) for further discussion). Applying a matching radius can be shown to improve performance regardless of whether the number of vehicles is the same as the number of drivers or not. For a system with remote drivers (whose number is fewer than the number of vehicles), it is possible, depending on the size of the radius, for service level to be non-monotonic in the number of drivers (see the example shown in Figure 2.10). More generally, if the matching radius is made sufficiently small, a system that was in the supply-limited regime could move into the supply-rich regime. In that case, although reducing the number of drivers always reduces service level, the initial decrease in the number of drivers has relatively limited impact on service level (see Figure 2.10 for an illustration of this effect).

Using the Nearest Dispatch Policy in Systems with Patient Customers. An alternative to the first dispatch policy we considered in the analysis of systems with patient customers is the so-called *nearest dispatch* policy (see for example Feng et al. (2021), Besbes et al. (2022) and Wang et al. (2022)). Under the nearest dispatch policy, instead of assigning a driver-vehicle pair to the customer who has been waiting the longest, a driver-vehicle pair is assigned to the customer who is nearest. Under such a policy, the dispatching policy is the same as the first dispatch policy when a customer arrives and there are multiple vehicles waiting. However, it is different when there are multiple customers waiting and a driver becomes available. Specifically, pick up times are now shorter the more there are customers waiting. In other words the service rate $\mu(m, q)$ in this case is decreasing in the number of customers in system, q, when $q \leq n$ and is increasing in q when $q > n$. Put differently, the service rate is sensitive to the thickness of both the demand and supply sides.

It is possible to extend our analysis of systems with patient customers to settings where

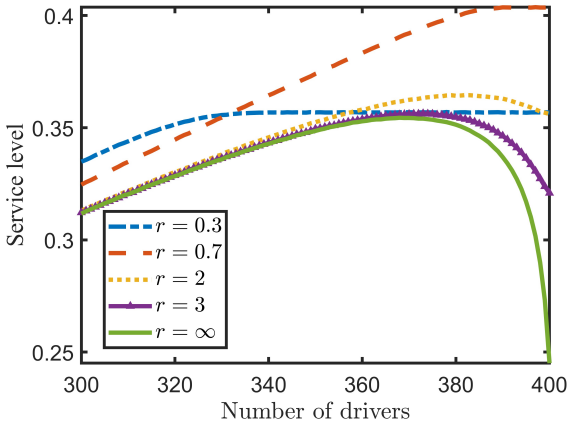


Figure 2.10: The impact of a matching radius on service level (The results pertain to a system where $m = 400$, the service region is a disk with a radius of 15 miles, vehicles maintain a constant speed of 0.5 miles per minute, customers arrive according to a Poisson process with a rate of 120 customers per minute, the origins and destinations of customers are uniformly distributed within the service region, and a customer is rejected if the distance between her origin and the nearest available vehicle exceeds the matching radius r)

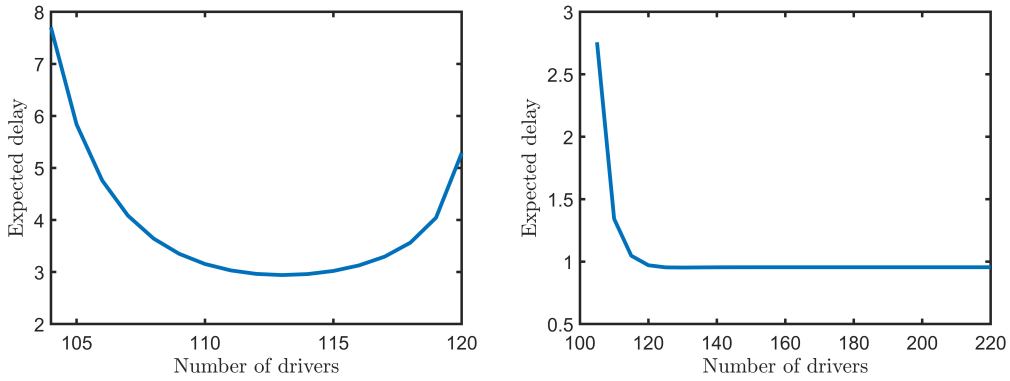
the nearest dispatch policy is used. In this case, the requirements placed on $\mu(m, q)$ are as follows:

- there is a function $\bar{\mu}(m, q)$ that satisfies Assumptions 2.3.1 and 2.3.2, such that $\mu(m, q) = \bar{\mu}(m, q)$ when $q \leq n$,
- $\mu(m, q)$ strictly increases in q when $q > n$, and
- $\lim_{q\to\infty} \mu(m, q) = \frac{1}{s}$ (i.e., pick-up time approaches zero as the number of customers waiting in the queue approaches infinity).

Lemma 2 below provides a condition for the system to be stable (the proof and additional discussion can be found in Appendix A.6.3) .

Lemma 2.5.1. *Under the nearest dispatching policy, the system is stable if and only if* $\frac{\lambda s}{n} < 1$.

The above lemma shows that the system is stable as long as $\frac{\lambda s}{n} < 1$. An implication of this (because $\frac{\lambda s}{n}$ is decreasing in n) is that it is no longer possible for reducing the



(a) Intermediate regime: $m = 120$, $\lambda = 10$ and $s = 10$

(b) Supply-rich regime: $m = 220$, $\lambda = 10$ and $s = 10$

Figure 2.11: The impact of the number of drivers on expected customer delay under the nearest dispatching policy (the results pertain to a system where $\mu(m, q)$ is given by (2.4) if $q \leq n$ and $\mu(m, q) = \left[\frac{s}{\sqrt{(m-n+1)(q-n)}} + s\right]^{-1}$ otherwise [17])

number of drivers to stabilize a system that would otherwise be unstable. However, in a stable system, it is possible to reduce the delay experienced by customers by reducing the number of drivers (i.e., an analogous result to Proposition 2.4.2). This effect is illustrated for an example system in Figure 2.11. As shown in panel (a) of Figure 2.11, expected customer delay can be non-monotone in the number of drivers, i.e., $E[W(m, n)]$ is first decreasing and then increasing in n. Moreover, as shown in panel (b) of Figure 2.11, it continues to be possible, when the supply of vehicles is large, to reduce the number of drivers without significantly increasing delay. In Corollary A.6.4 in Appendix A.6.3, we show that, in the limit, the reduction in the number of drivers is at least a half (i.e., an analogous result to Theorem 2.3.2.C).

[17] The approximation is asymptotically exact (up to a geometry dependent coefficient) when the origin-destination pair associated with each requested trip is uniformly distributed over a square service region. See Theorem 5 in Wang et al. (2022) for the validation and more discussions.

2.6 Concluding Comments

In this paper, we explored the novel concept of tele-driving and examined the extent to which the number of drivers relative to the number of vehicles can be reduced. Among our findings, we showed that, depending on the supply of vehicles relative to the workload, there is an opportunity for a system with remote drivers (appropriately sized) to significantly improve service level or significantly reduce the number of drivers without affecting service level. If customers are patient, we showed that a system with remote drivers can stabilize an otherwise unstable system or significantly reduce the number of drivers while maintaining a similar expected delay. Our analysis uncovered how the tradeoff between shorter service times and more servers, brought about by remote driving, shapes these outcomes.

In this paper, we focused on a single driver of efficiency (the reduction in the number of drivers). As mentioned, there may be other benefits from tele-driving, including eliminating inefficiencies associated with drivers acting strategically (in settings where the drivers can act independently), increasing access to areas that are perceived by drivers as less desirable, and increasing demand by making the service more attractive to certain users (e.g., those who may feel uncomfortable riding with a stranger). There may also be benefits to drivers (e.g., by removing the requirement that drivers own a vehicle, increasing labor participation as drivers may work remotely at locations that are most convenient to them, and reducing income variability because of the reduction in the spatial mismatch between supply and demand). Quantifying these benefits, though each is likely to require a different approach, could provide interesting avenues for future research.

Chapter 3

The Impact of Automation on Workers when Workers are Strategic: The Case of Ride-Hailing

3.1 Introduction

There is an ongoing debate, both in the public sphere and among scholars in various fields, as to whether the introduction of automation is harmful or beneficial to workers. As pointed out by the growing economics literature in this area (e.g., Korinek and Stiglitz (2020) and Jackson and Zafer (2019)), there are at least two counteracting forces at play: (1) a displacement effect (automation replacing workers) and (2) a productivity enhancement effect (automation making workers more efficient)[1]. For low-skill workers, the displacement effect usually outweighs productivity enhancement, leading to a decrease in worker welfare (Guo (2022)). This literature assumes that workers have no discretion in how they carry

[1]Other effects include, among others, a demand expansion effect (by making products more affordable, automation increases the demand for these products and the labor involved), a job creation effect (automation makes new businesses and new jobs possible), and an automation deepening effect (automation makes previously deployed automation more productive); see Acemoglu and Restrepo (2018a) for a detailed discussion and references therein.

their work and do not act strategically. In this paper, motivated by the behavior of drivers in ride-hailing (individual drivers decide whether or not to work based on the offered wage and where to locate themselves in anticipation of future fares), we study a setting where workers act strategically. Specifically, we examine, in the context of a ride-hailing service, the impact of automation on workers when the automation only substitutes for workers (i.e., there is displacement but no productivity enhancement) and workers are strategic.

Autonomous vehicle (AV) technology is an exciting new technology, though not fully mature yet, that some have argued is poised to transform the transportation landscape. How this transformation would take place and how it would affect various stakeholders (riders, drivers, and service providers) continues to be a subject of vigorous debate (Iyer and Alton (2019) and Lalley (2017)). A potentially important application of AV technology is ride hailing (the provisioning of transportation services on-demand). The ride hailing industry, which currently relies mostly on independent drivers using conventional vehicles (CVs), has shown a particular interest in AV technology, with several of the leading platforms, such as Uber, Lyft, and Didi, making substantial investments in the research and development of AV technology (Uber (2019), Lyft (2021) and Didi (2021)). However, under most scenarios, it is envisioned that the introduction of AVs will be gradual and that ride hailing platforms are likely to operate initially with a mixed fleet of both AVs and CVs, with the latter owned and operated by human drivers (Iyer and Alton (2019)).

In this paper, we consider a setting where a ride-hailing platform operates a mixed fleet of AVs and CVs. The platform seeks to fulfill transportation requests from customers, who arrive continuously over time, so as to maximize profit. The platform operates over a network consisting of multiple locations. The rate at which customers arrive varies depending on the origin and destination of the requested trips. Customer requests that cannot be immediately matched with a vehicle are considered lost. The platform charges customers a price per unit of travel time. The CVs are driven by independent drivers, who are heterogeneous in their opportunity cost. The platform pays drivers a fixed wage per unit time of service (drivers are paid only when transporting a customer). Drivers decide whether or not to work for the platform depending on the expected earnings from working for the platform and their outside options. The platform incurs a fixed cost for purchasing AVs. For simplicity, we assume no travel costs for AVs and CVs, though this

is not necessary for our analysis. Our findings remain valid when the travel costs are considered (e.g., the cost of fuel and of driving effort), provided that the travel cost for AVs is lower than or equal to that of CVs (see Table 3 of Bösch et al. (2018)).

Upon completing a trip transporting a customer, vehicles can either stay at the location where the trip terminated or reposition (drive empty) to another location. The repositioning of AVs is under the control of the platform while the repositioning of CVs is in the hands of the drivers who act strategically and reposition to the location that maximizes their expected earnings. When a customer request arises, the platform can also decide on whether to prioritize AVs or CVs in assigning the request to a vehicle. Finally, the platform also decides on how many AVs to acquire and how much to pay the drivers.

To study such a setting, we adopt a fluid model of a stylized network consisting of two locations with asymmetric demands. Such models have been widely used to study the dynamics of ride-hailing (see Braverman et al. (2019), Afèche et al. (2023) and Hosseini et al. (2021) and the references therein). As we discuss in Section 3.3, the network we consider, though simple, captures essential features of more complex settings, including imbalances in customer demand and vehicle supply across locations, strategic behavior on the part of the drivers, and multiple types of operational decisions on the part of the platform.

Among our findings, we show that, although the introduction of AVs can displace drivers and depress wages, there are settings where the introduction of AVs leads to higher wages and more drivers being hired. This is because the presence of AVs can incentivize drivers to choose more efficient relocation strategies, earning the platform more revenue and making the drivers more productive. We refer to this effect as an *incentive effect.* Surprisingly, this effect is present only when the platform prioritizes AVs in making work assignments. This result can be explained by the interplay of two counteracting forces with respect to the human drivers: (1) the introduction of AV provides the platform with an additional source of supply and renders human drivers substitutable; and (2) having access to and control over AVs enables the platform to influence the strategic behavior of CVs, thereby reducing the inefficiency from selfish behavior. In particular, we note the following (the statements will be made precise in subsequent sections of the paper).

- *The displacement effect.* Due to the heterogeneous opportunity cost of drivers, the

platform's marginal cost of recruiting CVs increases in the amount of CVs recruited. With the introduction of AVs, the platform is able to replace some human drivers (those with high opportunity costs) with some AVs, benefiting the platform but harming human drivers.

- *The incentive effect.* Human drivers who act strategically may not always make decisions that are in the best interest of the platform or the system as a whole. Specifically, drivers act selfishly in deciding whether to reposition or not. This can lead to sub-optimal deployment of resources. With the introduction of AVs, the platform has an additional source of supply that can be strategically deployed to influence the decision-making of human drivers, potentially resulting in better outcomes for both the platform and the driver (e.g., making drivers more productive, leading the platform to recruit more of them at higher wages).

Depending on the relative strength of these two forces, different outcomes are possible regarding driver welfare. Our analysis shows that this crucially depends on the driver pool size and the AV purchase cost. In particular, the incentive effect is more likely to dominate the displacement effect when (1) the driver pool size is moderate (if the driver pool size is large, the platform recruits a large number of drivers regardless of drivers' strategic behavior which diminishes the incentive effect; if the driver pool size is small, the recruitment cost for drivers is high which enhances the displacement effect), and (2) the AV purchase cost is moderate (if the AV purchase cost is low, the platform procures a relatively large number of AVs and the displacement effect is strong; if the AV purchase cost is high, the platform procures a small number of AVs and, thus, is less capable of deploying AVs to influence the decision-making of human drivers).

These effects manifest themselves differently depending on whether the platform prioritizes AVs, CVs, or neither. When the platform prioritizes AVs, it can easily incentivize drivers to choose more efficient repositioning strategies via demand allocation (e.g., the platform can decide how much demand is allocated to CVs), hence enhancing the incentive effect. When the platform prioritizes CVs or makes no distinction between AVs and CVs, it either has no influence on driver behavior, or can only influence driver behavior by employing AVs to compete with CVs, which may not be in the platform's best interest. Therefore, the introduction of AVs is a win-win for the platform and for the drivers only

if the platform prioritizes AVs in making work assignments.

The incentive effect of automation, when automation is partial and workers are strategic, is novel and, to our knowledge, has not been previously studied. This effect may carry over to other areas where automation is introduced. In particular, the results of this paper suggest that closer attention should be paid to how the introduction of automation may change the workers' incentives and how this feature may be used not only to improve productivity but also worker welfare.

The rest of the paper is organized as follows. In Section 2, we provide a review of related literature. In Section 3, we describe our model. In Section 4, we analyze the benchmark case of no AVs. In Section 5, we formulate the platform's problem and characterize the resulting equilibrium outcomes when AVs are present and demonstrate the impact of automation on drivers. Proofs for all the results, unless otherwise stated, are included in the Appendices.

3.2 Literature Review

This work is at the intersection of two streams of literature. The first is literature that studies spatial networks where resources move from one location to another in the process of servicing demand that is also spatially distributed. Of particular relevance is literature that is motivated by applications in on-demand transportation services, including ride hailing and vehicle sharing (e.g., bike sharing); see Benjaafar and Hu (2020), Hu (2021), and Freund et al. (2019) for recent reviews. Our work is related to streams within this literature that focus on the operational control of these networks, where control levers include the assignment of resources with customers, the spatial repositioning of resources so as to better match supply with demand, and the shaping of demand, indirectly through pricing or directly through admission control. Some of this literature, particularly as it relates to ride hailing, accounts for the fact that the control of resources is distributed and in the hands of individuals who are strategic in their decision making.

Below we briefly review papers that are most salient to our work. We focus on papers that take, as we do, a queueing network (and its associated fluid model approximation) perspective.

Repositioning. Braverman et al. (2019) consider a vehicle sharing network where a platform controls the repositioning of all vehicles (this is akin to a system with only AVs in

our setting). Using a fluid approximation, they show that the vehicle repositioning problem can be formulated as a linear program which can then be solved efficiently. Moreover, they prove that the optimal solution to this linear program specifies a repositioning strategy that is asymptotically optimal. Afèche et al. (2023) consider, like we do, a fluid model of a two-location, four-route network with strategic drivers. Strategic drivers are not controlled by the platform and make their own repositioning decisions to maximize their earnings (this corresponds to a system with only CVs in our setting). The platform maximizes profit by deciding on how much demand to accept from each location (i.e., the platform has control over admission). They characterize, in equilibrium, both the platform's optimal admission control and the drivers' optimal repositioning. In particular, they show that, under some conditions, it is optimal for the platform to reject demand in the low-demand location in order to incentivize drivers to reposition to the high-demand location. Hosseini et al. (2021) design a state-dependent vehicle repositioning policy based on structural properties of a fluid-based model. They provide numerical evidence showing that this state-dependent policy can outperform static policies. In contrast to these papers, we consider a setting with a mix of CVs and AVs with the repositioning of AVs under the control of the platform. We allow for the wage paid to drivers to be a decision made by the platform. We consider a broader range of decisions the platform can make, including how AVs should be repositioned, how to assign vehicles to incoming requests (vehicle assignment priorities) and how large the AV fleet size should be.

There is extensive literature that deals with vehicle repositioning in non-queueing contexts, including problems with a single period or under multiple discrete periods and without strategic drivers; see for example, Benjaafar et al. (2021b), Akturk et al. (2021), He et al. (2020), and Zhao et al. (2020). A comprehensive review of this literature can be found in Benjaafar et al. (2021b).

Admission control and matching. Özkan and Ward (2020) study the problem of matching customer requests with nearby drivers in the context of a ride hailing network. They use a fluid model approximation and show that a static matching is asymptotically optimal under heavy traffic. Banerjee et al. (2022b) consider a similar problem. They develop a family of state-dependent policies whose performance they show to improve exponentially as the number of vehicles scales to infinity. Kanoria and Qian (2020) study

network control, using levers that include admission control, dispatching, and pricing. They develop a class of control policies that are nearly optimal under certain conditions for the discrete time version of the problem they consider.

Dimensioning. George and Xia (2011) develop exact and approximate solution algorithms to determine the optimal fleet size in a vehicle sharing network where the objective is to maximize system profit. George et al. (2012) derive the exact-order asymptotic growth rate of system throughput as the number of vehicles increases. Benjaafar et al. (2022b) develop an approximation for the number of vehicles in a vehicle sharing network needed to guarantee a specified service level (the fraction of demand fulfilled) at each location. They show that this approximation is optimal under various asymptotic regimes. Besbes et al. (2022) study the problem of optimal service capacity for a ride-hailing system modeled as a single multi-server queue with a state-dependent service rate that account for pick up and travel times. For systems with strategic drivers where the vehicles are under the control of human drivers, service capacity is determined indirectly via the mechanism of the wage paid to drivers; see for example Taylor (2018), Benjaafar et al. (2022a) and Hu and Zhou (2020).

Lastly, our work contributes to the emerging literature on **autonomous vehicles** in transportation. Papers that consider the role of AVs in ride sharing and ride hailing systems include Siddiq and Taylor (2021), Lian and Van Ryzin (2022), Baron et al. (2022), Noh et al. (2021), Castro and Frazelle (2021) and Castro et al. (2023). The focus for many of these papers is the economics of AVs and on examining ownership structure and competition and do not, typically, account for the spatial features of these systems. Mirzaeian et al. (2020) use a state-dependent queueing system to study the impact of AVs on highway traffic. Hampshire et al. (2020) and Benjaafar et al. (2023a) consider AVs that are remotely controlled by human tele-operators. Using queuing models, they show that the use of tele-operators can, under some conditions, significantly reduce the ratio of human operators to vehicles. In the setting they consider the human operators are not strategic.

The second stream of related literature is from various fields that considers the impact of automation on human labor. Perhaps the literature of most relevance is from economics where the study of the effect of automation on labor welfare has been received increasing

attention; we refer the readers to Acemoglu and Restrepo (2018a), Mondolo (2021), Lu and Zhou (2021) and Filippi et al. (2023) for recent reviews. Some of this literature examines the impact of automation on workers empirically (see for examples Autor et al. (2003), Frey and Osborne (2017), Graetz and Michaels (2018), Bessen et al. (2019), Acemoglu and Restrepo (2020), Dauth et al. (2021), Dixon et al. (2021), Rio-Chanona et al. (2021), Guo (2022), and Brynjolfsson et al. (2022)) and some of it provides analytical models (see for example Acemoglu (1998), Benzell et al. (2015), Acemoglu and Restrepo (2018a), Acemoglu and Restrepo (2018b), Korinek and Stiglitz (2020), and Hémous and Olsen (2022)). As we mentioned in Section 3.1, much of this literature focuses on the tension between the various manifestations of the displacement and the productivity enhancement effects. The incentive effect we uncover in this paper does not appear to have been previously considered.

Finally, our work is related to recent literature that studies settings with a hybrid workforce, comprising both traditional employees and independent contractors (see for example (Dong and Ibrahim (2020), Chakravarty (2021), Lobel et al. (2021), He and Goh (2022), Hu (2022) and Castro and Frazelle (2022)). This literature primarily focuses on a firm's challenges pertaining to staffing, demand rationing, and supply prioritization and does not consider scenarios where the independent contractors possess task discretion and behave strategically. (There is literature that studies task discretion among workers but with a single type of employees; see Kostami (2023) and the references therein). An interpretation of the results of this paper is that hiring non-contractual workers, whose work is managed by the firm, can be beneficial to the independent contractors, as it mitigates the inefficiency from them behaving strategically.

3.3 Model Description

Consider a platform that operates a mixed fleet of AVs and CVs and let M and N denote the amount of AVs and CVs purchased and recruited respectively. The platform charges customers a price p per unit of travel time. That is, a customer pays amount pt_{ij} for a trip from location i to location j where t_{ij} is the duration of the trip from location i to location j. Customers generate demand for trips from location i to location j. If a customer arrives at a location and there are no empty vehicles available at that location, the customer leaves the system and the platform does not earn any revenue. The platform pays drivers a wage

w per unit of time the driver spends transporting a customer. Therefore, a driver earns wt_{ij} from serving demand that originates at location i and ends at location j.

We adopt a fluid model of a stylized network consisting of two locations (indexed by 1 and 2) and two cross-location routes (routes from one location to the other). We do not consider within-location routes for simplicity, though this is not necessary for our analysis. Fluid models have been widely used to study the dynamics of ride-hailing (see for example Braverman et al. (2019), Afèche et al. (2023) and Hosseini et al. (2021)) and to extract important qualitative features of these systems. The two-location network we consider, though simple, captures many of the essential features of more complex settings, including imbalances in customer demand and vehicle supply across locations, strategic behavior on the part of the drivers, and multiple types of operational decisions on the part of the platform (see also Afèche et al. (2023) for further discussion and motivation).

Upon completing service (a trip transporting a customer), vehicles can either stay at the location where the service terminated or drive empty to the other location. We denote by q_i^A and q_i^C the *volume* of AVs and CVs respectively queueing at location i and we let $q_i = q_i^A + q_i^C$ denote the sum of the two. When a customer arrives, the platform assigns a vehicle among the available AVs and CVs to the customer according to a specified assignment policy. Let W_i^A and W_i^C denote the delay experienced by AVs and CVs waiting to be matched with customers at location i. Let $\eta^C = (\eta_1^C, \eta_2^C)$ denote the drivers' repositioning strategy, where η_i^C is the probability that a CV drives empty to location j after completing a service that ended at location i, where $i \neq j \in \{1, 2\}$.

We assume that drivers make their own decisions regarding repositioning in order to maximize their earnings. We focus on the case where drivers adopt symmetric strategies. Therefore, we call $\eta^C = (\eta_1^C, \eta_2^C)$ a *CV equilibrium repositioning strategy* if it is the best response for every driver. For the AVs, we define $\eta^A = (\eta_1^A, \eta_2^A)$ similarly. The platform owns the AVs and controls the AVs' repositioning strategy η^A. Let ν_i^A and ν_i^C denote the repositioning rates of AVs and CVs, respectively, from location i to location j.

We consider a continuum of drivers with mass L. The drivers are heterogeneous in their opportunity costs with a uniform distribution over $[0, \overline{w}]$, where $\overline{w}$ is the maximal opportunity cost for drivers. Note that because $w \leq p$, where w is the wage the platform pays drivers per unit time of service, drivers with opportunity costs greater than p never

work for the platform. Therefore, we assume that $\overline{w} = p$. A driver works for the platform only if her expected earning in equilibrium exceeds her opportunity cost. The platform procures AVs at a fixed cost. Let I denote the AV purchase cost amortized over the AV's expected lifetime. We assume that $I \leq p$ (otherwise, the platform dose not procure any AVs).

The platform has several levers at its disposal. First, the platform determines the amount of AVs to purchase and the wage it pays drivers. Second, the platform controls the AV's repositioning. Lastly, the platform decides on the assignment of customer requests to vehicles (e.g., the platform can choose to prioritize AVs, CVs, or neither). We assume that drivers have (or can infer) full information, including the decisions made by the platform, when making their own decisions about whether to work for the platform and whether or not to reposition upon trip completion.

Let Λ_{ij} denote the potential demand rate from location i to location j for $i \neq j \in \{1, 2\}$. Without loss of generality, we assume $\Lambda_{12} < \Lambda_{21}$ and thus we call location 1 the *low-demand location* and location 2 the *high-demand location* (note that when $\Lambda_{12} = \Lambda_{21}$, no repositioning is needed). To avoid trivial cases, we assume that $\Lambda_{ij} > 0$. A demand request is lost if there is no available vehicle at location i upon arrival. We use λ_{ij} to denote the effective demand rate from location i to location j (i.e., the rate of fulfilled demand that originates at location i and ends at location j). Let λ_{ij}^A and λ_{ij}^C denote the demand rate from location i to location j fulfilled by AVs and CVs respectively and $\lambda_{ij} = \lambda_{ij}^A + \lambda_{ij}^C$. Let s_{ij}^A and s_{ij}^C denote, respectively the volume of AVs and CVs in service from location i to location j. By Little's law, $s_{ij}^A = \lambda_{ij}^A t_{ij}$ and $s_{ij}^C = \lambda_{ij}^C t_{ij}$. Let $s_{ij} = s_{ij}^A + s_{ij}^C$ refer to the total volume of vehicles in service from location i to location j. Denote by r_{ij}^A and r_{ij}^C the volume of AVs and CVs repositioning from location i to location j. By Little's law, $r_{ij}^A = \nu_{ij}^A t_{ij}$ and $r_{ij}^C = \nu_{ij}^C t_{ij}$. Let $r_{ij} = r_{ij}^A + r_{ij}^C$. Let s denote the pair (s_{ij}^A, s_{ij}^C), r the pair (r_{ij}^A, r_{ij}^C) and q the pair (q_i^A, q_i^C). We refer to (s, r, q) as the capacity allocation of the system. Let $a_i = \frac{\lambda_{ij}}{\Lambda_{ij}}$ for $j \neq i$ denote the service level (i.e., the fraction of demand that is fulfilled) at location i. Let $F_i = \frac{s_{ij}^A}{s_{ij}}$ for $j \neq i$ denote the fraction of demand that is fulfilled by AVs at location i.

In steady-state, η^A, η^C, and (s, r, q) must satisfy the following steady state fluid model

equations:

$$\Lambda_{ij} a_i = \frac{s_{ij}}{t_{ij}}, \quad i \neq j \in \{1,2\}, \tag{3.1}$$

$$a_i \in [0,1], \quad i \neq j \in \{1,2\}, \tag{3.2}$$

$$\frac{r_{ij}^A}{t_{ij}} = \eta_i^A \frac{s_{ji}}{t_{ji}} F_j, \quad i \neq j \in \{1,2\}, \tag{3.3}$$

$$\frac{r_{ij}^C}{t_{ij}} = \eta_i^C \frac{s_{ji}}{t_{ji}} (1 - F_j), \quad i \neq j \in \{1,2\}, \tag{3.4}$$

$$\Lambda_{ij} a_i F_i = \frac{r_{ji}^A}{t_{ji}} + (1 - \eta_i^A) \frac{s_{ji}}{t_{ji}} F_j, \quad i \neq j \in \{1,2\}, \tag{3.5}$$

$$\Lambda_{ij} a_i (1 - F_i) = \frac{r_{ji}^C}{t_{ji}} + (1 - \eta_i^C) \frac{s_{ji}}{t_{ji}} (1 - F_j), \quad i \neq j \in \{1,2\}, \tag{3.6}$$

$$(1 - a_i) q_i = 0, \quad i \in \{1,2\}. \tag{3.7}$$

$$s_{12} F_1 + s_{21} F_2 + (r_{12}^A + r_{21}^A) + (q_1^A + q_2^A) = M, \quad \text{and} \tag{3.8}$$

$$s_{12}(1 - F_1) + s_{21}(1 - F_2) + (r_{12}^C + r_{21}^C) + (q_1^C + q_2^C) = N. \tag{3.9}$$

Equation (3.1) is a result of Little's law. Equation (3.2) specifies that the service level at location i is within the range $[0,1]$. Equation (3.3) and equation (3.4) are the repositioning flow balance equations for AVs and CVs respectively, i.e., $\nu_i^A = \eta_i^A \lambda_{ji}^A$ and $\nu_i^C = \eta_i^C \lambda_{ji}^C$ for $i \neq j \in \{1,2\}$. Equation (3.5) and equation (3.6) state that the rates of outflow and inflow at location i must be equal for both AVs and CVs. Equation (3.7) guarantees that the demand originating at location i can only be lost if there are no vehicles queueing at location i. Equation (3.8) and equation (3.9) state that the amount of AVs and CVs in service, being repositioned, and queueing must be equal to the fleet size of AVs and CVs respectively.

The quantity of drivers recruited is determined by the wage w the platform pays and drivers' utilization (i.e.,the fraction of time drivers expect to be serving customers). In particular, the supply consists of CVs is determined by the fraction of drivers whose opportunity cost is less than their expected earnings (i.e., the effective wage) $\hat{w}$. Let $\rho = \frac{s_{12}^C + s_{21}^C}{N}$ denote drivers' utilization, then $\hat{w} = \rho w$. Because the fraction of drivers whose opportunity cost is smaller than the effective wage is $\frac{\hat{w}}{\overline{w}}$ and recall that we denote by L the driver pool

size, the supply of drivers (and equivalently CVs) in equilibrium satisfies

$$N = \frac{\hat{w}}{\overline{w}} L. \tag{3.10}$$

We close this section by defining driver welfare, which we denote by DW. We define the driver welfare as the aggregate income of workers net of the opportunity costs. That is,

$$DW = L \int_0^{\overline{w}} \frac{\max(\hat{w} - y, 0)}{\overline{w}} dy = L \frac{\hat{w}^2}{2\overline{w}}. \tag{3.11}$$

In the next two sections, we formulate the platform's problem and characterize the resulting equilibrium outcomes. We do so first for the benchmarks case of no AVs (i.e. no automation). We then consider the case where AVs are present and compare outcomes in both cases, particularly with regard to driver welfare.

3.4 The Platform's Problem: The Case of No AVs

In a system without AVs, the platform's only decision is the wage w it pays drivers. The platform does so to maximize profit. Thus, the platform's problem can be stated as follows:

$$\begin{aligned} \text{(Problem I)} \qquad \max_{w} \quad & \Pi^C = (p - w)(s_{12}^C + s_{21}^C) \\ \text{subject to} \quad & (3.1), (3.2), (3.4), (3.6), (3.7), (3.9), (3.10), \\ & \eta^C \text{ is a CV equilibrium repositioning strategy,} \\ & M = 0 \text{ and } F_k = 0 \text{ for } k \in \{1, 2\}. \end{aligned}$$

Before proceeding further with the solution to the above problem, we make the following observation. The maximum demand (and the associated minimum capacity needed to fulfill this demand) consists of two components. The first component is demand that can be fulfilled without any vehicle repositioning. The maximum demand (for trips from location 1 to 2 and 2 to 1) that could be fulfilled without repositioning is given by $2\Lambda_{12}$ (this is because $\Lambda_{12} < \Lambda_{21}$) and the corresponding minimum amount of vehicles (drivers) needed to service this demand is

$$C_1 = (t_{12} + t_{21})\, \Lambda_{12}. \tag{3.12}$$

Note that while all the demand from location 1 (the low-demand location) to location 2 (the high-demand location) can be fulfilled, an amount $\Lambda_{21} - \Lambda_{12}$ from location 2 to location 1 goes unfulfilled. To fulfill this demand requires that drivers reposition empty vehicles from location 1 to location 2 at rate $\Lambda_{21} - \Lambda_{12}$. The corresponding minimum driver capacity needed to fulfill this demand is given by:

$$C_2 = (\Lambda_{21} - \Lambda_{12})(t_{21} + t_{12}), \tag{3.13}$$

This driver capacity consists of drivers repositioning empty from location 1 to 2 (an amount equal to $(\Lambda_{21} - \Lambda_{12})(t_{12})$) and drivers transporting customers from location 2 to 1 (an amount equal to $(\Lambda_{21} - \Lambda_{12})(t_{21})$).

A graphic illustration of this minimum capacity is provided in Figure 3.1. In the remainder of the paper, we refer to the demand associated with C_1 ($2\Lambda_{12}$) as *type-1* demand and demand associated with C_2 ($\Lambda_{21} - \Lambda_{12}$) as *type-2* demand.

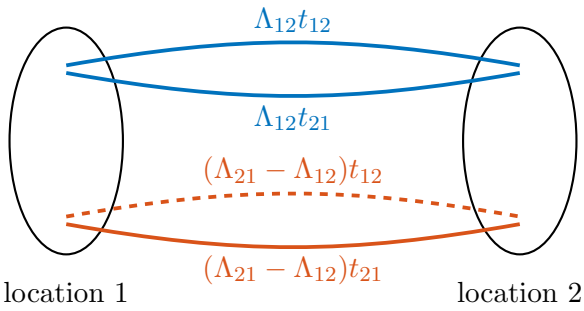


Figure 3.1: An illustration of the minimum capacity needed to fulfill the maximum demand. The orange dashed arc represents capacity associated with repositioning

Next, we let

$$\gamma = \frac{t_{21}}{t_{21} + t_{12}}, \tag{3.14}$$

which has the interpretation of the maximum utilization of drivers who reposition from location 1 to location 2. The parameters C_1, C_2, and γ are useful in characterizing the optimal solution of the platform problem per the theorem below.

Theorem 3.4.1. *There exists an optimal solution for the platform's problem I under which*
(i) $N = \min(\frac{L}{2}, C_1)$, $q_1^C = q_2^C = 0$, *and* $\eta_1^C = \eta_2^C = 0$ *if* $L < L^C$; *and*

(ii) $N = \min(\frac{\gamma L}{2}, C_1 + C_2 + q_1^*)$, $\eta_1^C \in (0,1)$, $\eta_2^C = 0$, $q_1^C = q_1^*$, *and* $q_2^C = 0$ *otherwise, where*

$$L^C = \begin{cases} \frac{2C_1\left(1+\sqrt{1-\gamma^2}\right)}{\gamma^2}, & \text{if } C_2 \geq \frac{\sqrt{1-\gamma^2}}{\gamma} C_1, \\ \frac{(C_2+q_1^*)(C_2+2C_1+q_1^*)}{\gamma C_2}, & \text{otherwise,} \end{cases} \tag{3.15}$$

$$\text{and} \quad q_1^* = \frac{t_{12}}{t_{21}} C_1. \tag{3.16}$$

Moreover, the number of drivers recruited weakly increases in the labor pool size L.

Theorem 3.4.1 indicates that when the driver pool size is sufficiently small (below the threshold L^C), the platform chooses to recruit at most C_1 drivers with these drivers choosing not to reposition (i.e., $\eta_1^C = \eta_2^C = 0$) and not to queue (i.e., $q_1^C = q_2^C = 0$). In this case, only demand of type-1 would be fulfilled with the amount fulfilled being at most $2\Lambda_{12}$. The fact that the platform chooses not to fulfill all demand can be explained by the high cost of drivers when the driver pool size is small (recall that drivers are heterogeneous in their opportunity costs, implying an increasing marginal driver cost).

When the driver pool size is large (above the threshold L^C), the platform may recruit more drivers than the amount needed to cover all the demand (i.e., choose N in excess of $C_1 + C_2$). In this case, some drivers choose to reposition from location 1 to 2 (i.e., $\eta_1^C \in (0,1)$) while others (choose) to wait at location 1, the low-demand location. Perhaps consistent with intuition, the volume of drivers preferring to wait at the low-demand location, rather than relocate to the high-demand location, as specified by $q_1^* = \frac{t_{12}}{t_{21}} C_1$, is increasing in t_{12} and C_1, and decreasing in t_{21}.

The fact that the platform recruits more drivers than what is needed to cover all the demand can be explained by the fact that drivers are strategic and would prefer to wait if the wait is not too long. (In Lemma B.1.1, we show that it is optimal for a driver to reposition from the low-demand location to the high-demand location with a positive probability only if $q_1^C \geq q_1^*$). In other words, in order to incentivize drivers to reposition away from location 1 so as to fulfill more of the demand from location 2, the platform must allow some drivers to idly queue up at location 1. The platform is willing to incur the associated cost (in the form of a higher wage) if the labor pool size is sufficiently large. If not, the platform forgoes type-2 demand and only fulfills type-1 demand.

These results suggest that regardless of whether the labor pool size is small or large, the platform ends up leaving something on the table: unfulfilled demand when the labor pool size is small and excess supply of drivers when the labor pool size is large. Drivers are similarly affected: lower wages offered and fewer drivers recruited when the labor pool is small and lower utilization when the labor pool size is large. As we discuss next, this *efficiency loss* can be attributed to the strategic behavior of drivers.

The Centralized System – To assess the efficiency loss due to drivers behaving strategically, let us consider the case of a centralized system in which the platform has control over the repositioning of drivers (i.e., η^C is no longer a CV equilibrium repositioning strategy anymore). In this case, the platform solves the following problem:

$$\begin{aligned} \text{(Problem II)} \qquad \max_{w,\eta^C} \quad & \Pi^C = (p-w)(s_{12}^C + s_{21}^C) \\ \text{subject to} \quad & (3.1), (3.2), (3.4), (3.6), (3.7), (3.9), (3.10), \\ & M = 0 \text{ and } F_k = 0 \text{ for } k \in \{1,2\}. \end{aligned}$$

Lemma 3.4.1. *There exists a unique optimal solution for the platform's Problem II under which*

(i) $N = \min(\frac{L}{2}, C_1)$, $q_1^C = q_2^C = 0$, *and* $\eta_1^C = \eta_2^C = 0$ *if* $L < \frac{2C_1}{\gamma}$; *and*

(ii) $N = \min(\frac{\gamma L}{2}, C_1 + C_2)$, $\eta_1^C \in (0,1)$, $\eta_2^C = 0$, *and* $q_1^C = q_2^C = 0$ *otherwise.*

Moreover, the number of drivers recruited weakly increases in the labor pool size L.

Comparing Lemma 3.4.1 with Theorem 3.4.1, we observe that (i) the threshold on L above which drivers start to reposition and fulfill type-2 demand in a centralized system (i.e., $\frac{2C_1}{\gamma}$) is smaller than that when drivers are strategic (i.e., L^C); and (ii) drivers reposition even though no drivers are queuing up at location 1. Moreover, the platform never invests in capacity in excess of $C_1 + C_2$.

Let Π_I^C (Π_{II}^C) denote the optimal value of Problem I (II) and DW_I^C (DW_{II}^C) the corresponding driver welfare. Under the centralized system, the platform is obviously always (weakly) better off (i.e.,$\Pi_{II}^C \geq \Pi_I^C$). In the proposition below, we show that both the platform and the drivers can be strictly better off in the centralized system.

Proposition 3.4.1. *When* $L \in (\frac{2C_1}{\gamma}, L^C)$, $DW_{II}^C > DW_I^C$ *and* $\Pi_{II}^C > \Pi_I^C$.

Proposition 3.4.1 shows that, by giving the platform control over the repositioning of drivers, both platform profit and driver welfare can be higher. An explanation is as follows. By controlling the repositioning of drivers the platform can eliminate driver queueing at location 1. This makes drivers more productive, allowing the platform to hire more of them and pay them a higher wage, resulting in more demand being fulfilled and a higher net profit for the platform.

The results in Lemma 3.4.1 and Proposition 3.4.1 show how the strategic interactions between the platform and the drivers and among the drivers themselves can result in outcomes that are less advantageous to all parties involved. In the next section, we show how the introduction of AVs, though short of direct control of driver decision making, can lead to improved outcomes for all.

3.5 The System with AVs

In this section, we consider the case where the platform may deploy a mixed fleet of CVs and AVs. In deciding how to assign incoming demand to available vehicles, the platform may choose to prioritize AVs, CVs, or neither (i.e., choose randomly among available vehicles). In Appendix B.4, we show that the AV-prioritized policy dominates the other two policies from the platform's perspective. Therefore, for simplicity, we first consider the case where the platform adopts the AV-prioritized policy. We defer the analysis of the other two assignment policies to the end of this section.

Under an AV-prioritized policy, the system from the perspective of drivers is equivalent to one in which the demand fulfilled by AVs is removed. If, in steady state, $q_i^A > 0$, the expected delay experienced by AVs and CVs is given by $W_i^A = \frac{q_i^A}{\Lambda_{ij}}$ where $j \neq i$ and $W_i^C = +\infty$ respectively. Otherwise, $W_i^A = 0$ and $W_i^C = \frac{q_i^C}{\Lambda_{ij} - \lambda_{ij}^A}$. Therefore, demand from location i is assigned to CVs only if no AVs are queued at location i. That is, we have

$$s_{ij}^C q_i^A = 0 \quad \text{for} \quad i \neq j \in \{1, 2\}. \tag{3.17}$$

The platform's problem can now be stated as follows:

$$\begin{aligned} \text{(Problem A)} \qquad \max_{M,w,\eta^A} \quad & \Pi = p(s_{12}^A + s_{21}^A) + (p-w)(s_{12}^C + s_{21}^C) - M \cdot I, \\ \text{subject to} \quad & \text{(3.1)–(3.10), (3.17) and} \\ & \eta^C \text{ is a CV equilibrium repositioning strategy.} \end{aligned}$$

In Theorem 3.5.1 below, we characterize the solution to the platform's problem (recall F_i corresponds to the fraction of effective demand fulfilled by AVs at location i and q_1^* is defined in (3.16)).

Theorem 3.5.1. *There exists an optimal solution for the platform's problem A under which there exists a positive threshold L^A on the driver pool size such that*

(i) $N = \min(\frac{IL}{2p}, C_1)$, $\eta_1^C = \eta_2^C = 0$ *and* $q_1^C = q_2^C = 0$ *if* $L \leq L^A$*; and*

(ii) $N = \begin{cases} \min(\frac{\gamma L}{2}, C_2) & \text{if } L \in (L^A, \frac{2C_2 p}{\gamma I}] \\ \min(\frac{\gamma IL}{2p}, C_1 + C_2 + q_1^*) & \text{if } L > \max(L^A, \frac{2C_2 p}{\gamma I}), \end{cases}$

$\eta_1^C \in (0,1]$, $\eta_2^C = 0$, $(1-F_1)((\Lambda_{21} - \Lambda_{12}) - \nu_{12}^C) = 0$, $q_1^C = (1-F_1)q_1^*$ *and* $q_2^C = 0$ *otherwise. Moreover, the number of drivers recruited N weakly increases in the labor pool size L.*

Theorem 3.5.1 indicates that when the driver pool size is below the threshold L^A, the platform chooses to recruit at most C_1 drivers with these drivers choosing not to reposition (i.e., $\eta_1^C = \eta_2^C = 0$) and not to queue up (i.e., $q_1^C = q_2^C = 0$). When the driver pool size is above the threshold L^A, the platform may recruit more drivers than the amount needed to cover all the demand (i.e., the platform may choose N in excess of $C_1 + C_2$). In this case, two situations may arise: (1) drivers do not queue up at location 1 and always reposition ($q_1^C = 0$ and $\eta_1^C = 1$) and all demand from location 1 is fulfilled by the AVs ($F_1 = 1$) or (2) drivers queue up at location 1 with a queue size equal to $q_1^C = (1-F_1)q^* > 0$ and drivers fulfill all the type-2 demand ($F_1 < 1$, $\nu_{12}^C = \Lambda_{21} - \Lambda_{12}$).

Comparing Theorem 3.5.1 with Theorem 3.4.1, we observe that, in the presence of AVs, (i) drivers may reposition even though no drivers are queuing up at location 1 (the case when all demand of type-1 is fulfilled by AVs) and (ii) in the case where there is a queue, the queue size is smaller than the one without AVs. In other words, while the platform

may continue to leave something on the table (e.g., foregoing some demand), it is less likely to do so. More importantly, the platform is able to use more productively the drivers it recruits by inducing them to reposition more and by having fewer of them idle [2]. This is possible because the platform can now deploy enough AVs to discourage drivers from not repositioning (e.g., by prioritizing AVs in assigning demand at location 1 (the low-demand location), the platform can make queueing at location 1 less desirable). We refer to this effect as the *incentive effect*.

The introduction of AVs obviously always (weakly) improves profit for the platform (the platform can always choose not to invest in any AVs). In Theorem 3.5.2, we show that, under some conditions, the introduction of AVs also improves outcomes for drivers.

Let DW^A (DW_I^C) and N^A (N_I^C) denote the driver welfare and the amount of drivers recruited in systems with and without AVs respectively. Also, let w^A (w^C) and $(r_{12}^C)^A$ ($(r_{12}^C)^C$) denote the wage paid to drivers and the volume of drivers repositioning from location 1 (the low-demand location) to location 2 (the high-demand location) in a system with (without) AVs.

Theorem 3.5.2. *$DW^A > DW_I^C$, $w^A > w^C$, and $N^A > N_I^C$ if and only if $L^C > L^A$ and $L \in (L^A, L^C)$, where L^A and L^C are defined in Theorem 3.5.1 and Theorem 3.4.1 respectively. Moreover, $(r_{12}^C)^A > (r_{12}^C)^C = 0$ for $L \in (L^A, L^C)$.*

Theorem 3.5.2 shows that drivers can also benefit from the introduction of AVs with more workers recruited earning higher wages and enjoying higher welfare. This occurs when the labor pool size is in the interval $L \in (L^A, L^C)$ (we observe numerically that the width of this interval is largest when the purchase cost of AVs is moderate; see Figure 3.2) [3]. An explanation is as follows. By deploying AVs in sufficient number and giving them priority in fulfilling type-1 demand, the platform incentives drivers to reposition and fulfill some type-2 demand. This incentive effect makes drivers more profitable to the platform, leading it to hire more of them and, in doing so, paying them more.

[2] In Lemma B.1.1, we show that, in a system without AVs, a single driver is willing to reposition from location 1 (the low-demand location) to location 2 (the high-demand location) only if $q_1^C \geq q_1^*$. In a system with AVs, this queueing threshold is reduced to $(1 - F_1)q_1^*$. Notice that this threshold is decreasing in F_1 (the fraction of type-1 demand fulfilled by AVs), implying that the platform may decrease the queueing by drivers by choosing to fulfill more type-1 demand using AVs.

[3] On a related note, the introduction of AVs always (weakly) improves the customer service level, with the service level being strictly higher when $L \in (L^A, L^C)$.

In summary, the presence of AVs, in addition to giving the platform an alternative source of supply, can be used to mitigate the impact of harmful strategic behavior on the part of drivers. As long as (1) the AV purchase cost is not too low (so that the displacement effect is strong) or too high (so that the incentive effect is weak), and (2) the driver pool size is not too small (so that the displacement effect is strong) and not too large (so that the incentive effect is weak), this can also be beneficial to drivers [4].

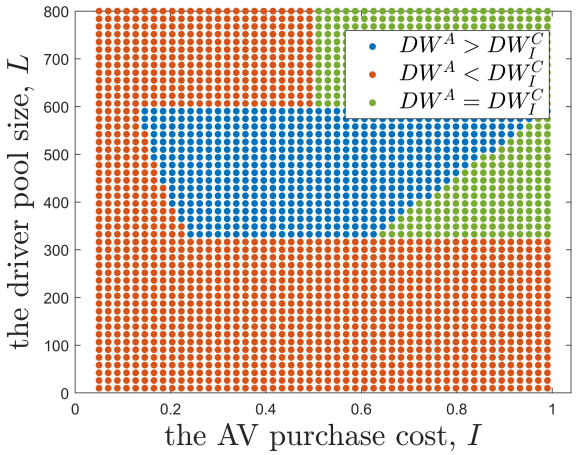


Figure 3.2: Driver welfare in systems with and without AVs. Model parameters: $\Lambda_{12} = 20$, $\Lambda_{21} = 200$, $t_{12} = t_{21} = 1$ and $p = 1$.

We conclude this section by showing that a strict improvement in driver welfare with the introduction of AVs is only possible when the platform used an AV-prioritized policy. Specifically, per Proposition 3.5.1 below, driver welfare cannot be improved if the platform prioritizes CVs nor can it be improved if it does not distinguish between the two (i.e., assigns demand randomly)[5]. Let Π^R (Π^{CP}) and DW^R (DW^{CP}) denote the platform profit and driver welfare respectively under the platform's optimal strategy when the platform adopts

[4] In Appendix B.6, we extend the analysis to a system with location-dependent pricing where the platform can also decide on whether to charge customers a higher price and pay workers a higher wage for trips originating from the high-demand location. We observe numerically that our qualitative results remain intact.

[5] Detailed discussion of the problem formulation, the platform's optimal strategy, and the corresponding outcomes under the random assignment and the CV-prioritized policies are provided in Appendix B.2 and Appendix B.3 respectively.

the random assignment policy (CV-prioritized policy).

Proposition 3.5.1. *For any AV purchase cost I and driver pool size L, $DW^R \leq DW_I^C$ and $DW^{CP} \leq DW_I^C$.*

The results in Proposition 3.5.1 can be explained by the fact that, in comparison to the AV-prioritized policy, the CV-prioritized policy and the random assignment policy, give the platform less control over the strategic behavior of drivers (i.e., the incentive effect is weak). Specifically, under the CV-prioritized policy, the system is equivalent to one without AVs from the driver's perspective. Therefore, the platform cannot use AVs to influence drivers' behavior. Under the random assignment policy, the platform can only deploy AVs to compete with CVs for type-1 demand, which is ineffective since the queuing threshold cannot be reduced, and costly because excess supply of AVs must be wasted queuing at the low-demand location.

3.6 Concluding Comments

In this paper, we examined the impact of introducing AVs on the welfare of human drivers. We did so, using an equilibrium model that accounts for the spatial features of demand and for the strategic behavior of human divers. Our findings reveal a nuanced relationship between the introduction of AVs and its impact on driver welfare. While AVs can lead to driver displacement and wage depression, we also identify scenarios where their introduction results in higher wages and more drivers being hired. We show that these results can be explained by the interplay of two counteracting effects resulting from AV implementation: a displacement effect (which hurts human drivers) and an incentive effect (which benefits human drivers). The relative strength of these effects crucially depends on the costs of AVs and CVs, with the incentive effect outweighing the displacement effect when the costs of both AVs and CVs are moderate. The dependency of outcomes on the cost structures of AVs and CVs opens the door for possible regulatory interventions that can induce more socially desirable outcomes (e.g., a regulator may affect these costs via subsidies, taxes or the use of direct limits on the mix of AVs and CVs deployed). Furthermore, we highlight the importance of work assignment prioritization. Perhaps surprisingly, we show that human drivers can only be better off with the introduction of AVs when the

platform prioritizes AVs. This highlights an operational lever that can be used to steer the adoption of automation in a more desirable direction for workers. Lastly, the insights gained from this paper may extend beyond the realm of ride-hailing and hold relevance for other applications involving a mix of automation and human workers where workers are strategic. The findings in this paper may also be of relevance to hybrid workplaces, where some of the workers are traditional employees and others are independent contractors.

Chapter 4

Do Workers and Customers Benefit from Competition between On-Demand Service Platforms?

4.1 Introduction

The use of on-demand services, such as ride-hailing and food delivery, is rapidly increasing, leading to a surge in the number of platforms operating in this industry (Sheromova (2020)). Projections indicate that the market value of on-demand platforms will reach approximately $335 billion by 2025 (Mansuri (2022)). These new platforms are entering into competition with incumbent players that have been in the business for a longer time. For instance, the rivalry between Uber and Lyft began when Lyft entered the ride-hailing market three years after Uber did. Furthermore, new market entrants may attract different customer segments compared to the incumbent, offering a closer match to their preferences. An instance of this phenomenon can be observed in the ride-hailing industry, where Uber caters more to businesspeople, while Lyft emphasizes friendly and interactive service (Farrington (2022)). Similarly, in the food delivery sector, Slice concentrates solely on pizza delivery (Slice (2022)), while Chowbus specializes in delivering Asian cuisine (Chowbus (2022)).

In this work, we explore the impact of the entrance of a new on-demand platform on both workers and customers. Specifically, we analyze the competition between two

on-demand service platforms, the incumbent and the entrant. The new entrant positions itself differently from the existing incumbent by catering to a different niche market on the demand-side, resulting in customer heterogeneity in platform preference modeled through a Hotelling line. In other words, in the setting we consider, competition is accompanied by *market expansion*, as the entry of a new platform brings with it customers who otherwise might not participate (due to customers' preferences for features offered by one platform but not the other). The workers are heterogeneous in the income derived from their outside options. The platforms, who are profit-seeking, compete for both customers and workers by deciding on prices to charge customers and wages to pay workers. Customers are sensitive to both price and congestion, which decreases in the supply of workers and increases in demand. Workers are independent agents who decide on whether or not to work for one (*single-homing*) or both platforms (*multi-homing*). If they decide to work for one of the platforms (or both), they forego income associated with their outside option. Workers are paid by the platforms only when they are busy (i.e., they are not compensated when they are idle). We compare the outcomes of this scenario to those when the incumbent operates as a monopolist before the entry of a new platform.

The setting after the entry of a new platform can be viewed as one involving competition in a two-sided market where the two platforms compete for both supply and demand, with supply affecting demand and vice-versa. Such a competition gives rise to several important questions. How does competition affect wages and prices and does competition necessarily lead to higher wages and lower prices? Does competition make workers busier or does it lead to more worker idleness? More importantly, does competition necessarily lead to higher worker welfare and higher consumer surplus? If not, under what conditions does competition harm either workers or customers?

In this paper, we address these questions. In particular, we identify conditions under which both workers and consumers are worse off under competition than under monopoly. We show that this can arise when the worker pool size is sufficiently large and customer *stickiness* (the strength of preference of customers for one platform over another) is moderate. Specifically, we obtain the following results.

- There exists a symmetric duopoly equilibrium when the worker pool size is large and customer stickiness is not too low. Otherwise, the competition becomes too intense

either on the customer or worker side, and the two platforms may not be able to coexist.

- Depending on customer stickiness, competition between on-demand service platforms can result in either higher or lower worker welfare (relative to a monopoly), with worker welfare being lower when customer stickiness is moderate. The wage workers earn is always lower, even though their workload is always higher (i.e., they are busier).

- Similarly, depending on customer stickiness, competition can result in either higher or lower consumer surplus, with consumer surplus being lower (relative to a monopoly) when customer stickiness is moderate. In this case, customers pay a higher price and experience more congestion.

That it is possible for competition to harm both workers and customers can be explained as follows. When the pool of potential workers is sufficiently large and so is customer stickiness, the equilibrium consists of workers choosing to work for both platforms (i.e., workers are multi-homing). Generally, a platform that offers a higher wage benefits from more worker supply which can be leveraged into more revenue. However, when workers work for both platforms, a platform that offers a higher wage increases worker supply not only for itself but for the other platform. This puts a downward pressure on the marginal benefit derived from a wage increase, possibly resulting in a lower wage relative to the monopoly case. This effect, which we refer to as the multi-homing effect, is particularly pronounced when (1) customer stickiness is not too high, so that the market expansion effect (resulting from the new entrant occupying a different niche market) is not too strong (market expansion could benefit workers as it increases their workload) and (2) customer stickiness is not too low, as low customer stickiness intensifies the competition for customers, making more supply more valuable. Although workers are busier in the region where worker welfare is lower, the increase in their workload is not sufficient to overcome the decrease in the wages (this is possible when customer stickiness is neither too high nor too low). Hence, perhaps paradoxically, in this regime, workers earn less even though they work more. On the customer side, when customer stickiness is neither too high nor too low, platforms find it profitable to forego customers who favor the competitor in exchange for higher prices. The

net effect is lower consumer surplus not only because of the higher prices but also because of the higher delay. Customers always experience more congestion under competition because of the associated market expansion effect that is not matched with a corresponding increase in labor supply due to the lower wages being offered.

The rest of our paper is organized as follows. In Section 2, we discuss related literature. In section 3, we describe our model. In Section 4, we provide the equilibrium analysis. In Section 5, we compare outcomes before and after the entry of a new platform. In Section 6, we offer concluding comments. Proofs for all the results, unless otherwise stated, are included in the Appendix.

4.2 Literature Review

Our work is related to the growing operations management literature on on-demand service platforms. Reviews of this literature can be found in Benjaafar and Hu (2020), Hu (2021), and Chen et al. (2019) and the references therein. The focus of this literature has been on settings with a single platform that operates as a monopolist. Literature that considers two or more competing platforms is less extensive.

Bernstein et al. (2021) study competition between on-demand service platforms under two settings: one in which there is a dedicated pool of workers for each platform (single-homing), and one in which all workers work for both platforms (multi-homing). The platforms decide on price only, with the wage being an exogenously-specified fraction of price. Under these assumptions, they show that (1) "surge pricing" (relative to fixed pricing) benefits workers and consumers and (2) single-homing (relative to multi-homing) benefits customers and workers. Our results complement theirs in two ways. First, we focus on comparing outcomes under competition and monopoly to highlight the impact of a new entrant. We show that these outcomes with respect to worker welfare and consumer surplus crucially depend on customer stickiness (customer stickiness does not appear to affect the comparisons of multi-homing and single homing in Bernstein et al. (2021)). Second, we consider a more general model setting where (1) the platforms set prices and wages independently[1] (instead of wage being a fixed fraction of price as assumed in Bernstein et al.

[1]This appears to be consistent with recent trends in how on-demand service platforms decide on prices and wages; see for example Garg and Nazerzadeh (2021).

(2021)), and (2) worker's decision on which platform to operate through is an equilibrium outcome rather than an exogenous requirement. We show that our setting leads to different results compared to those presented in Bernstein et al. (2021).

Nikzad (2022) studies a setting similar to ours with two platforms competing for workers and customers via prices and wages, while they assume that workers earn income from their outside option when they are idle (our assumption of workers not generating income when they are idle appears consistent with the reality of certain on-demand services, such as those involved in ride-hailing and home deliveries). They show that competition increases wages and improves worker welfare but can increase prices and reduce average consumer surplus when the labor pool size is moderate. Because workers earn outside income when they are idle, they make a decision on whether or not to work through the platforms based only on the wage offered and independently of the extent to which they expect to be busy. In our work, workers are sensitive to both the wage (per service rendered) and workload.

Cohen and Zhang (2022) study competition and "coopetition" between two on-demand service platforms. They assume that customers and workers choose between the platforms according to a multinomial logit choice model. They show that competition (relative to a monopoly setting where both platforms are owned by a single entity) results in lower prices and higher wages. They also show that all parties, namely the platforms, customers, and workers, can benefit from coopetition (an arrangement involving the two platforms collaborating on a joint service and sharing profits).

Chen et al. (2021) study a two-period setting where on-demand service platforms compete for workers via bonuses, where bonuses are given out to workers who work for the same platform in both periods. They show that whether bonuses are offered depends on the worker stickiness (strength of preference by workers for one platform over the other). They also show that, depending on worker stickiness, offering bonuses (relative to a setting where offering bonuses is not an option) can reduce both profit and social welfare.

Bai and Tang (2022) identify conditions under which competition between two on-demand service platforms leads to both platforms being profitable. They show that this is possible when customers are heterogeneous in their sensitivity to delay, the platforms have exclusive customers or workers, or the platforms employ time-dependent pricing. They do not compare, as we do, outcomes with and without competition with regard to prices,

wages, and consumer and worker welfare.

Wu et al. (2020) study competition between two on-demand service platforms and compare settings where (i) workers and customers move sequentially and (ii) workers and customers move simultaneously. They show that the two settings yield different outcomes. Their model assumes customers are indifferent between the two platforms, all workers are active, and the labor pool size is exogenous. They also do not study as we do the impact of competition on labor welfare and consumer surplus.

Siddiq and Taylor (2022b) study competition between two on-demand service platforms when one of the platforms has access to autonomous vehicles. They study how the presence of automated vehicles affects equilibrium outcomes in terms of prices and platform profit. Among their findings is that the equilibrium profit of one platform may decrease in its rival's cost of acquiring autonomous vehicles.

Bakos and Halaburda (2020) examine a two-sided market with two competing platforms where the mass of participants on one side generates externalities for the other side (which implies that a platform can maximize its total profits by subsidizing one of the sides). They use a Hotelling line model for both sides and demonstrate that permitting agents to multi-home on both sides weakens (or eliminates) the benefit of subsidy.

Ahmadinejad et al. (2020) examine a scenario where the time it takes to fulfill a service request is dependent on the availability of idle workers. They find that competition can lead to a phenomenon known as the wide-goose chase (WGC) [2]. This effect is pronounced when customers are highly sensitive to delay, and it can be avoided otherwise.

Zhang et al. (2022) compare three wage schemes in the competition between two on-demand platforms: (1) platforms first select wages and then prices, (2) platforms first select commission rates and then prices, and (3) platforms select prices and wages simultaneously. Their analysis evaluates the performance of these wage schemes in terms of platform profit, consumer surplus, worker welfare, and social welfare, and identifies conditions under which one wage scheme outperforms another. Hu and Liu (2021) examine a similar problem, but with two additional wage scheme settings: (1) platforms first select price and then wage, and (2) platforms commit to transaction capacity and then select price and wage simultaneously.

[2] In which workers end up traveling long distances to pick up far-away customers, resulting in longer wait times and lower throughput (see Castillo et al. (2021)).

Our work is also related to the economics literature on competition in two-sided markets; see for example Rochet and Tirole (2003), Caillaud and Jullien (2003), Armstrong (2006), Rochet and Tirole (2006), and Wright (2012) and the references therein. The settings considered in this literature do not have an on-demand feature in the sense that the utility of individuals on either side of the market is affected by the size of one side relative to the other. For example, in our setting, the size of demand *relative* to the size of worker supply determines the fraction of demand that a worker captures, and hence, their earnings. Similarly, the size of demand relative to the size of worker supply determines the delay experienced by a customer.

4.3 Problem Formulation

We consider the competition between an incumbent and a new entrant for both customers and workers. For convenience, we denote the imcumbent as platform 1 and the new entrant as platform 2. Platform i, where $i \in \{1, 2\}$, makes two decisions: the price p_i it charges customers for fulfilling each service request, and the wage w_i it pays workers for carrying out each service request. We use $\mathbf{P} = (p_1, p_2, w_1, w_2)$ to denote a strategy profile of the two platforms. The two platforms decide on their prices and wages simultaneously. For convenience, we use i, $i \in \{1, 2\}$, to denote one platform and $j = 3 - i$ to denote the other.

There is a unit mass of infinitesimal customers. Customers are heterogeneous in their preferences for the two platforms, which can be captured by a Hotelling line model. That is, customers are uniformly *located* between the two platforms, the distance between which is normalized to 1. All else being equal, customers prefer to receive service from a platform nearby. We use x, $x \in [0, 1]$, to denote the location of a customer on the Hotelling line. The customer (she) with location x (referred to as customer x) incurs a *traveling* cost tx to receive service from platform 1 and a traveling cost $t(1 - x)$ to receive service from platform 2, where $t > 0$ is a scaling parameter that accounts for customer *stickiness* (or strength of customer preference). A customer can choose to receive service from one of the two platforms, or fulfill her need from an outside option. A customer's utility of receiving service from platform i is determined by (i) the price p_i she pays for the service; (ii) the congestion cost she experiences at platform i, denoted by C_i and determined in equilibrium as a function of the realized demand and worker supply (more on this later); and (iii) the

traveling cost to platform i. Specifically, customer x receives utility $u_1(x) = v - tx - p_1 - C_1$ if she chooses platform 1 and $u_2(x) = v - t(1-x) - p_2 - C_2$ if she chooses platform 2, where v is the nominal value derived from having a service request fulfilled by one of the platforms. Otherwise, she receives utility from the outside option.We use u_0 to denote the customer utility derived from the outside option, which is the same for all customers. Because we will vary the value of t and analyze its impact on outcomes, we can normalize $v - u_o$ to 1 without loss of generality. Let λ_i denote the realized demand of platform i, then

$$\lambda_1 = \max\{x \in [0,1] : 1 - tx - p_1 - C_1 \geq 1 - t(1-x) - p_2 - C_2 \text{ and } 1 - tx - p_1 - C_1 \geq 0\} \quad (4.1)$$

$$\lambda_2 = \min\{x \in [0,1] : 1 - t(1-x) - p_2 - C_2 \geq 1 - tx - p_1 - C_1 \text{ and } 1 - t(1-x) - p_2 - C_2 \geq 0\}. \quad (4.2)$$

We consider a continuum of workers with mass M. We refer to M as the worker pool size. Workers are heterogeneous in the incomes they earn from their outside options, which are uniformly distributed from 0 to 1. We assume workers are indifferent between the two platforms. This assumption is reasonable when considering applications such as ride-hailing and home delivery. It is also consistent with treatment elsewhere in the literature; see for example, Rochet and Tirole (2003), Nikzad (2022), Teh et al. (2022) and Ahmadinejad et al. (2020).

A worker (he) can choose to take the outside option, to work for one platform, or work for both platforms. We say that a worker is *single-homing* if he works for only one platform, and a worker is *multi-homing* if he works for both platforms. For the worker with opportunity cost y (referred to as worker y), if he chooses the outside option, he drives utility normalized to 0. If he chooses to work for platform i only, $i \in \{1, 2\}$, he drives utility $w_i\rho_i$, where ρ_i is the amount of work (or workload) worker y receives. The realized demand at platform i is uniformly rationed among the workers who work for platform i so that workers whose platform-joining decisions are the same receive the same amount of work. Hence, in a setting where demand and worker supply are measured per unit time, ρ_i has the interpretation of the fraction of time a worker is busy working for platform i and

w_i is the wage a worker earns per unit time the worker is busy [3]. If he chooses to work for both platforms, he accepts service requests from both platforms and his utility is given by $w_1\rho_{b1} + w_2\rho_{b2} - y$, where ρ_{bi} is the amount of work the worker receives from platform i (or equivalently the fraction of time the worker is busy working for platform i) by working for "both" platforms [4]. We let $\rho_b = \rho_{b1} + \rho_{b2}$. Workers choose whichever option yields the highest utility. If there are multiple options that generate the same highest utility, we assume that a worker prefers working for more platforms. That is, all else being equal, a worker prefers diversifying his sources of income [5]. The key results in the paper (i.e., Theorem 4.4.1 and Theorem 4.5.2) continue to hold if a worker prefers working for fewer platforms. For convenience, we refer to workers who choose not to take the outside option as *active* workers.

Let S_i, $i \in \{1, 2\}$, denote the realized supply of workers who work for platform i only, and let S_b denote the realized supply of workers who work for both platforms. Define $\hat{w}_i = w_i\rho_i$ and $\hat{w}_b = w_1\rho_{b1} + w_2\rho_{b2}$. Then, $\hat{w}_i$ and $\hat{w}_b$ correspond respectively to the income of a worker (or his *effective* wage) if he works for platform i only or for both platforms.

Given a strategy profile $\mathbf{P} = (p_1, p_2, w_1, w_2)$ of the two platforms, customers and workers make decisions simultaneously. Let $\mathbf{A}|_{\mathbf{P}} = (\lambda_1, \lambda_2, S_1, S_2, S_b)|_{\mathbf{P}}$ denote a market allocation under the strategy profile $\mathbf{P} = (p_1, p_2, w_1, w_2)$. We say that $\mathbf{A}|_{\mathbf{P}}$ is a *subgame equilibrium* if no customer or worker has an incentive to deviate from her/his current action with the market allocation $\mathbf{A}$ under the strategy profile $\mathbf{P}$. As workers prefer working for more platforms (all else being equal), for any strategy profile $\mathbf{P}$, there are only three possible types of subgame equilibria: (i) there are no active workers and no customers, that is $\lambda_1 = \lambda_2 = S_1 = S_2 = S_b = 0$; (ii) all active workers work for the same platform, say

[3] Formally, if demand is measured in terms of service requests per unit time (with each request requiring, on average, one unit of time) and supply is measured in terms of the number of service requests per unit time that can be fulfilled by the workers that choose to work (also per unit time), then the ratio of demand to supply corresponds to the fraction of time a worker is busy; see also Benjaafar et al. (2021a) for additional details.

[4] We assume there is no extra cost related to multi-homing. This is consistent perhaps with what we see in ride hailing and food deliveries because it is easy for workers to switch between platforms. However, in settings where multi-homing is costly to workers, the equilibrium analysis is more complicated and the results derived in this paper may not apply.

[5] Worker utilization may fluctuate over short periods of time. Working for more platforms allows workers to smooth income. The preference for multi-homing is consistent with assumptions made elsewhere in the literature and with observed practice in certain applications such as ride-hailing; see for example Ahmadinejad et al. (2020).

platform i, i.e., $\lambda_i > 0$, $S_i > 0$ and $\lambda_j = S_j = S_b = 0$; and (iii) all active workers work for both platforms, that is $S_1 = S_2 = 0$, $\lambda_1 > 0$, $\lambda_2 > 0$, and $S_b > 0$. Note that there is no subgame equilibrium such that both platforms have dedicated workers, i.e., $S_1 > 0$ and $S_2 > 0$. This is because, for any strategy profile $\mathbf{P}$ such that $w_i \geq w_j$, it is more profitable for a worker to work for both platforms than to work for platform j only.

Similar to Bernstein et al. (2021), we assume that the congestion cost C_i is a function of worker workload, denoted by $c(\rho)$ with ρ being the workload. Under type (ii) subgame equilibrium where platform i has a positive market share, $C_i = c(\rho_i)$. Under type (iii) subgame equilibrium, $C_i = C_j = c(\rho_b)$. We assume that the function $c(\rho)$ is strictly increasing and convex in ρ (i.e., $c'(\rho) > 0$ and $c''(\rho) > 0$), $c(0) = 0$ and $1 \leq c(1) < \infty$. Note that by (4.1)–(4.2), $c(1) \geq 1$ implies that the demand must be strictly less than supply under any subgame equilibrium.

The amount of labor supply under different scenarios can be specified as follows. Under type (ii) subgame equilibrium,

$$S_i = M w_i \frac{\lambda_i}{S_i}, \tag{4.3}$$

since only workers with incomes from their outside options lower than $w_i \frac{\lambda_i}{S_i}$ are willing to work for platform i. Rearranging terms in (4.3) leads to $S_i = \sqrt{M w_i \lambda_i}$. Similarly, under type (iii) subgame equilibrium, $S_b = M \frac{w_1\lambda_1 + w_2\lambda_2}{S_b}$, or equivalently, $S_b = \sqrt{M(w_1\lambda_1 + w_2\lambda_2)}$.

Platform i decides on p_i and w_i so as to maximize its profit. In order to analyze the equilibrium, we consider the optimization problem faced by platform i given the strategy of platform j:

$$\begin{aligned} \max_{p_i, w_i} \quad & \lambda_i(p_i - w_i), \\ \text{subject to} \quad & \mathbf{A}|_{\mathbf{P}} \ \text{ is a subgame equilibrium.} \end{aligned} \tag{4.4}$$

We close this section by introducing two important metrics, namely consumer surplus and worker welfare. Consumer surplus, denoted by CS, is the aggregate utility derived by customers. First, note that λ_i is the demand for platform i, which consists of customers whose distance to platform i is no larger than λ_i (recall that customers are uniformly distributed along the Hotelling line and the total mass of customers is 1). That is, customers

in locations $x \in [0, \lambda_1]$ choose service from platform 1 and customers in locations $x \in [1-\lambda_2, 1]$ choose service from platform 2. Then, consumer surplus CS can be expressed as

$$CS = \int_0^{\lambda_1} [1 - tx - p_1 - c(\rho)]dx + \int_{1-\lambda_2}^{1} [1 - t(1-x) - p_2 - c(\rho)]dx, \tag{4.5}$$

where worker workload $\rho = \rho_i = \frac{\lambda_i}{S_i}$ under type (ii) subgame equilibrium given platform i has a positive market share and $\rho = \rho_b = \frac{\lambda_1+\lambda_2}{S_b}$ under type (iii) subgame equilibrium.

Worker welfare, denoted by LW, is aggregate worker utility:

$$LW = M \int_0^1 \max(\hat{w} - y, 0)dy = \frac{M\hat{w}^2}{2}, \tag{4.6}$$

where $\hat{w} = \hat{w}_i$ under type (ii) subgame equilibrium given platform i has a positive market share and $\hat{w} = \hat{w}_b$ under type (iii) subgame equilibrium.

4.4 Equilibrium Analysis

We first consider the subgame equilibrium among customers and workers given a strategy profile $\mathbf{P}$ of the two platforms. Then we examine the equilibrium between platforms. Throughout, we use the term "equilibrium" to refer to Nash equilibrium.

4.4.1 Subgame Equilibrium Analysis

Recall from Section 4.3 that, for any strategy profile $\mathbf{P}$, there exists three possible types of subgame equilibria: (i) $\lambda_1 = \lambda_2 = S_1 = S_2 = S_b = 0$; (ii) $\lambda_i > 0$, $S_i > 0$ and $\lambda_j = S_j = S_b = 0$; and (iii) $S_1 = S_2 = 0$, $\lambda_1 > 0$, $\lambda_2 > 0$, and $S_b > 0$. For the sake of convenience, following Nikzad (2022), we refer to $\mathbf{A}|_{\mathbf{P}}$ a *non-trivial subgame equilibrium* if both platforms have a positive market share, i.e., $\lambda_1 > 0$ and $\lambda_2 > 0$; and a *trivial subgame equilibrium* otherwise. Therefore, type (iii) subgame equilibra are non-trivial while type (i) and type (ii) are trivial.

Since workers and customers move simultaneously, for any strategy profile $\mathbf{P}$, there always exists a type (i) trivial subgame equilibrium. For a type (ii) trivial subgame equilibrium to exist, customer demand and labor supply must satisfy the following set of

equations:

$$\lambda_i = \max\{\lambda \in [0,1] : 1 - t\lambda - p_i - C(\frac{\lambda}{S_i}) \geq 0\}, \tag{4.7}$$

$$S_i = \sqrt{M w_i \lambda_i}, \quad \text{and} \tag{4.8}$$

$$\lambda_j = S_j = S_b = 0, \tag{4.9}$$

where (4.8) follows from the analysis in Section 4.3. One can check that for $p_i < 1$ and $w_i > 0$, there exists a type (ii) trivial subgame equilibrium involving platform i, $i \in \{1, 2\}$.

Now let us consider the existence of a type (iii) subgame equilibrium. Under this type, there are two possible market-coverage outcomes: (a) the demand market is not fully covered, i.e., $\lambda_1 + \lambda_2 < 1$; and (b) the demand market is fully covered, i.e., $\lambda_1 + \lambda_2 = 1$. Under outcome (a), the marginal customer for platform i (that is, the customer whose distance to platform i equals λ_i) is indifferent between seeking service from platform i or choosing the outside option. This implies that $1 - t\lambda_i - p_i - C(\frac{\lambda_i + \lambda_j}{S_b}) = 0$, for $i = 1, 2$. Formally, for a type (iii) subgame equilibrium with partial market coverage to exist, customer demand and worker supply must satisfy the following set of equations:

$$1 - t\lambda_i - p_i - C(\frac{\lambda_i + \lambda_j}{S_b}) = 0, \qquad i \in \{1, 2\} \tag{4.10}$$

$$\lambda_i + \lambda_j < 1, \quad \text{and} \tag{4.11}$$

$$S_b = \sqrt{M(w_1\lambda_1 + w_2\lambda_2)}, \tag{4.12}$$

where (4.12) follows from the analysis in Section 4.3.

Under outcome (b), the market is fully covered, i.e., $\lambda_1 + \lambda_2 = 1$. In this case, the marginal customer for platform i is indifferent between seeking service from platform 1 and seeking service from platform 2, with the corresponding utility being no lower than the utility derived from the outside option. That is, $1 - t\lambda_i - p_i - C(\frac{\lambda_1 + \lambda_2}{S_b}) = 1 - t(1 - \lambda_i) - p_j - C(\frac{\lambda_1 + \lambda_2}{S_b}) \geq 0$. Hence, for a type (iii) subgame equilibrium with full market coverage to exist, customer demand and worker supply must satisfy the following set of equations:

$$\lambda_i = \frac{t + p_j - p_i}{2t}, \qquad i \in \{1, 2\} \tag{4.13}$$

$$1 - t\lambda_i - p_i - C(\frac{1}{S_b}) \geq 0, \qquad i \in \{1,2\} \text{ and} \tag{4.14}$$

$$S_b = \sqrt{M(w_1\lambda_1 + w_2\lambda_2)}, \tag{4.15}$$

where equation (4.15) follows from the analysis in Section 4.3.

Lemma 4.4.1. *Given a strategy profile* $\mathbf{P} = (p_1, p_2, w_1, w_2)$ *with* $w_i, p_i \in (0,1)$*, there always exists a trivial type (i) equilibrium and a trivial type (ii) subgame equilibrium involving platform* i *for* $i \in \{1,2\}$*. Depending on the parameters, a non-trivial type (iii) subgame equilibrium may or may not exist, and it is possible to have multiple non-trivial subgame equilibra* [6]*.*

By Lemma 4.4.1, it is possible for multiple subgame equilibria to exist given a strategy profile $\mathbf{P}$ of the platforms. Therefore, we need to apply a selection rule to refine the Nash equilibrium in the subgame so as to make a clearer prediction. We adopt the following refinement rule, which appears to be in the spirit of treatments elsewhere in the literature; see for example Nikzad (2022).

Refinement Rule: For any strategy profile $\mathbf{P}$, if there exists a non-trivial subgame equilibrium, then customers and workers form a non-trivial subgame equilibrium. In the case that there exist multiple non-trivial subgame equilibria, customers and workers form a non-trivial subgame equilibrium with the highest worker welfare. If there does not exist a non-trivial subgame equilibrium, customers and workers form a trivial subgame equilibrium with the highest worker welfare.

Under the refinement rule, we select the subgame equilibrium based on worker welfare rather than by consumer surplus. We do so for two important reasons. First, a higher worker welfare means a higher income for each worker. The same may not be true for customers because they differ in their preferences for the platforms. Second, in reality, workers may have more market power as they are the service providers. However, our main results continue to hold if we were to select based on consumer surplus rather than worker welfare.

[6] Sufficient conditions for the existence of a non-trivial subgame equilibrium (and for the existence of multiple non-trivial subgame equilibra) are provided in Appendix C.1.2. From the simulation results, there may only exist trivial subgame equilibra when $|w_1 - w_2|$ is large and M is small

4.4.2 Analysis of the Full Game

With the Refinement Rule specified in section 4.4.1, we have a unique prediction of the subgame equilibrium among customers and workers for any strategy profile **P** of the two platforms. Based on this, we are able to characterize the equilibrium of the full game (i.e., the competitive equilibrium between the two platforms). We say that an equilibrium is a *duopoly equilibrium* if the underlying subgame equilibrium is non-trivial. In Theorem 4.4.1, we provide results for the existence of non-trivial symmetric equilibrium.

Theorem 4.4.1. *There exists a threshold $\overline{M}(t)$ which depends on t, such that when $M > \overline{M}(t)$,*

(i) if $t \in (0, \frac{2}{3}]$, there does not exist a symmetric pure-strategy Nash equilibrium;

(ii) if $t \in (\frac{2}{3}, 1)$, there exists a unique non-trivial symmetric pure-strategy Nash equilibrium such that the demand-side market is fully covered; and

(iii) if $t \in [1, +\infty)$, there exists a unique non-trivial pure-strategy Nash equilibrium such that the demand-side market is partially covered.

Theorem 1 shows that for an equilibrium to exist the worker pool size and customer stickiness must be sufficiently large. Also, whether the market is fully covered or not depends on customer stickiness, with the market being fully covered when customer stickiness is moderate. When t is small (i.e., $t < \frac{2}{3}$), the competition on the demand side can become too intense for the coexistence of both platforms. As shown in Appendix C.3.1, when $t < \frac{2}{3}$, a platform can make more profit by reducing the price to attract more customers and increasing the wage to attract more workers, leading to a "winner-takes-all" outcome. Similarly, even when $t > \frac{2}{3}$, a symmetric duopoly equilibrium is not guaranteed when the worker pool size is small, i.e., $M < \overline{M}(t)$, due to the intense competition on the worker's side.

4.5 The Impact of Competition

In this section, we compare outcomes for workers and customers before and after the entry of a new platform with respect to price, wage, workload, worker welfare and consumer surplus. We identify conditions under which competition results in worse outcomes for workers and customers.

The Monopoly Case. The system before the entry of a new platform (i.e., the monopoly case) is constructed by removing one of the platforms from the original duopoly model. Specifically, the monopolist (incumbent) solves the following problem:

$$\max_{p,w} \quad \lambda(p-w) \tag{4.16}$$

$$\text{subject to} \quad \lambda = \max\{\lambda \in [0,1] : 1 - t\lambda - p - c(\frac{\lambda}{S}) \geq 0\}, \tag{4.17}$$

$$S = \sqrt{Mw\lambda}. \tag{4.18}$$

In Theorem 4.5.1, we establish the existence and uniqueness of the optimal strategy for the monopoly incumbent.

Theorem 4.5.1. *For $t > 0$, there exists a unique optimal strategy in the monopoly case for all $M > 0$. Moreover, under the optimal strategy,*

(i) if $t \geq \frac{1}{2}$, the demand market is partially covered;

(ii) if $t \in (0, \frac{1}{2})$, there exists a threshold $\tilde{M}$ such that the market is fully covered when $M \geq \tilde{M}$, and the market is partially covered otherwise.

In what follows, we use the superscript m to denote the outcomes (under the optimal strategy) before the entry of a new platform (the monopoly case), and use the superscript d to denote the the outcomes (under the non-trivial symmetric equilibrium) after the entry of a new platform. These outcomes are compared in Theorem 4.5.2.

Theorem 4.5.2. *For $t > \frac{2}{3}$, there exists $\overline{M}^*(t) \geq \overline{M}(t)$ such that for $M \geq \overline{M}^*(t)$,*

- *$\rho^d > \rho^m$ (this result holds for all $M \geq \overline{M}(t)$) and $w^d < w^m$;*
- *$LW^d < LW^m$ if $t < \frac{2+\sqrt{10}}{6}$ and $LW^d \geq LW^m$ otherwise;*
- *$p^d > p^m$ if $t < 1$ and $p^d < p^m$ otherwise;*
- *$CS^d < CS^m$ if $t < \frac{\sqrt{2}}{2}$ and $CS^d \geq CS^m$ otherwise.*

Theorem 4.5.2 shows that it is possible for competition to harm both workers and customers. For workers, this is the case when customer stickiness is moderate (neither too low nor too high). The result can be explained as follows. Under competition, workers

multi-home. If workers multi-home, then an increase in labor supply for one platform translates into an increase in labor supply for the other platform. This diminishes the competitive advantage a platform gains from paying a higher wage and securing more supply (i.e., because supply is "shared", the resulting reduction in delay is enjoyed by both platforms and, hence, may not be an effective means for attracting more customers). This is particularly the case when (1) customer stickiness is not too high, so that the market expansion effect (resulting from the new entrant occupying a different niche market) is not too strong (market expansion could benefit workers as it increases their workload) and (2) customer stickiness is not too low, as low customer stickiness intensifies the competition for customers, making more supply more valuable so that the two platforms fail to co-exist. Note that, although workers are busier in the region where worker welfare is lower, the increase in their workload is not sufficient to overcome the decrease in the wages (this is possible when customer stickiness is neither too high nor too low). Hence, perhaps paradoxically, in this regime, workers earn less even though they work more.

Similarly, when customer stickiness is moderate (as specified in Theorem 4.5.2), customers end up paying higher prices, experiencing more congestion, and realizing a lower surplus. This can be explained as follows. Under competition, customers have heterogeneous preferences for the platforms. If a customer favors (is located closer to) platform 1, then platform 2 would need to set its price significantly lower than that of platform 1 (to overcome the higher traveling cost) to attract that customer. Because of this, a platform may choose to forego the market where its rival has a competitive advantage and, instead, cater to nearby customers, charging a higher price (since they do not have to compensate for the higher traveling costs). This would hold when (1) customer stickiness is not too high (if the stickiness is high, a monopoly platform would also cater to nearby customers, making it difficult for platforms under competition to charge even higher prices), and (2) customer stickiness is not too low (when the stickiness is low, customers are relatively indifferent between the two platforms, which intensifies the competition and the two platforms fail to co-exist). Customers always experience more congestion under competition because of the associated market expansion effect that is not matched with a corresponding increase in labor supply due to the lower wages being offered, as previously explained.

A possible policy implication from the above results is that, as both workers and customers can benefit from competition only when customer stickiness is high ($t > \frac{2+\sqrt{10}}{6}$), a social planner seeking to enhance worker welfare and consumer surplus may prefer promoting competition in such scenarios, while being more cautious otherwise.

4.6 Concluding Remarks

In this work, we study how does the entry of a new platform affect workers and customers. The new entrant differentiates itself from the incumbent by occupying a different niche of the market. Competition is often viewed as being socially desirable. The results in this paper suggest that some caution is warranted when competition is between service platforms that compete for both workers and customers and when workers may multi-home. In particular, we identify conditions under which competition between two platforms leads to worse outcomes for workers and customers. It does so by highlighting factors (namely, the multi-homing of workers and the stickiness of customers) that may drive the equilibrium toward such outcomes. The results of this paper highlights important forces that may affect outcomes under competition. Knowing how these forces come into play could be useful to both platforms and policy makers as they consider the implications of competition on profit and social welfare.

References

Acemoglu, D. (1998). Why do new technologies complement skills? directed technical changes and wage inequality. *The Quarterly Journal of Economics*, 113(4):1055–1089.

Acemoglu, D. and Restrepo, P. (2018a). Artificial intelligence, automation and work. Working Paper 24196, National Bureau of Economic Research.

Acemoglu, D. and Restrepo, P. (2018b). The race between man and machine: Implications of technology for growth, factor shares, and employment. *American Economic Review*, 108(6):1488–1542.

Acemoglu, D. and Restrepo, P. (2020). Robots and jobs: Evidence from us labor markets. *Journal of Political Economy*, 128(6):2188–2244.

Afèche, P., Liu, Z., and Maglaras, C. (2022). Ride-hailing networks with strategic drivers: The impact of platform control capabilities on performance. Working paper, University of Toronto.

Afèche, P., Liu, Z., and Maglaras, C. (2023). Ride-hailing networks with strategic drivers: The impact of platform control capabilities on performance. Forthcoming in *Manufacturing & Service Operations Management*.

Ahmadinejad, A., Nazerzadeh, H., Saberi, A., Skochdopole, N., and Sweeney, K. (2020). Competition in ride-hailing markets. Working paper, Stanford University.

Akturk, D., Candogan, O., and Gupta, V. (2021). Network inventory management: Approximate optimality in large-scale systems. Working paper, University of Chicago.

Altman, E., Jiménez, T., and Koole, G. (2001). Optimal admission control in resource-sharing system. *IEEE Transactions on Communications*, 49(9):1659–1668.

Armstrong, M. (2006). Competition in two-sided markets. *The RAND Journal of Economics*, 37(3):668–691.

Ashkrof, P., Homem de Almeida Correia, G., Cats, O., and Arem, B. (2022). On the relocation behaviour of ride-sourcing drivers. Working paper, Delft University of Technology.

Autor, D. H., Levy, F., and Murnane, R. J. (2003). The skill content of recent technological change: an empirical exploration. *The Quarterly Journal of Economics*, 118(4):1279–1333.

Bai, J. and Tang, C. S. (2022). Can two competing on-demand service platforms be profitable? *International Journal of Production Economics*, 250:108672.

Bakos, Y. and Halaburda, H. (2020). Platform competition with multihoming on both sides: Subsidize or not? *Management Science*, 66(12):5599–5607.

Banerjee, S., Freund, D., and Lykouris, T. (2022a). Pricing and optimization in shared vehicle systems: An approximation framework. *Operations Research*, 70(3):1783–1805.

Banerjee, S., Kanoria, Y., and Qian, P. (2022b). Large deviations optimal scheduling of closed queueing networks. Working paper, Cornell University.

Baron, O., Berman, O., and Nourinejad, M. (2022). Introducing autonomous vehicles: Adoption patterns and impacts on social welfare. *Manufacturing & Service Operations Management*, 24(1):352–369.

Batt, R. J. and Terwiesch, C. (2012). Doctors under load: An empirical study of state-dependent service times. Technical report, The Wharton School, University of Pennsylvania.

Benjaafar, S., Ding, J.-Y., Kong, G., and Taylor, T. (2021a). Labor welfare in on-demand service platforms. *Manufacturing & Service Operations Management*, 24(1):110–124.

Benjaafar, S., Ding, J.-Y., Kong, G., and Taylor, T. A. (2022a). Labor welfare in on-demand service platforms. *Manufacturing & Service Operations Management*, 24(1):110–124.

Benjaafar, S. and Hu, M. (2020). Operations management in the age of the sharing economy: What is old and what is new? *Manufacturing & Service Operations Management*, 22(1):93–101.

Benjaafar, S., Jiang, D., Li, X., and Li, X. (2021b). Dynamic inventory repositioning in on-demand rental networks. Forthcoming in *Management Science*.

Benjaafar, S. and Shen, X. (2023). Pricing in on-demand (and one-way) vehicle sharing networks. Forthcoming in *Operations Research*.

Benjaafar, S., Wang, Z., and Yang, X. (2023a). Human in the loop automation: Ride-hailing with remote (tele-) drivers. Working paper, University of Minnesota.

Benjaafar, S., Wang, Z., and Yang, X. (2023b). The impact of automation on workers when workers are strategic: The case of ride-hailing. Working paper, University of Minnesota.

Benjaafar, S., Wu, S., Liu, H., and Gunnarsson, E. (2022b). Dimensioning on-demand vehicle sharing systems. *Management Science*, 68(2):1218–1232.

Benzell, S. G., Kotlikoff, L. J., LaGarda, G., and Sachs, J. D. (2015). Robots are us: some economics of human replacement. *NBER Working Paper 20941*.

Bernstein, F., DeCroix, G., and Keskin, N. B. (2021). Competition between two-sided platforms under demand and supply congestion effects. *Manufacturing & Service Operations Management*, 23(5):1043–1061.

Besbes, O., Castro, F., and Lobel, I. (2022). Spatial capacity planning. *Operations Research*, 70(2):1271–1291.

Bessen, J., Goos, M., Salomons, A., and Van den Berge, W. (2019). Automatic reaction-what happens to workers at firms that automate? *Law and Economics Research Paper*.

Blanco, S. (2021). Startup's autonomy workaround: 'teledrivers' to operate cars from remote location. `https://www.caranddriver.com/news/a37648114/vay-autonomous-teledriver-startup/`.

Bösch, P. M., Becker, F., Becker, H., and Axhausen, K. W. (2018). Cost-based analysis of autonomous mobility services. *Transport Policy*, 64:76–91.

Braverman, A., Dai, J., Liu, X., and Ying, L. (2019). Empty-car routing in ridesharing systems. *Operations Research*, 67(5):1437–1452.

Brynjolfsson, E., Liu, M., and Westerman, G. (2022). Do computers reduce the value of worker persistence. *Journal of Management Information Systems*, 39(1):41–67.

Caillaud, B. and Jullien, B. (2003). Chicken & egg: Competition among intermediation service providers. *The RAND Journal of Economics*, 34(2):309–328.

Cao, J. and Qi, W. (2022). Stall economy: The value of mobility in retail on wheels. Forthcoming in *Operations Research*.

Castillo, J. C., Knoepfle, D. T., and Weyl, E. G. (2021). Matching in ride hailing: Wild goose chases and how to solve them. Working paper, Stanford University.

Castro, F. and Frazelle, A. (2021). Getting out of your own way: Introducing autonomous vehicles on a ride-hailing platform. Working paper, UCLA Anderson School of Management.

Castro, F. and Frazelle, A. E. (2022). Getting out of your own way: Managing human drivers and autonomous vehicles on a ride-hailing platform. Working paper, University of California, Los Angeles (UCLA).

Castro, F., Gao, J., and Martin, S. (2023). Autonomous vehicles in ride-hailing and the threat of spatial inequalities. Working paper, UCLA Anderson School of Management.

Chakravarty, A. K. (2021). Blending capacity on a rideshare platform: Independent and dedicated drivers. *Production And Operations Management*, 30(8):2522–2546.

Chan, C. W., Farias, V. F., and Escobar, G. J. (2017). The impact of delays on service times in the intensive care unit. *Management Science*, 63(7):2049–2072.

Cheah, J. and Smith, M. (1994). Generalized m/g/c/c state dependent queueing models and pedestrian traffic flows. *Queueing Systems*, 15:365–386.

Chen, L., Cui, Y., Liu, J., Liu, X., and Liu, X. (2021). Bonus competition in the gig economy. Working paper, Cornell University.

Chen, M. and Hu, M. (2022). Courier dispatch in on-demand delivery. Forthcoming in *Management Science*.

Chen, Y.-J., Dai, T., Korpeoglu, C. G., Körpeoğlu, E., Sahin, O., Tang, C. S., and Xiao, S. (2019). Innovative online platforms: Research opportunities. *Manufacturing & Service Operations Management*, 22(3):430–445.

Chowbus (2022). `https://www.chowbus.com/`.

Coco (2022). `https://www.cocodelivery.com/`.

Cohen, M. and Zhang, R. P. (2022). Competition and coopetition for two-sided platforms. *Production and Operations Management*, 31(5):1997–2014.

Dauth, W., Findeisen, S., Suedekum, J., and Woessner, N. (2021). The adjustment of labor markets to robots. *Journal of the European Economic Association*, 16(6):3104–3153.

Daw, A., Hampshire, R. C., and Pender, J. (2020). Beyond safety drivers: Staffing a teleoperations system for autonomous vehicles. Working paper, University of Southern California, Marshall School of Business.

Delasay, M., Ingolfsson, A., and Kolfal, B. (2016). Modeling load and overwork effects in queueing systems with adaptive service rates. *Operations Research*, 64(4):867–885.

Delasay, M., Ingolfsson, A., Kolfal, B., and Schultz, K. (2019). Load effect on service times. *European Journal of Operational Research*, 279(3):673–686.

Didi (2021). Form f-1 registration statement. `https://www.sec.gov/Archives/edgar/data/1764757/000104746921001194/a2243272zf-1.htm`.

Dixon, J., Hong, B., and Wu, L. (2021). The robot revolution: Managerial and employment consequences for firms. Forthcoming in *Management Science*.

Doll, G., Ebel, E., Heineke, K., and Wiemuth, C. (2020). Private autonomous vehicles: The other side of the robo-taxi story. *McKinsey & Company*.

Dong, J., Feldman, P., and Yom-Tov, G. B. (2015). Service systems with slowdowns: Potential failures and proposed solutions. *Operations Research*, 63(2):305–324.

Dong, J. and Ibrahim, R. (2020). Managing supply in the on-demand economy: Flexible workers, full-time employees, or both? *Operations Research*, 68(4):1238–1264.

Farrington, R. (2022). The ultimate lyft vs. uber comparison (for drivers and riders). `https://thecollegeinvestor.com/20641/ultimate-lyft-vs-uber-comparison-drivers-riders/`.

Feng, G., Kong, G., and Wang, Z. (2021). We are on the way: Analysis of on-demand ride-hailing systems. *Manufacturing & Service Operations Management*, 23(5):1237–1256.

Filippi, E., Bannò, M., and Trento, S. (2023). Automation technologies and their impact on employment: A review, synthesis and future research agenda. *Technological Forecasting and Social Change*, 191.

Freund, D., Henderson, S. G., and Shmoys, D. B. (2019). Bike sharing. In Hu, M., editor, *Sharing Economy: Making Supply Meet Demand*, volume 6 of *Springer Series in Supply Chain Management*, pages 435–459. Springer.

Frey, C. B. and Osborne, M. A. (2017). The future of employment: How susceptible are jobs to computerisation? *Technological Forecasting and Social Change*, 114:254–280.

Garg, N. and Nazerzadeh, H. (2021). Driver surge pricing. *Management Science*, 68(5):3219–3235.

Ge, Y., Knittel, MacGregor, C. R., MacKenzie, D., and Zoepf, S. (2016). Racial and gender discrimination in transportation network companies. `https://www.nber.org/papers/w22776`.

George, D. K. and Xia, C. H. (2011). Fleet-sizing and service availability for a vehicle rental system via closed queueing networks. *European Journal of Operational Research*, 211(1):198–207.

George, D. K., Xia, C. H., and Squillante, M. S. (2012). Exact-order asymptotic analysis for closed queueing networks. *Journal of Applied Probability*, 49(2):503–520.

Graetz, G. and Michaels, G. (2018). Robots at work. *The Review of Economics and Statistics*, 100(5):753–768.

Guo, K. (2022). Automation, skill and job creation. *Working Paper*.

Hampshire, R. C., Bao, S., Lasecki, W. S., Daw, A., and Pender, J. (2020). Beyond safety drivers: Applying air traffic control principles to support the deployment of driverless vehicles. *PLOS One*, 15(5).

He, E. J. and Goh, J. (2022). Profit or growth? dynamic order allocation in a hybrid workforce. *Management Science*, 68(8):5891–5906.

He, L., Hu, Z., and Zhang, M. (2020). Robust repositioning for vehicle sharing. *Manufacturing & Service Operations Management*, 22(2):241–256.

Hosseini, M., Milner, J., and Romero, G. (2021). Dynamic relocations in car-sharing networks. Working paper, Rotman School of Management.

Hu, M. (2021). From the classics to new tunes: A neoclassical view on sharing economy and innovative marketplaces. *Production & Operations Management*, 30(6):1668–1685.

Hu, M. (2022). Spatial or temporal pooling solves wild goose chase. Working paper, University of Toronto.

Hu, M. and Liu, Y. (2021). Precommitments in two-sided market competition. Working paper, Rotman School of Management, University of Toronto,.

Hu, M. and Zhou, Y. (2020). Price, wage, and fixed commission in on-demand matching. Working paper, University of Toronto.

Hémous, D. and Olsen, M. (2022). The rise of the machines: automation, horizontal innovation, and income inequality. *American Economic Journal: Macroeconomics*, 14(1):179–223.

Iyer, C. and Alton, R. (2019). The race for autonomous ride-hailing: Developing a strategy for success. Technical report, Christensen Institute.

Jackson, M. O. and Zafer, K. (2019). How automation that substitutes for labor affects production networks, growth, and income inequality. *Growth, and Income Inequality.*

Kanoria, Y. (2021). Dynamic spatial matching. Working paper, Columbia Business School.

Kanoria, Y. and Qian, P. (2020). Blind dynamic resource allocation in closed networks via mirror backpressure. Working paper, Columbia Business School.

Kc, D. S. and Terwiesch, C. (2009). Impact of workload on service time and patient safety: An econometric analysis of hospital operations. *Management science*, 55(9):1486–1498.

Korinek, A. and Stiglitz, J. E. (2020). Steering technological progress. *NBER Conference on the Economics of AI.*

Kostami, V. (2023). If you love your agents, set them free: Task discretion in online workplaces. Forthcoming in *Management Science*.

Lalley, C. (2017). Will self-driving cars replace uber and lyft drivers? https://www.policygenius.com/blog/will-self-driving-cars-replace-uber-lyft-drivers/.

Lian, Z. and Van Ryzin, G. (2022). Capturing the benefits of autonomous vehicles in ride-hailing: The role of dispatch platforms and market structure. Working paper, Cornell University.

Lobel, I., Martin, S., and Song, H. (2021). Employees versus contractors: An operational perspective. Working paper, New York University.

Lu, Y. and Zhou, Y. (2021). A review on the economics of artificial intelligence. *Journal of Economic Surveys*, 35(4):1045–1072.

Lyft (2021). Lyft (website). https://self-driving.lyft.com/level5/.

Mandelbaum, A. (1995). State-dependent queues: approximations and applications. *Stochastic networks*, 71:239–282.

Mandelbaum, A. and Pats, G. (1998). State-dependent stochastic networks. part i. approximations and applications with continuous diffusion limits. *The Annals of Applied Probability*, 8(2):569–646.

Mansuri, S. (2022). 13 service industries that drive the on-demand economy. https://www.peerbits.com/blog/service-industries-that-drive-the-on-demand-economy.html.

Mirzaeian, N., Cho, S.-H., and Scheller-Wolf, A. (2020). A queueing model and analysis for autonomous vehicles on highways. *Management Science*, 67(5):2904–2923.

Mondolo, J. (2021). The composite link between technological changes and employment: A survey of the literature. *Journal of Economic Surveys*, 36(4):1027–1068.

Nikzad, A. (2022). Thickness and competition in ride-sharing markets. Working paper, University of Southern California.

Noh, D., Tunca, T. I., and Xu, Y. (2021). Evolution of ride services: From taxicabs to ride hailing and self-driving cars. Working paper, University of California, San Diego.

O'Brien, S. A. (2022). Uber releases safety data: 998 sexual assault incidents including 141 rape reports in 2020. https://www.cnn.com/2022/06/30/tech/uber-safety-report/index.html#:~:text=Even%20one%20report%20is%20one,that%20of%20the%20first%20report).

Özkan, E. and Ward, A. R. (2020). Dynamic matching for real-time ride sharing. *Stochastic Systems*, 10(1):29–70.

Pegoraro, R. (2021). This driverless car-sharing service uses remote human 'pilots', not ai. https://www.fastcompany.com/90653650/halo-driverless-car-sharing-service.

PhantomAuto (2022). https://phantom.auto/.

Rio-Chanona, R. M., Mealy, P., Beguerisse-Díaz, M., Lafond, F., and Farmer, J. D. (2021). Occupational mobility and automation: a data-driven network model. *Journal of the Royal Society Interface*, 18(174).

Rochet, J.-C. and Tirole, J. (2003). Platform competition in two-sided markets. *Journal of The European Economic Association*, 1(4):990–1029.

Rochet, J.-C. and Tirole, J. (2006). Two-sided markets: a progress report. *The RAND Journal of Economics*, 37(3):645–667.

Ross, S. (1996). *Stochastic processes*. Wiley series in probability and statistics: Probability and statistics. Wiley.

Rudin, W. (1976). *Principles of mathematical analysis*, volume 3. McGraw-hill New York.

Santi, P., Resta, G., Szell, M., Sobolevsky, S., Strogatz, S. H., and Ratti, C. (2014). Quantifying the benefits of vehicle pooling with shareability networks. *PNAS*, 111(37):13290–13294.

Sawers, P. (2020). Smooth teleoperator: The rise of the remote controller. https://venturebeat.com/2020/08/17/smooth-teleoperator-the-rise-of-the-remote-controller/.

Sawers, P. (2022). Einride will now develop human-driven trucks as part of transition to full autonomy. https://venturebeat.com/2020/04/22/einride-will-now-develop-human-driven-trucks-as-part-of-transition-to-full-autonomy/.

Sheromova, V. (2020). 10 growth trends for on-demand service platforms in 2020-2021. https://exyte.com/blog/10-growth-trends-for-on-demand-service-platforms-2020.

Siddiq, A. and Taylor, T. A. (2021). Ride-hailing platforms: Competition and autonomous vehicles. Working paper, University of California, Berkeley.

Siddiq, A. and Taylor, T. A. (2022a). Ride-hailing platforms: Competition and autonomous vehicles. *Manufacturing & Service Operations Management*, 24(3):1511–1528.

Siddiq, A. and Taylor, T. A. (2022b). Ride-hailing platforms: Competition and autonomous vehicles. *Manufacturing & Service Operations Management*, 24(3):1511–1528.

Slice (2022). `https://slicelife.com/`.

Starship (2022). `https://www.starship.xyz/`.

Stidham, S. (1985). Optimal control of admission to a queueing system. *IEEE Transactions on Automatic Control*, 30(8):705–713.

Stidham, S. (2002). Analysis, design, and control of queueing systems. *Operations Research*, 50(1):197–216.

Tang, Y., Guo, P., Tang, C. S., and Wang, Y. (2021). Gender-related operational issues arising from on-demand ride-hailing platforms: Safety concerns and system configuration. *Production and Operations Management*, 30(10):3481–3496.

Taylor, T. A. (2018). On-demand service platforms. *Manufacturing & Service Operations Management*, 20(4):704–720.

Teh, T.-H., Liu, C., Wright, J., and Zhou, J. (2022). Multihoming and oligopolistic platform competition. *Amer. Econom. J.: Microeconom.*

Robotics Tomorrow (2022). Using romote operation to help solve labor shortage issues in the supply chain. `https://www.roboticstomorrow.com/article/2022/07/using-remote-operation-to-help-solve-labor-shortage-issues-in-the-supply-chain/19124`.

Uber (2019). Form s-1 registration statement. `https://www.sec.gov/Archives/edgar/data/1543151/000119312519103850/d647752ds1.htm`.

Wang, G., Zhang, H., and Zhang, J. (2022). On-demand ride-matching in a spatial model with abandonment and cancellation. Forthcoming in *Operations Research*.

Wang, J. and Zhou, Y.-P. (2018). Impact of queue configuration on service time: Evidence from a supermarket. *Management Science*, 64(7):3055–3075.

Waserhole, A. and Jost, V. (2016). Pricing in vehicle sharing systems: optimization in queuing networks with product forms. *EURO Journal on Transportation and Logistics*, 5:293–320.

Wright, J. (2012). Why payment card fees are biased against retailers. *The RAND Journal of Economics*, 43(4):761–780.

Wu, S., Xiao, S., and Benjaafar, S. (2020). Two-sided competition between on-demand service platforms. Working paper, University of Minnesota.

Xu, Z., Yin, Y., and Ye, J. (2020). On the supply curve of ride-hailing systems. *Transportation Research Part B: Methodological*, 132:29–43.

Zhang, C., Chen, J., and Raghunathan, S. (2022). Two-sided platform competition in a sharing economy. *Management Science*, 68(12):8909–8932.

Zhang, T. (2020). Toward automated vehicle teleoperation: Vision, opportunities, and challenges. *IEEE Internet of Things Journal*, 7(12):11347 – 11354.

Zhao, L., Liu, Z., and Hu, P. (2020). Dynamic repositioning for vehicle sharing with setup costs. *Operations Research Letters*, 48(6):792–797.

Zhong, Y., Gopalakrishnan, R., and Ward, A. (2022). Behavior-aware queueing: The finite-buffer setting with many strategic servers. Working paper, The University of Chicago Booth School of Business.

Örmeci, E. L., Burnetas, A., and van der Wal, J. (2001). Admission policies for a two class loss system. *Stochastic Models*, 17(4):513–539.

Appendix A

Appendices for Chapter 2

A.1 Preliminary Results

In this section, we prove Lemma A.1.1, which is used to prove various other results. We also characterize $\gamma_1(\lambda)$, $\gamma_2(\lambda)$, q_1, q_2 and q_3 which are introduced in Proposition 2.3.1. Recall that we define $\rho(q) = \frac{\lambda}{q\mu(m,q)}$ in (2.1).

Lemma A.1.1. *$\rho(q)$ is strictly convex and $\frac{1}{\rho(q)}$ is strictly concave. Moreover, there exist $\gamma_1(\lambda)$ and $\gamma_2(\lambda)$ with $\gamma_1(\lambda) < \gamma_2(\lambda)$ such that*

(i) if $m < \gamma_1(\lambda)$, $\rho(q) > 1$ for $1 \le q \le m$;

(ii) if $\gamma_1(\lambda) < m < \gamma_2(\lambda)$, there exist q_1 and q_2 with $q_1 < q_2$ such that $\rho(q) < 1$ for $q_1 < q < q_2$ and $\rho(q) \ge 1$ otherwise (the equality is achieved if and only if $q = q_1$ or $q = q_2$); and

(iii) if $m > \gamma_2(\lambda)$, there exist q_3 such that $\rho(q) < 1$ for $q > q_3$ and $\rho(q) \ge 1$ otherwise (the equality is achieved if and only if $q = q_3$).

Proof of Lemma A.1.1. We first show that $\rho(q)$ is strictly convex:

$$\rho'(q) = \frac{\partial}{\partial q}\left(\frac{\lambda}{q\mu(m,q)}\right) = -\frac{\lambda\left(\mu(m,q) + q\frac{\partial\mu(m,q)}{\partial q}\right)}{[q\mu(m,q)]^2}, \quad \text{and}$$

$$\begin{aligned}\rho''(q) &= \frac{\partial^2}{\partial q^2}\left(\frac{\lambda}{q\mu(m,q)}\right) \\ &= -\frac{\lambda\left[[q\mu(m,q)]^2\left(2\frac{\partial\mu(m,q)}{\partial q} + q\frac{\partial^2\mu(m,q)}{\partial q^2}\right) - 2q\mu(m,q)\left(\mu(m,q) + q\frac{\partial\mu(m,q)}{\partial q}\right)^2\right]}{[q\mu(m,q)]^4} > 0,\end{aligned}$$

where the inequality holds because $\frac{\partial\mu(m,q)}{\partial q} < 0$ (as $\mu(m,q)$ is deceasing in q) and $\frac{\partial^2\mu(m,q)}{\partial q^2} < 0$ (as $\mu(m,q)$ is strictly concave in q). Similarly, we can show that $\frac{1}{\rho(q)} = \frac{q\mu(m,q)}{\lambda}$ is strictly concave in q.

For ease of exposition, we shall use the notation $\rho(m,q)$ to indicate the dependence of $\rho(q)$ on m throughout the remaining of this section. We then show the following three results.

(1) When $\lambda > \max\limits_{q\in\{1,2,\ldots,\bar{m}\}} q\mu(\bar{m},q)$, $\min\limits_{q\in[1,m]} \rho(m,q) = 1$ admits a unique solution on m, which we denote by $\gamma_1(\lambda)$. For convenience, let $q_{\min}(m) = \arg\min\limits_{q\in[1,m]} \rho(m,q)$. To prove the statement, it suffices to show that $\rho(m, q_{\min}(m))$ is decreasing in m. Because $\rho(m,q) = \frac{\lambda}{q\mu(m,q)}$ is decreasing in m and strictly convex in q, we have $\rho(m+1, q_{min}(m+1)) \leq \rho(m+1, q_{min}(m)) < \rho(m, q_{min}(m))$ as desired.

(2) When $\lambda > \max\limits_{q\in\{1,2,\ldots,\bar{m}\}} q\mu(\bar{m},q)$, $\rho(m,m) = 1$ admits a unique solution on m which we denote by $\gamma_2(\lambda)$. To prove this statement, it suffices to show that $\rho(m,m) = \frac{\lambda}{m\mu(m,m)}$ is decreasing in m, which immediately follows from Assumption 2.3.1 ($\mu(m,m)$ is invariant in m).

(3) $\gamma_1(\lambda) < \gamma_2(\lambda)$. This is because $\rho(m, q_{\min}(m))$ and $\rho(m,m)$ are both decreasing in m, and $\rho(m, q_{\min}(m)) < \rho(m,m)$ when $m > \bar{m}$.

With the above results, we consider the following three cases.

Case (i) $m < \gamma_1(\lambda)$. Because $\min\limits_{q\in[1,m]} \rho(m,q) > 1$, we have $\rho(q) > 1$ for all $1 \leq q \leq m$.

Case (ii) $\gamma_1(\lambda) < m < \gamma_2(\lambda)$. Because (a) $\rho(m,q)$ is strictly convex with respect to q, (b) $\min\limits_{q\in[1,m]} \rho(m,q) < 1$, (c) $\rho(m,m) > 1$, and (d) $\rho(m,1) > 1$ when $\lambda > \frac{1}{s}$, we conclude that $\rho(m,q) = 1$ admits two roots which we denote by q_1 and q_2 with $q_1 < q_2$. Therefore,

$\rho(m,q) < 1$ if $q_1 < q < q_2$ and $\rho(m,q) \geq 1$ otherwise.

Case (iii) $m > \gamma_2(\lambda)$. Because (a) $\rho(m,q)$ is strictly convex in q, (b) $\rho(m,m) < 1$ and (c) $\rho(m,1) > 1$ when $\lambda > \frac{1}{s}$, we conclude that $\rho(m,q) = 1$ admits a unique solution which we denote by q_3. Moreover, $\rho(m,q) < 1$ if $q > q_3$ and $\rho(m,q) \geq 1$ otherwise. ■

A.2 Proof of Proposition 2.3.1 and 2.3.2

We first prove Proposition 2.3.2. By (2.2)–(2.3), we have

$$\pi_{m,n}(q+1) = \rho(q+1)\pi_{m,n}(q), \tag{A.1}$$

where $\rho(q)$ is defined in (2.1). By Lemma A.1.1, we can show the following results.

(1) If $m < \gamma_1(\lambda)$, $\rho(q) > 1$ for $1 \leq q \leq n$. It follows that $\pi_{m,n}(q)$ is increasing in q for $q \in \{1, \cdots, n\}$. Therefore, $\pi_{m,n}(q)$ is unimodal with the mode at n.

(2) If $\gamma_1(\lambda) < m < \gamma_2(\lambda)$, $\rho(q) < 1$ for $q \in (q_1, q_2)$ and $\rho(q) > 1$ for $q \in (0, q_1)$ and $q \in (q_2, n]$. Therefore, if $n < q_1$, $\pi_{m,n}(q)$ is increasing in q for $q \in \{1, \cdots, n\}$ and thus $\pi_{m,n}(q)$ is unimodal with the mode at n. If $q_1 < n < q_2$, $\pi_{m,n}(q)$ is increasing in q for $q \in (0, q_1)$ and decreasing in q for $q \in (q_1, n]$. Therefore, $\pi_{m,n}(q)$ is unimodal with the mode at $\lfloor q_1 \rfloor$. If $n > q_2$, $\pi_{m,n}(q)$ is increasing in q for $q \in (0, q_1) \cup (q_2, n]$, and it is decreasing in q for $q \in (q_1, q_2)$. Therefore, $\pi_{m,n}(q)$ is bimodal with one mode at $\lfloor q_1 \rfloor$ and the other at n.

(3) If $m > \gamma_2(\lambda)$, $\rho(q) > 1$ for $q \in (0, q_3)$ and $\rho(q) < 1$ for $q \in (q_3, m]$. Therefore, if $n < q_3$, $\pi_{m,n}(q)$ is increasing in q for $q \in \{1, \cdots, n\}$ and thus $\pi_{m,n}(q)$ is unimodal with the mode at n. If $n > q_3$, $\pi_{m,n}(q)$ is increasing in q for $q \in (0, q_3)$ and decreasing in q for $q \in (q_3, n]$. Therefore, $\pi_{m,n}(q)$ is unimodal with the mode at $\lfloor q_3 \rfloor$.

The proof for Proposition 2.3.1 is a special case of that for Proposition 2.3.2 with $m = n$.

A.3 Proof of Theorem 2.3.1.A, 2.3.1.B and 2.3.1.C

In this section, we prove Theorem 2.3.1.A – Theorem 2.3.1.C. We first introduce a useful recursive result for $\pi_{m,n}(n)$ with respect to n per Lemma A.3.1.

Lemma A.3.1. *Given any m, for $n \in \{1, \cdots, m\}$, we have $\pi_{m,n-1}(n-1) > \pi_{m,n}(n)$ if and only if $\pi_{m,n-1}(n-1) > \frac{\rho(n)-1}{\rho(n)}$, and $\pi_{m,n-1}(n-1) > \pi_{m,n}(n)$ if and only if $\pi_{m,n}(n) > \frac{\rho(n)-1}{\rho(n)}$.*

Proof of Lemma A.3.1. By (2.2)–(2.3), for $n \geq 1$, we have

$$\pi_{m,n}(n) = \frac{\prod_{q=1}^{n} \rho(q)}{1 + \sum_{q=1}^{n} \prod_{k=1}^{q} \rho(k)}, \tag{A.2}$$

where $\rho(q)$ is defined in (2.1). By some algebra, $\pi_{m,n}(n)$ can be rewritten as

$$\pi_{m,n}(n) = \frac{\overbrace{\prod_{q=1}^{n-1} \rho(q)}^{a} + \overbrace{[\rho(n) - 1] \prod_{q=1}^{n-1} \rho(q)}^{c}}{\underbrace{1 + \sum_{q=1}^{n-1} \prod_{k=1}^{q} \rho(k)}_{b} + \underbrace{\prod_{q=1}^{n} \rho(q)}_{d}}. \tag{A.3}$$

Observe that $\pi_{m,n-1}(n-1) = \frac{a}{b}$, and $\frac{c}{d} = \frac{\rho(n)-1}{\rho(n)}$, where a, b, c and d are illustrated in (A.3). Then Lemma A.3.1 follows from the fact that given $a, b, c, d > 0$, $\frac{a}{b} > \frac{a+c}{b+d} \Leftrightarrow \frac{a}{b} > \frac{c}{d}$, and $\frac{a+c}{b+d} > \frac{c}{d} \Leftrightarrow \frac{a}{b} > \frac{a+c}{b+d}$. ∎

In the following subsections, we provide proofs for Theorem 2.3.1.A, 2.3.1.B and 2.3.1.C.

A.3.1 Proof of Theorem 2.3.1.A

By Lemma A.1.1, because $\pi_{m,1}(1) = \frac{\rho(1)}{1+\rho(1)} > \frac{\rho(2)-1}{\rho(2)}$ (as $\rho(1) > \rho(2)$ by Assumption 2.3.2), $\pi_{m,2}(2) < \pi_{m,1}(1)$. Therefore, either (a) $\pi_{m,n}(n)$ is decreasing in n for all $n \in \{1, \cdots, m\}$, or (b) there exists $\tilde{n}_1$ such that $\pi_{m,n}(n) > \pi_{m,n+1}(n+1)$ for $n \leq \tilde{n}_1 - 1$ and $\pi_{m,\tilde{n}_1}(\tilde{n}_1) \leq \pi_{m,\tilde{n}_1+1}(\tilde{n}_1 + 1)$. In scenario (b), we show that $\pi_{m,n}(n)$ is increasing in n for $n \geq \tilde{n}_1 + 1$ in the following steps.

Step (i). Suppose (for a contradiction) that $\rho(\tilde{n}_1+1) \leq \rho(\tilde{n}_1)$. Because $\pi_{m,\tilde{n}_1-1}(\tilde{n}-1) > \pi_{m,\tilde{n}_1}(\tilde{n}_1)$, by Lemma A.3.1, we have $\pi_{m,\tilde{n}_1}(\tilde{n}_1) > \frac{\rho(\tilde{n}_1)-1}{\rho(\tilde{n}_1)} \geq \frac{\rho(\tilde{n}_1+1)-1}{\rho(\tilde{n}_1+1)}$, where the last inequality is due to $\rho(\tilde{n}_1 + 1) \leq \rho(\tilde{n}_1)$. This implies that $\pi_{m,\tilde{n}_1+1}(\tilde{n}_1 + 1) < \pi_{m,\tilde{n}_1}(\tilde{n}_1)$ by Lemma A.3.1, which leads to a contradiction. Therefore, we must have $\rho(\tilde{n}_1 + 1) > \rho(\tilde{n}_1)$.

Step (ii). We show that $\rho(q)$ is increasing in q for $q \in \{\tilde{n}_1, \cdots, m\}$. This is because (1) $\rho(q)$ is strictly convex by Lemma A.1.1, and (2) $\rho(\tilde{n}_1 + 1) > \rho(\tilde{n}_1)$ by the analysis in step (i).

Step (iii). We show that $\pi_{m,n}(n)$ is increasing in n for $n \in \{\tilde{n}_1 + 1, \cdots, m\}$. Because $\pi_{m,\tilde{n}_1}(\tilde{n}_1) \leq \pi_{m,\tilde{n}_1+1}(\tilde{n}_1 + 1)$, by Lemma A.3.1, we have $\pi_{m,\tilde{n}_1+1}(\tilde{n}_1 + 1) \leq \frac{\rho(\tilde{n}_1+1)-1}{\rho(\tilde{n}_1+1)} < \frac{\rho(\tilde{n}_1+2)-1}{\rho(\tilde{n}_1+2)}$, where the last inequality is due to $\rho(\tilde{n}_1 + 2) > \rho(\tilde{n}_1 + 1)$ by step (ii). This further implies that $\pi_{m,\tilde{n}_1+2}(\tilde{n}_1 + 2) > \pi_{m,\tilde{n}_1+1}(\tilde{n}_1 + 1)$ by Lemma A.3.1. By applying this argument recursively, the desired result follows.

To summarize, we have shown that either $\pi_{m,n}(n)$ is decreasing in n for all $n \in \{1, \cdots, n\}$ (scenario (a)), or $\pi_{m,n}(n)$ is first decreasing and then increasing in n (scenario (b)). To check which scenario the system lies in, it suffices to compare $\pi_{m,m}(m)$ and $\pi_{m,m-1}(m-1)$. If $\pi_{m,m}(m) > \pi_{m,m-1}(m-1)$, there must exist a unique $\tilde{n}_1 \leq m-1$ such that $\pi_{m,n}(n)$ is decreasing in n if $n \leq \tilde{n}_1$ and it is increasing in n if $n > \tilde{n}_1$. Note that

$$SL(m, n+1) = \frac{\overbrace{1 + \sum_{i=1}^{n-1} \prod_{k=1}^{i} \rho(k)}^{a} + \overbrace{\prod_{k=1}^{n} \rho(k)}^{c}}{\underbrace{1 + \sum_{i=1}^{n} \prod_{k=1}^{i} \rho(k)}_{b} + \underbrace{\prod_{k=1}^{n+1} \rho(k)}_{d}}.$$

Observe that $\frac{a}{b} = SL(m, n)$ and $\frac{c}{d} = \frac{1}{\rho(n+1)} = \frac{(n+1)\mu(m,n+1)}{\lambda}$. By virtue of the following relation: $\frac{a+c}{b+d} < \frac{a}{b} \Leftrightarrow \frac{a+c}{b+d} > \frac{c}{d}$ given $a, b, c, d > 0$, we can obtain that $\pi_{m,m}(m) > \pi_{m,m-1}(m-1)$ if and only if $SL(m, m) > \frac{m\mu(m,m)}{\lambda}$, which is Condition (2.5).

A.3.2 Proof of Theorem 2.3.1.B

We prove Theorem 2.3.1.B in three steps. Recall the characterization of q_1 and q_2 in Lemma A.1.1. In step (i), we show that $\pi_{m,n}(n)$ is decreasing in n if $1 \leq n < q_1$; in step (ii), we show that $\pi_{m,n}(n)$ is decreasing in n for $q_1 \leq n \leq q_2$; and in step (iii), we show that if (2.5) holds, there exists $\tilde{n}_2 > q_2$ such that $\pi_{m,n}(n)$ is decreasing in n if $q_2 \leq n \leq \tilde{n}_2$ and increasing in n if $\tilde{n}_2 < n \leq m$.

Step (i). First, we note that $\rho(q)$ is decreasing in q for $1 \leq q < q_1$. This is because

by Lemma A.1.1, $\rho(q)$ is strictly convex and $\rho(q) = 1$ admits two solutions q_1 and q_2 with $q_1 < q_2$ given $\gamma_1(\lambda) < m < \gamma_2(\lambda)$. Because $\pi_{m,2}(2) < \pi_{m,1}(1)$ (see Proof of Theorem 2.3.1.A), by Lemma A.3.1, we have $\pi_{m,2}(2) > \frac{\rho(2)-1}{\rho(2)} > \frac{\rho(3)-1}{\rho(3)}$, where the last inequality is due to the fact that $\rho(q)$ is decreasing for $1 \leq q < q_1$. Then by Lemma A.3.1, we have $\pi_{m,3}(3) < \pi_{m,2}(2)$. By applying this argument recursively, we can obtain that $\pi_{m,n}(n)$ is decreasing in n for $1 \leq n < q_1$.

Step (ii). By Lemma A.1.1, when $\gamma_1(\lambda) < m < \gamma_2(\lambda)$, $\rho(q) < 1$ for $q_1 < q < q_2$. Recall from (A.2) that $\pi_{m,n}(n) = \frac{\prod_{q=1}^{n} \rho(q)}{\sum_{q=1}^{n} \prod_{k=1}^{q} \rho(k)}$. For $q_1 < n < q_2$, because $\prod_{q=1}^{n} \rho(q)$ is decreasing in n and $\sum_{q=1}^{n} \prod_{k=1}^{q} \rho(k)$ is increasing in n, it follows that $\pi_{m,n}(n)$ is decreasing in n.

Step (iii). First, by the same argument as in step (i), $\rho(q)$ is increasing in q for $q_2 < q \leq m$. Suppose there exists $\tilde{n}_2 \in (\hat{q}, m)$ such that $\pi_{m,n}(n) > \pi_{m,n+1}(n+1)$ for $n \leq \tilde{n}_2 - 1$ and $\pi_{m,\tilde{n}_2}(\tilde{n}_2) \leq \pi_{m,\tilde{n}_2+1}(\tilde{n}_2 + 1)$. Then we have $\pi_{m,\tilde{n}_2+1}(\tilde{n}_2 + 1) \leq \frac{\rho(\tilde{n}_2+1)-1}{\rho(\tilde{n}_2+1)} < \frac{\rho(\tilde{n}_2+2)-1}{\rho(\tilde{n}_2+2)}$, where the first inequality follows from Lemma A.3.1 and the second inequality follows from the fact that $\rho(q)$ is increasing in q for $q_2 < q \leq m$. Then by Lemma A.3.1, we have $\pi_{\tilde{n}_2+2}(m) > \pi_{\tilde{n}_2+1}(m)$. By applying this argument recursively, we have that $\pi_{m,n}(n)$ is increasing in n for $n \in \{\tilde{n}_2, \cdots, m\}$.

Therefore, if $\pi_{m,m}(m) > \pi_{m,m-1}(m-1)$, there must exist a unique integer $\tilde{n}_2 \leq m-1$ such that $\pi_{m,n}(n)$ is increasing in n for $n \geq \tilde{n}_2$ and it is decreasing in n for $n < \tilde{n}_2$; otherwise $\pi_{m,n}(n)$ is non-increasing in n for $n \in \{1, \cdots, m\}$.

A.3.3 Proof of Theorem 2.3.1.C

By Lemma A.1.1, when $m > \gamma_2(\lambda)$, $\rho(q)$ is decreasing in q for $1 \leq q < q_3$ and $\rho(q) < 1$ for $q_3 < q \leq m$. Therefore, by applying the same analysis in step (i) of the proof of Theorem 2.3.1.B, we can obtain that $\pi_{m,n}(n)$ is decreasing in n for $1 \leq n < q_3$. By applying the same analysis in step (ii) of the proof of Theorem 2.3.1.B, we can obtain that $\pi_{m,n}(n)$ is decreasing in n for $q_3 < n \leq m$.

Next, we obtain a lower bound for $\pi_{m,n}(n)$. For $m > \gamma_2(\lambda)$ and $n > q_3$, by (2.2)–(2.3),

we have

$$\begin{aligned}\pi_{m,n}(n) &= \frac{\prod_{k=1}^{n}\rho(k)}{1+\sum_{q=1}^{n}\prod_{k=1}^{q}\rho(k)} \\ &< \frac{\prod_{k=1}^{n}\rho(k)}{\sum_{q=\lfloor q_3\rfloor}^{n}\prod_{k=1}^{q}\rho(k)} \\ &\overset{(a)}{\leq} \frac{\prod_{k=1}^{n}\rho(k)}{(n-\lfloor q_3\rfloor)\prod_{k=1}^{n}\rho(k)} \\ &= \frac{1}{n-\lfloor q_3\rfloor}.\end{aligned}$$

where (a) follows from the fact that $\prod_{k=1}^{q}\rho(k)$ is decreasing in q as $\rho(q) < 1$ for $q_3 < q \leq m$ (see Lemma A.1.1). Therefore, we have $SL(m,n) = 1 - \pi_{m,n}(n) > 1 - \frac{1}{n-\lfloor q_3\rfloor}$.

A.4 Proof of Theorem 2.3.2.A, 2.3.2.B and 2.3.2.C

In this section, we prove the asymptotic results for systems with inpatient customers. We prove Theorem 2.3.2.A, 2.3.1.B and 2.3.2.C in Appendix A.4.3, A.4.2 and A.4.3, respectively. In Appendix A.4.4, we provide comparisons between finite system ratios (i.e., service level ratio in the supply-limited regime and driver-to-vehicle ratio in the supply-rich regime) derived from simulations of example systems and their asymptotic bounds.

We first characterize $\rho(q)$, $\gamma_1(\lambda)$, $\gamma_2(\lambda)$, q_1, q_2, and q_3 when $\mu(m,q)$ is given by (2.4).

$$\rho(q) = \frac{\lambda}{q}\left(\frac{s}{\sqrt{m-q+1}} + s\right). \tag{A.4}$$

For convenience, define

$$RHS(q) = \frac{q}{\lambda s} - \frac{1}{\sqrt{m-q+1}}.$$

Observe that $RHS(q) \leq 1$ implies $\rho(q) \geq 1$ and vise versa. Therefore, it is equivalent to

investigate $RHS(q)$. Because

$$RHS'(q) = \frac{1}{\lambda s} - \frac{1}{2}(m-q+1)^{-3/2}, \text{ and}$$
$$RHS''(q) = -\frac{3}{4}(m-q+1)^{-5/2} < 0,$$

$RHS(q)$ is strictly concave and the maximum is achieved at $q_{max} = m + 1 - \left(\frac{\lambda s}{2}\right)^{2/3}$. (i) If $RHS(q_{\max}) < 1$, which is equivalent to $m < \lambda s + 3\left(\frac{\lambda s}{2}\right)^{2/3} - 1$, we have $RHS(q) < 1$ for all $0 \le q \le m$. (ii) If $RHS(q_{max}) > 1$ and $RHS(m) < 1$, which is equivalent to $\lambda s + 3\left(\frac{\lambda s}{2}\right)^{2/3} - 1 < m < 2\lambda s$, because $RHS(\lambda s) < 1$, $RHS(q) = 1$ admits two roots in $[\lambda s, m]$ which we denote by q^* and $\hat{q}$ with $q^* < \hat{q}$. Therefore, $RHS(q) > 1$ for $q^* < q < \hat{q}$ and $RHS(q) \le 1$ otherwise. (iii) If $RHS(q_{max}) > 1$ and $RHS(m) > 1$, which is equivalent to $m > 2\lambda s$, the equation $RHS(q) = 1$ admits a unique root in $[\lambda s, m]$ which we denote by q^* (we will show later that it has the same expression as the smaller root in case (ii)). Therefore, $RHS(q) < 1$ if $q < q^*$ and $RHS(q) \ge 1$ otherwise. Per definitions of $\gamma_1(\lambda)$, $\gamma_2(\lambda)$, q_1, q_2 and q_3, we have $q_1 = q_3 = q^*$, $q_2 = \hat{q}$,

$$\gamma_1(\lambda) = \lambda s + 3\left(\frac{\lambda s}{2}\right)^{2/3} - 1 \text{ and } \gamma_2(\lambda) = 2\lambda s. \tag{A.5}$$

We then solve for the roots of $RHS(q) = 1$, given $m > \lambda s + 3(\frac{\lambda s}{2})^{2/3} - 1$. Note that $RHS(q) = 1$ is equivalent to $(q - \lambda s)\sqrt{m - q + 1} = \lambda s$. Taking the square of both sides of the equation, we can obtain the following cubic equation:

$$F(q) = q^3 - [(m+1) + 2\lambda s]q^2 + [2(m+1)\lambda s + (\lambda s)^2]q - m(\lambda s)^2 = 0. \tag{A.6}$$

It remains to find the roots of $F(q) = 0$ that are greater than λs and less than m. Let $a = 1$, $b = -[(m+1) + 2\lambda s]$, $c = 2(m+1)\lambda s + (\lambda s)^2$ and $d = -m(\lambda s)^2$. Denote

$$A = \frac{3ac - b^2}{3a^2} = -\frac{1}{3}(m + 1 - \lambda s)^2,$$
$$B = \frac{27a^2d - 9abc + 2b^3}{27a^3} = \frac{1}{27}\left[-2(m+1-\lambda s)^3 + 27(\lambda s)^2\right], \text{ and}$$
$$\delta = \left(\frac{B}{2}\right)^2 + \left(\frac{A}{3}\right)^3 = \frac{(\lambda s)^2}{108}\left[-4(m+1-\lambda s)^3 + 27(\lambda s)^2\right].$$

Because $m > \lambda s + 3(\frac{\lambda s}{2})^{2/3} - 1$, we have

$$\delta < \frac{(\lambda s)^2}{108}\left[-4\left(3\left(\frac{\lambda s}{2}\right)^{2/3}\right)^3 + 27(\lambda s)^2\right] = 0,$$

which implies that (A.6) admits three distinct real roots. Let $r = \sqrt{-\left(\frac{A}{3}\right)^3} = \frac{1}{27}(m+1-\lambda s)^3$ and

$$\theta = \frac{1}{3}\arccos\left(-\frac{B}{2r}\right) = \frac{1}{3}\arccos\left(1 - \frac{27(\lambda s)^2}{2(m+1-\lambda s)^3}\right). \tag{A.7}$$

Then, the three roots can be expressed as:

$$\begin{aligned} q_{r1} &= 2\sqrt[3]{r}\cos(\theta) - \frac{b}{3a}, \\ q_{r2} &= 2\sqrt[3]{r}\cos(\theta + \frac{2}{3}\pi) - \frac{b}{3a}, \quad \text{and} \\ q_{r3} &= 2\sqrt[3]{r}\cos(\theta + \frac{4}{3}\pi) - \frac{b}{3a}. \end{aligned}$$

In what follows, we will show that $q^* = q_{r3}$ and $\hat{q} = q_{r1}$. Because $m > \lambda s + 3(\frac{\lambda s}{2})^{2/3} - 1$, we have $-1 < \left(1 - \frac{27(\lambda s)^2}{2(m+1-\lambda s)^3}\right) < 1$, and thus $\theta \in (0, \frac{1}{3}\pi)$, where θ is given in (A.7). It follows that $q_{r2} < q_{r3} < q_{r1}$. Moreover, for a cubic equation which admits three real roots, we have

$$q_{r1} + q_{r2} + q_{r3} = -\frac{b}{a} = (m+1) + 2\lambda s > 0, \quad \text{and} \quad q_{r1} \cdot q_{r3} \cdot q_{r3} = -\frac{d}{a} = m(\lambda s)^2 > 0.$$

Therefore, we either have three positive roots, or one positive root and two negative roots.

When $\lambda s + 3\left(\frac{\lambda s}{2}\right)^{2/3} - 1 < m < 2\lambda s$, $RHS(q) = 1$ admits two positive roots in $[\lambda s, m]$. Therefore, $F(q) = 0$ must admit three positive roots. Because $q_{r2} < \lambda s$, we have $q^* = q_{r3}$ and $\hat{q} = q_{r1}$.

When $m > 2\lambda s$, because $F(m) = 2\lambda s - m < 0$ and $F(q) \to \infty$ as $q \to \infty$, there must exist a real root that is greater than m. Moreover, because $RHS(q) = 1$ admits a unique root in $[\lambda s, m]$. It follows that $F(q) = 0$ admits three positive roots and $q^* = q_{r3}$.

Therefore, we can obtain that

$$q_1 = q_3 = q_{r3} = \frac{2}{3}(m+1-\lambda s)\cos(\theta+\frac{4}{3}\pi) + \frac{1}{3}(m+1+2\lambda s) \text{ and} \tag{A.8}$$

$$q_2 = q_{r1} = \frac{2}{3}(m+1-\lambda s)\cos(\theta) + \frac{1}{3}(m+1+2\lambda s), \tag{A.9}$$

where θ is given in (A.7).

We then provide proofs for Theorem 2.3.2.A, 2.3.2.B and 2.3.2.C in the following subsections.

A.4.1 Proof of Theorem 2.3.2.A

Recall that $m_\lambda = \lfloor \alpha\lambda s \rfloor$ for $\alpha \in (0,1)$. Also recall that we denote by $SL(m,n)$ and $\pi_{m,n}(q)$ the service level and the probability of having q customers in a system with m vehicles and n drivers, and $SL(m,n) = 1 - \pi_{m,n}(n)$.

We first consider $SL(m_\lambda, m_\lambda)$. By (2.2)–(2.3), we have $\pi_{m_\lambda,m_\lambda}(m_\lambda - i) = \frac{\pi_{m_\lambda,m_\lambda}(m_\lambda)}{\prod_{k=0}^{i-1}\rho(m_\lambda-k)}$. It follows that

$$\sum_{q=1}^{m_\lambda} \pi_{m_\lambda,m_\lambda}(q) = \pi_{m_\lambda,m_\lambda}(m_\lambda)\left[1 + \sum_{i=0}^{m_\lambda-1}\prod_{k=0}^{i}\frac{1}{\rho(m_\lambda-k)}\right] = 1.$$

Because $m_\lambda = \lfloor \alpha\lambda s \rfloor$, by (A.4), we have $\rho(k) > \frac{1}{\alpha}$ for $1 \le k \le m_\lambda$. For convenience, let $a = \frac{1}{\alpha}$. Define $\tilde{\pi}_{m_\lambda,m_\lambda}(m_\lambda)$ as follows:

$$\tilde{\pi}_{m_\lambda,m_\lambda}(m_\lambda) = \frac{1}{\left[1 + \sum_{i=0}^{m_\lambda-1}\frac{1}{\rho(m_\lambda)a^i}\right]}. \tag{A.10}$$

Then we must have $\tilde{\pi}_{m_\lambda,m_\lambda}(m_\lambda) < \pi_{m_\lambda,m_\lambda}(m_\lambda)$ and $\lim_{\lambda\to\infty}\tilde{\pi}_{m_\lambda,m_\lambda}(m_\lambda) = \frac{2a-2}{2a-1}$.

We then consider $SL(m_\lambda, n)$. Let $H(n) = \frac{1}{\rho(n)} = \frac{n}{\lambda s\left(\frac{1}{\sqrt{m_\lambda-n+1}}+1\right)}$. By taking the first order derivative, we can obtain that

$$H'(n) = \frac{1}{\lambda s\left(\frac{1}{\sqrt{m_\lambda-n+1}}+1\right)^2}\left[(m_\lambda-n+1)^{-1/2} + 1 - \frac{n}{2}(m_\lambda-n+1)^{-3/2}\right].$$

Let $g(n) = (m_\lambda - n + 1)^{-1/2} + 1 - \frac{n}{2}(m_\lambda - n + 1)^{-3/2}$. We have $g'(n) = -\frac{3n}{4}(m_\lambda - n + 1)^{-5/2} < 0$. Because $g(0) = (m_\lambda + 1)^{-1/2} + 1 > 0$ and $g(m_\lambda) = 2 - \frac{m_\lambda}{2} < 0$ for $m_\lambda > 4$, $H(n)$ is first increasing and then decreasing. Hence, there exists a unique n^* which maximizes $H(n)$, and n^* is the solution to $g(n) = 0$.

Because $g(n) = 0$ is equivalent to $3(m_\lambda - n + 1) + 2(m_\lambda - n + 1)^{3/2} = m_\lambda + 1$, by letting $y = (m_\lambda - n + 1)^{1/2}$, it suffices to obtain the unique positive root of the following equation:

$$2y^3 + 3y^2 - (m_\lambda + 1) = 0.$$

By some algebra, we can obtain the root $y_r = \sqrt[3]{\frac{2m_\lambda+1}{8} + \frac{\sqrt{m_\lambda^2+m_\lambda}}{4}} + \sqrt[3]{\frac{2m_\lambda+1}{8} - \frac{\sqrt{m_\lambda^2+m_\lambda}}{4}} - \frac{1}{2}$. It follows that

$$n^* = m_\lambda + 1 - y^2 = (m_\lambda + 1) - \left[\sqrt[3]{\frac{2m_\lambda + 1}{8} + \frac{\sqrt{m_\lambda^2 + m_\lambda}}{4}} + \sqrt[3]{\frac{2m_\lambda + 1}{8} - \frac{\sqrt{m_\lambda^2 + m_\lambda}}{4}} - \frac{1}{2}\right]^2. \tag{A.11}$$

Notice that n^* depends on m_λ. For convenience, we do not express the dependence in the notation explicitly. We then show that for any $0 < \epsilon < 1$, $\lim_{\lambda\to\infty} \sum_{i=0}^{\lfloor(1-\epsilon)n^*\rfloor} \pi_{m_\lambda,\lfloor n^*\rfloor}(i) = 0$. Let $z = \lfloor(1-\epsilon)n^*\rfloor$ and let Δ be a fixed positive integer. We have

$$\begin{aligned}
\limsup_{\lambda\to\infty} \sum_{i=0}^{z} \pi_{m_\lambda,\lfloor n^*\rfloor}(i) &= \limsup_{\lambda\to\infty} \left[\frac{1 + \sum_{i=1}^{z}\prod_{k=1}^{i}\rho(k)}{1 + \sum_{i=1}^{\lfloor n^*\rfloor}\prod_{k=1}^{i}\rho(k)}\right] \\
&\overset{(a)}{\le} \limsup_{\lambda\to\infty} \left[\frac{1}{1 + \sum_{i=1}^{\lfloor n^*\rfloor}\prod_{k=1}^{i}\rho(k)} + \frac{\sum_{i=1}^{z}\prod_{k=1}^{i}\rho(k)}{\sum_{i=z+\Delta}^{\lfloor n^*\rfloor}\prod_{k=1}^{i}\rho(k)}\right] \\
&\overset{(b)}{\le} \limsup_{\lambda\to\infty} \left[\frac{1}{1 + \sum_{i=1}^{\lfloor n^*\rfloor}\prod_{k=1}^{i}\rho(k)} + \frac{z\prod_{k=1}^{z}\rho(k)}{(\lfloor n^*\rfloor - z - \Delta)\prod_{k=1}^{z+\Delta}\rho(k)}\right] \\
&\le \limsup_{\lambda\to\infty} \left[\frac{1}{1 + \sum_{i=1}^{\lfloor n^*\rfloor}\prod_{k=1}^{i}\rho(k)} + \frac{z}{(\lfloor \epsilon n^*\rfloor - \Delta)\prod_{k=z+1}^{z+\Delta}\rho(k)}\right] \\
&\overset{(c)}{\le} \limsup_{\lambda\to\infty} \left[\frac{1}{1 + \sum_{i=1}^{\lfloor n^*\rfloor}\prod_{k=1}^{i}\rho(k)} + \frac{z}{\lfloor \epsilon n^*\rfloor - \Delta}[\rho(z+\Delta)]^{-\Delta}\right],
\end{aligned} \tag{A.12}$$

where (a), (b) and (c) follow from Lemma A.1.1. Because $\rho(k) > 1$ for $1 \leq i \leq \lfloor n^* \rfloor$ and $n^* \to \infty$ as $\lambda \to \infty$, we have $\lim_{\lambda\to\infty} \left[\frac{1}{1+\sum_{i=1}^{\lfloor n^* \rfloor} \prod_{k=1}^{i} \rho(k)} \right] = 0$. Because

$$\lim_{\lambda\to\infty} \rho(z+\Delta) = \lim_{\lambda\to\infty} \frac{\lambda s}{\lfloor (1-\epsilon)n^* \rfloor + \Delta} \left(\frac{1}{\sqrt{m_\lambda - \lfloor (1-\epsilon)n^* \rfloor - \Delta + 1}} + 1 \right) = \frac{a}{1-\epsilon},$$

and Δ can be an arbitrarily large but fixed integer, the desired result follows.

Because $\lim_{\lambda\to\infty} \sum_{i=0}^{z} \pi_{m_\lambda, \lfloor n^* \rfloor}(i) = 0$, we have

$$\lim_{\lambda\to\infty} \left[\sum_{i=z}^{\lfloor n^* \rfloor} \pi_{m_\lambda, \lfloor n^* \rfloor}(i) \right] = \lim_{\lambda\to\infty} \left(\pi_{m_\lambda, \lfloor n^* \rfloor}(\lfloor n^* \rfloor) \left[1 + \sum_{i=0}^{\lfloor n^* \rfloor - z - 1} \prod_{k=0}^{i} \frac{1}{\rho(\lfloor n^* \rfloor - k)} \right] \right) = 1.$$

Let

$$\hat{\pi}_{m_\lambda, \lfloor n^* \rfloor}(\lfloor n^* \rfloor) = \pi_{m_\lambda, \lfloor n^* \rfloor}(\lfloor n^* \rfloor) \frac{1 + \sum_{i=0}^{\lfloor n^* \rfloor - z - 1} \prod_{k=0}^{i} \frac{1}{\rho(\lfloor n^* \rfloor - k)}}{1 + \sum_{i=0}^{\lfloor n^* \rfloor - z - 1} \left(\frac{1}{\rho(z)} \right)^{i+1}}. \tag{A.13}$$

We have

$$\lim_{\lambda\to\infty} \hat{\pi}_{m_\lambda, \lfloor n^* \rfloor}(\lfloor n^* \rfloor) = 1 - \frac{1-\epsilon}{a}.$$

Because $\frac{1}{\rho(q)}$ is concave by Lemma A.1.1 and it is maximized at n^*, we have

$$\frac{1 + \sum_{i=0}^{\lfloor n^* \rfloor - z - 1} \prod_{k=0}^{i} \frac{1}{\rho(\lfloor n^* \rfloor - k)}}{1 + \sum_{i=0}^{\lfloor n^* \rfloor - z - 1} \left(\frac{1}{\rho(z)} \right)^{i+1}} > 1,$$

and thus $\frac{1}{\rho(z)} < \frac{1}{\rho(i)}$ for $z < i < n^*$. It follows that $\hat{\pi}_{m_\lambda, \lfloor n^* \rfloor}(\lfloor n^* \rfloor) > \pi_{m_\lambda, \lfloor n^* \rfloor}(\lfloor n^* \rfloor)$. Therefore, $\liminf_{\lambda\to\infty} \frac{SL(m_\lambda, \lfloor n^* \rfloor)}{SL(m_\lambda, m_\lambda)} \geq \lim_{\lambda\to\infty} \frac{1-\hat{\pi}_{m_\lambda, \lfloor n^* \rfloor}(\lfloor n^* \rfloor)}{1-\tilde{\pi}_{m_\lambda, m_\lambda}(m_\lambda)} = (2 - \frac{1}{a})(1-\epsilon) = (2-\alpha)(1-\epsilon)$. Because ϵ can be arbitrarily small, the desired result follows.

A.4.2 Proof of Theorem 2.3.2.B

Recall from Proposition 2.3.1 that, when $\gamma_1(\lambda) < m < \gamma_2(\lambda)$, the stationary distribution $\pi_{m,m}(q)$ for $q \in \{0, \cdots, m\}$ is bimodal with one mode at $\lfloor q_1 \rfloor$ and the other at m. Also recall that $m_\lambda = \lfloor \alpha \lambda s \rfloor$, where $\alpha \in (1, 2)$. To prove Theorem 2.3.2.B, it suffices to show

that $\frac{\pi_{m_\lambda,m_\lambda}(\lfloor q_1 \rfloor)}{\pi_{m_\lambda,m_\lambda}(m_\lambda)} \to \infty$ as $\lambda \to \infty$. By (2.2)–(2.3), we have

$$\frac{\pi_{m_\lambda,m_\lambda}(\lfloor q_1 \rfloor)}{\pi_{m_\lambda,m_\lambda}(m_\lambda)} = \frac{1}{\prod_{k=\lfloor q_1 \rfloor+1}^{m_\lambda} \rho(k)} = \frac{1}{\left[\prod_{k=\lfloor q_1 \rfloor+1}^{\lfloor q_2 \rfloor} \rho(k)\right]\left[\prod_{k=\lfloor q_2 \rfloor+1}^{m_\lambda} \rho(k)\right]}.$$

We prove Theorem 2.3.2.B in two steps. In step (i), we show that $\limsup_{\lambda\to\infty} \prod_{k=\lfloor q_2 \rfloor+1}^{m_\lambda} \rho(k) < \infty$. In step (ii), we show that $\lim_{\lambda\to\infty} \prod_{k=\lfloor q_1 \rfloor+1}^{\lfloor q_2 \rfloor} \rho(k) = 0$.

Step (i). By Lemma A.1.1, for $q_2 < q \le m_\lambda$, $\rho(q) \le \rho(m_\lambda) = \frac{2\lambda s}{m_\lambda}$. It suffice to show that $\limsup_{\lambda\to\infty} [m_\lambda - \lfloor q_2 \rfloor] < \infty$. By Lemma A.1.1 we have $\rho(q_2) = \frac{\lambda s}{q_2}\left(\frac{1}{\sqrt{m_\lambda - q_2 + 1}} + 1\right) = 1$. For any fixed positive number k, let $q_k = m_\lambda - k$. Then, we have $\lim_{\lambda\to\infty} \frac{\lambda s}{q_k} = \frac{1}{\alpha}$. Therefore, for $k > \frac{4}{(\alpha-1)^2}$, $\limsup_{\lambda\to\infty} \rho(q_k) < \frac{1+\alpha}{2\alpha} < 1$ and thus $q_k < q_2$ for sufficiently large λ by Lemma A.1.1. It follows that $\limsup_{\lambda\to\infty} [m_\lambda - \lfloor q_2 \rfloor] < \infty$.

Step (ii). By (A.8) – (A.9), for any small $\epsilon > 0$, $q_2(1-\epsilon) - q_1(1+\epsilon) = \Theta(m)$. Because $\rho(q_1) = \frac{\lambda s}{q_1}\left(\frac{1}{\sqrt{m - q_1 + 1}} + 1\right) = 1$. We have

$$\lim_{\lambda\to\infty} \rho((1+\epsilon)q_1) = \lim_{\lambda\to\infty} \frac{\lambda s}{(1+\epsilon)q_1}\left(\frac{1}{\sqrt{m_\lambda - (1+\epsilon)q_1 + 1}} + 1\right) = \frac{1}{1+\epsilon}.$$

Moreover, $\rho(q_2) = \frac{\lambda s}{q_2}\left(\frac{1}{\sqrt{m_\lambda - q_2 + 1}} + 1\right) = 1$. From step (i), for $k > \frac{4}{(\alpha-1)^2}$ and $q_k = m_\lambda - k$, we have $q_k < q_2$ and $\rho(q_k) = \frac{\lambda s}{q_k}\left(\frac{1}{\sqrt{m_\lambda - q_k + 1}} + 1\right) < 1$ when λ is sufficiently large. It follows that $\frac{\lambda s}{q_2}\left(\frac{1}{\sqrt{m_\lambda - q_k + 1}} + 1\right) < 1$, which implies that $\frac{\lambda s}{(1-\epsilon)q_2} < \frac{1}{(\frac{1}{\sqrt{k+1}}+1)(1-\epsilon)}$. Because $m_\lambda - (1-\epsilon)q_2 \to \infty$ as $\lambda \to \infty$, $\limsup_{\lambda\to\infty} \rho((1-\epsilon)q_2) \le \frac{1}{(\frac{1}{\sqrt{k+1}}+1)(1-\epsilon)}$. By Lemma A.1.1, we have

$$\begin{aligned}
\limsup_{\lambda\to\infty}\left[\prod_{k=\lfloor q_1 \rfloor+1}^{\lfloor q_2 \rfloor} \rho(k)\right] &\le \limsup_{\lambda\to\infty}\left[\prod_{k=\lfloor (1+\epsilon)q_1 \rfloor}^{\lfloor (1-\epsilon)q_2 \rfloor} \rho(k)\right] \\
&\le \limsup_{\lambda\to\infty} \left\{\max[\rho(\lfloor (1+\epsilon)q_1 \rfloor), \rho(\lfloor (1-\epsilon)q_2 \rfloor)]\right\}^{\lfloor (1-\epsilon)q_2 \rfloor - \lfloor (1+\epsilon)q_1 \rfloor} \\
&\le \limsup_{\lambda\to\infty}\left[\max\left(\frac{1}{1+\epsilon}, \frac{1}{(\frac{1}{\sqrt{k+1}}+1)(1-\epsilon)}\right)\right]^{\lfloor (1-\epsilon)q_2 \rfloor - \lfloor (1+\epsilon)q_1 \rfloor}.
\end{aligned}$$

Because ϵ can be arbitrarily small, the desired result follows.

A.4.3 Proof of Theorem 2.3.2.C

Recall that $m_\lambda = \lfloor \alpha\lambda s \rfloor$ for $\alpha > 2$. Theorem 2.3.2.C can be implied by the following lemma.

Lemma A.4.1. *Let $m_\lambda = \lfloor \alpha\lambda s \rfloor$ for $\alpha > 2$. For any small $\epsilon > 0$ and $n_\epsilon = \lfloor (1+\epsilon)q_3 \rfloor$, we have*

$$\lim_{\lambda\to\infty} \left[\sum_{i=\lfloor (1+\frac{\epsilon}{2})q_3 \rfloor}^{n_\epsilon} \pi_{m_\lambda,n_\epsilon}(i) + \sum_{i=0}^{\lfloor (1-\frac{\epsilon}{2})q_3 \rfloor} \pi_{m_\lambda,n_\epsilon}(i) \right] = 0,$$

where q_3 is given in (A.8).

We first show that $\lim_{\lambda\to\infty} \sum_{i=\lfloor (1+\frac{\epsilon}{2})q_3 \rfloor}^{n_\epsilon} \pi_{m_\lambda,n_\epsilon}(i) = 0$. For convenience, let $z = \lfloor (1+\frac{\epsilon}{2})q_3 \rfloor$ and let Δ be a fixed positive integer. By (2.2)–(2.3), we have

$$\begin{aligned}
\limsup_{\lambda\to\infty} \sum_{i=z}^{n_\epsilon} \pi_{m_\lambda,n_\epsilon}(i) &= \limsup_{\lambda\to\infty} \left[\frac{\sum_{i=z}^{n_\epsilon} \prod_{k=1}^{i} \rho(k)}{1 + \sum_{i=1}^{n_\epsilon} \prod_{k=1}^{i} \rho(k)} \right] \\
&\le \limsup_{\lambda\to\infty} \left[\frac{\sum_{i=z+\Delta}^{n_\epsilon} \prod_{k=1}^{i} \rho(k)}{\sum_{i=\lfloor q_3 \rfloor}^{z} \prod_{k=1}^{i} \rho(k)} \right] \\
&\le \limsup_{\lambda\to\infty} \left[\frac{(n_\epsilon - z - \Delta) \prod_{i=1}^{z+\Delta} \rho(k)}{(z - \lfloor q_3 \rfloor) \prod_{k=1}^{z} \rho(k)} \right] \\
&= \limsup_{\lambda\to\infty} \left[\frac{(n_\epsilon - z - \Delta) \prod_{k=z+1}^{z+\Delta} \rho(k)}{z - q_3} \right] \\
&\le \limsup_{\lambda\to\infty} \frac{\frac{\epsilon}{2}q_3 - \Delta}{\frac{\epsilon}{2}q_3} \left[\max\{\rho(z+1), \rho(z+\Delta)\} \right]^\Delta ,
\end{aligned}$$

where the above inequalities are due to Lemma A.1.1. Because $\rho(q_3) = \frac{\lambda s}{q_3}(\frac{1}{\sqrt{m_\lambda - q_3 + 1}} + 1) = 1$ and $m_\lambda - q_3 \to \infty$ as $\lambda \to \infty$. We have $\lim_{\lambda\to\infty} \rho(z+\Delta) = \frac{1}{1+\frac{\epsilon}{2}} < 1$. Moreover, because Δ can be an arbitrarily large but fixed integer, the desired result follows.

We then show that $\lim_{\lambda\to\infty} \sum_{i=0}^{\lfloor (1-\frac{\epsilon}{2})q_3 \rfloor} \pi_{m_\lambda,n_\epsilon}(i) = 0$. Abusing notation, let $z = \lfloor (1-\frac{\epsilon}{2})q_3 \rfloor$. Let Δ be fixed positive integer. By Lemma A.1.1 and the analysis in the Proof of Theorem

2.3.2.A (see (A.12)), we have

$$\limsup_{\lambda\to\infty}\sum_{i=0}^{z}\pi_{m_\lambda,n_\epsilon}(i)\le\limsup_{\lambda\to\infty}\left[\frac{1}{1+\sum_{i=1}^{n_\epsilon}\prod_{k=1}^{i}\rho(k)}+\frac{z}{\lfloor\frac{\epsilon}{2}q_3\rfloor-\Delta}[\rho(z+\Delta)]^{-\Delta}\right],$$

where $\rho(\cdot)$ is given in (A.4). By Lemma A.1.1, $\lim_{\lambda\to\infty}\frac{1}{1+\sum_{i=1}^{n_\epsilon}\prod_{k=1}^{i}\rho(k)}=0$. Because $\rho(q_3)=\frac{\lambda s}{q_3}\left(\frac{1}{\sqrt{m-q_3+1}}+1\right)=1$ and $\lim_{\lambda\to\infty}\frac{1}{\sqrt{m_\lambda-q_3+1}}=0$, we have $\lim_{\lambda\to\infty}\frac{\lambda s}{q_3}=1$. Then, we must have

$$\lim_{\lambda\to\infty}\rho(z+\Delta)=\lim_{\lambda\to\infty}\frac{\lambda s}{z+\Delta}\left(\frac{1}{\sqrt{m_\lambda-z-\Delta+1}}+1\right)=\frac{1}{1-\frac{\epsilon}{2}}.$$

Because Δ can be an arbitrarily large but fixed integer, the desired result follows.

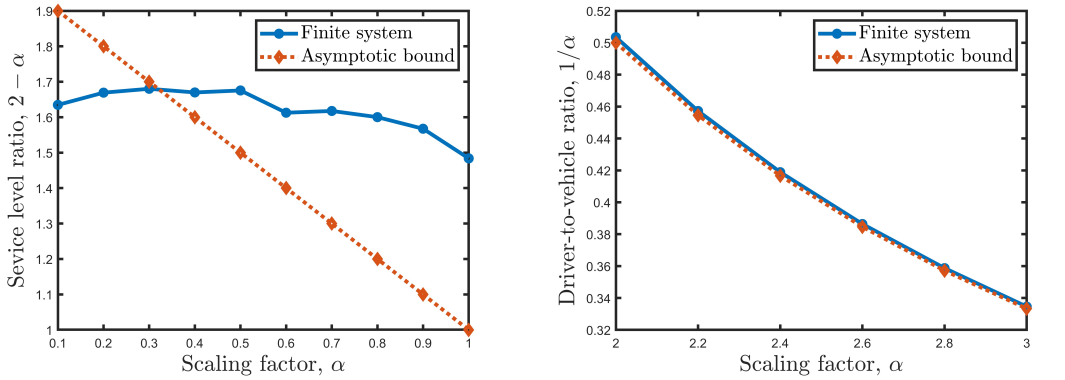
Because $\lim_{\lambda\to\infty}\frac{q_3}{\lambda s}=1$, Theorem 2.3.2.C follows immediately as $\limsup_{\lambda\to\infty}\frac{(1+\epsilon)q_3}{m_\lambda}\le\frac{1}{\alpha}(1+\epsilon)$, and ϵ can be arbitrarily small.

A.4.4 Comparisons between Asymptotic Bounds and Finite System Ratios

In this section, we provide comparisons between finite system ratios (i.e., service level ratio in the supply-limited regime and driver-to-vehicle ratio in the supply-rich regime) derived from simulations of example systems and their asymptotic bounds.

Panel (a) in Figure A.1 presents a comparison between the asymptotic lower bound and the simulation-derived ratio of the optimal service level for a tele-driving system relative to a system with an equal number of drivers and vehicles. The simulation is based on an example system with $\mu(m,q)=\left[\frac{s}{\sqrt{m-q+1}}+s\right]^{-1}$, $\lambda=1000$, $s=10$, and $m=\lfloor\alpha\lambda s\rfloor$. The results demonstrate that when α takes large values, the asymptotic lower bound underestimates the actual improvement in service level observed in finite systems. Conversely, for small values of α, the asymptotic lower bound overestimates the improvement in finite systems (significantly large system sizes are required for the service level ratio to surpass the lower bound).

Panel (b) in Figure A.1 presents a comparison between the asymptotic upper bound and the simulation-derived minimum driver-to-vehicle ratio required to maintain a service level of at least 99% relative to a system with an equal number of drivers and vehicles. The



(a) Service level ratio (supply-limited regime) (b) Driver-to-vehicle ratio (supply-rich regime)

Figure A.1: Impact of α on service level ratio (supply-limited regime) and driver-to-vehicle ratio (supply-rich regime)

simulation is based on an example system with $\mu(m,q) = \left[\frac{s}{\sqrt{m-q+1}} + s\right]^{-1}$, $\lambda = 1000$, $s = 10$, and $m = \lfloor \alpha\lambda s \rfloor$. The results indicate that the asymptotic bound closely approximates the actual driver-to-vehicle ratio observed in finite systems.

A.5 Proofs for Systems with Patient Customers

A.5.1 Proof of Lemma 2.4.1

The system with patient customers is stable if and only if the utilization of the system is less than 1, which is equivalent to that there exists a unique stationary distribution of the underlying Markov chain as it is irreducible and aperiodic. Therefore, the system is stable if and only if the term defined in (2.6) is positive and finite, which is equivalent to $\rho(n) < 1$.

A.5.2 Proof of Proposition 2.4.2

By virtue of Little's Law, it suffices to show the same monotonicity result for the long run average number of customers in system, which we denote by $E[Q(m,n)]$. Let $f(n) =$

$\Pi_{i=1}^{n}\rho(i)$. We have

$$E[Q(m,n)] = \frac{\overbrace{\sum_{i=1}^{n} if(i)}^{a} + \overbrace{\sum_{i=1}^{\infty}(n+i)f(n)[\rho(n)]^i}^{c}}{\underbrace{1+\sum_{i=1}^{n} f(i)}_{b} + \underbrace{f(n)\frac{\rho(n)}{1-\rho(n)}}_{d}}, \tag{A.14}$$

and

$$E[Q(m,n+1)] = \frac{\overbrace{\sum_{i=1}^{n} if(i)}^{a} + \overbrace{(n+1)f(n+1) + \sum_{i=1}^{\infty}(n+1+i)f(n+1)[\rho(n+1)]^i}^{e}}{\underbrace{1+\sum_{i=1}^{n} f(i)}_{b} + \underbrace{f(n+1)+f(n+1)\frac{\rho(n+1)}{1-\rho(n+1)}}_{f}}. \tag{A.15}$$

Recall that we define $\tilde{q}$ in Section 2.4 as the unique solution to $\frac{d\rho(q)}{dq} = 0$. Because $\rho(q)$ is strictly convex by Lemma A.1.1, $\rho(n)$ is decreasing when $n < \tilde{q}$ and it is increasing when $n > \tilde{q}$. Define a, b, c, d, e and f as illustrated in (A.14)–(A.15). When $n+1 < \tilde{q}$, we have

$$\frac{c}{d} = \frac{\frac{n\rho(n)+\frac{\rho(n)}{1-\rho(n)}}{1-\rho(n)}}{\frac{\rho(n)}{1-\rho(n)}} = n + \frac{1}{1-\rho(n)} > n + \frac{1}{1-\rho(n+1)} = \frac{\frac{n+1+\frac{\rho(n+1)}{1-\rho(n+1)}}{1-\rho(n+1)}}{1+\frac{\rho(n+1)}{1-\rho(n+1)}} = \frac{e}{f},$$

and $d > f$. Therefore, we have $E[Q(m,n)] > E[Q(m,n+1)]$. By a similar argument, we can show that when $n > \tilde{q}$, $E[Q(m,n)] < E[Q(m,n+1)]$ and thus the desired result follows.

A.5.3 Proof of Proposition 2.4.3

Recall that $m_\lambda = \lfloor \alpha\lambda s \rfloor$ for $\alpha > 1$, $n_\lambda = \lfloor \beta m_\lambda \rfloor$ for $\beta \in (0,1]$ and the stability condition in Lemma 2.4.1 is satisfied. To prove Proposition 2.4.3, it suffices to show that for any small

$\delta > 0$,

$$\lim_{\lambda\to\infty} \frac{1}{\lambda} \left[\sum_{i=0}^{\lfloor (1-\delta)q^* \rfloor} i\pi_{m_\lambda,n_\lambda}(i) + \sum_{i=\lfloor (1+\delta)q^* \rfloor}^{\infty} i\pi_{m_\lambda,n_\lambda}(i) \right] = 0. \tag{A.16}$$

We first show that $\lim_{\lambda\to\infty} \frac{1}{\lambda} \sum_{i=0}^{\lfloor (1-\delta)q^* \rfloor} i\pi_{m_\lambda,n_\lambda}(i) = 0$. Let $z = \lfloor (1-\delta)q^* \rfloor$ and let Δ be a fixed positive integer. We have

$$\begin{aligned}
\limsup_{\lambda\to\infty} \sum_{i=0}^{z} \pi_{m_\lambda,n_\lambda}(i) &= \limsup_{\lambda\to\infty} \left[\frac{1 + \sum_{i=1}^{z} \prod_{k=1}^{i} \rho(k)}{1 + \sum_{i=1}^{n_\lambda} \prod_{k=1}^{i} \rho(k) + \frac{\rho(n_\lambda)}{1-\rho(n_\lambda)} \prod_{k=1}^{n_\lambda} \rho(k)} \right] \\
&\le \limsup_{\lambda\to\infty} \left[\frac{1 + \sum_{i=1}^{z} \prod_{k=1}^{i} \rho(k)}{1 + \sum_{i=1}^{n_\lambda} \prod_{k=1}^{i} \rho(k)} \right] \\
&\le \limsup_{\lambda\to\infty} \left[\frac{1}{1 + \sum_{i=1}^{n_\lambda} \prod_{k=1}^{i} \rho(k)} + \frac{\sum_{i=1}^{z} \prod_{k=1}^{i} \rho(k)}{\sum_{i=z+\Delta}^{\lfloor q^* \rfloor} \prod_{k=1}^{i} \rho(k)} \right] \\
&\le \limsup_{\lambda\to\infty} \left[\frac{1}{1 + \sum_{i=1}^{n_\lambda} \prod_{k=1}^{i} \rho(k)} + \frac{z}{\lfloor \delta q^* \rfloor - \Delta} [\rho(z+\Delta)]^{-\Delta} \right],
\end{aligned}$$

where the last inequality follows from Lemma A.1.1. Because $\rho(q^*) = \frac{\lambda s}{q^*} \left(\frac{1}{\sqrt{m_\lambda - q^* + 1}} + 1 \right) = 1$ and $m_\lambda - q^* \to \infty$ as $\lambda \to \infty$, we have $\lim_{\lambda\to\infty} \rho(z+\Delta) = \frac{1}{1-\delta}$. Moreover, because Δ can be an arbitrarily large but fixed integer, we have $\lim_{\lambda\to\infty} \sum_{i=0}^{z} \pi_{m_\lambda,n_\lambda}(i) = 0$, which implies that $\lim_{\lambda\to\infty} \frac{1}{\lambda} \sum_{i=0}^{z} i\pi_{m_\lambda,n_\lambda}(i) \le \lim_{\lambda\to\infty} \frac{n_\lambda}{\lambda} \sum_{i=1}^{z} \pi_{m_\lambda,n_\lambda}(i) = \alpha\beta s \lim_{\lambda\to\infty} \sum_{i=0}^{z} \pi_{m_\lambda,n_\lambda}(i) = 0$.

We then show that $\lim_{\lambda\to\infty} \frac{1}{\lambda} \sum_{i=\lfloor (1+\delta)q^* \rfloor}^{n_\lambda - 1} i\pi_{m_\lambda,n_\lambda}(i) = 0$. Abusing notation, let $z =$

$\lfloor (1+\delta)q^* \rfloor$. We have

$$\begin{aligned}
\limsup_{\lambda\to\infty} \sum_{i=z}^{n_\lambda-1} \pi_{m_\lambda,n_\lambda}(i) &= \limsup_{\lambda\to\infty} \frac{\sum_{i=z}^{n_\lambda-1} \prod_{k=1}^{i} \rho(k)}{1+\sum_{i=1}^{n_\lambda} \prod_{k=1}^{i} \rho(k) + \frac{\rho(n_\lambda)}{1-\rho(n_\lambda)} \prod_{k=1}^{n_\lambda} \rho(k)} \\
&\le \limsup_{\lambda\to\infty} \frac{\sum_{i=z}^{n_\lambda-1} \prod_{k=1}^{i} \rho(k)}{\sum_{i=\lfloor q^* \rfloor}^{z} \prod_{k=1}^{i} \rho(k)} \\
&\overset{\text{(a)}}{\le} \limsup_{\lambda\to\infty} \frac{\sum_{i=z}^{n_\lambda-1} \prod_{k=1}^{i} \rho(k)}{(z-\lfloor q^* \rfloor) \prod_{k=1}^{z} \rho(k)} \\
&\le \limsup_{\lambda\to\infty} \frac{\sum_{i=z+\Delta}^{n_\lambda-1} \prod_{k=1}^{i} \rho(k)}{(z-\lfloor q^* \rfloor) \prod_{k=1}^{z} \rho(k)} \\
&\overset{\text{(b)}}{\le} \limsup_{\lambda\to\infty} \frac{(n_\lambda-1-z-\Delta) \prod_{k=1}^{z+\Delta} \rho(k)}{(z-\lfloor q^* \rfloor) \prod_{k=1}^{z} \rho(k)} \\
&= \limsup_{\lambda\to\infty} \frac{n_\lambda-1-z-\Delta}{z-\lfloor q^* \rfloor} \prod_{k=z+1}^{z+\Delta} \rho(k) \\
&\overset{\text{(c)}}{\le} \limsup_{\lambda\to\infty} \frac{n_\lambda-1-z-\Delta}{z-\lfloor q^* \rfloor} \left[\max\{\rho(z+1), \rho(z+\Delta)\}\right]^{\Delta},
\end{aligned}$$

where inequality (a), (b) and (c) are due to Lemma A.1.1. Because $\limsup_{\lambda\to\infty} \frac{n_\lambda-1-z-\Delta}{z-\lfloor q^* \rfloor} < \infty$, $\lim_{\lambda\to\infty} \rho(z+\Delta) = \frac{1}{1+\delta} < 1$ and Δ can be an arbitrarily large but fixed integer, we have $\lim_{\lambda\to\infty} \sum_{i=z}^{n_\lambda-1} \pi_{m_\lambda,n_\lambda}(i) = 0$. It follows that

$$\limsup_{\lambda\to\infty} \frac{1}{\lambda} \sum_{i=z}^{n_\lambda-1} i\pi_{m_\lambda,n_\lambda}(i) \le \limsup_{\lambda\to\infty} \frac{n_\lambda}{\lambda} \sum_{i=z}^{n_\lambda-1} \pi_{m_\lambda,n_\lambda}(i) = \alpha\beta s \limsup_{\lambda\to\infty} \sum_{i=z}^{n_\lambda-1} \pi_{m_\lambda,n_\lambda}(i) = 0.$$

Lastly, we show that $\lim_{\lambda\to\infty} \frac{1}{\lambda} \sum_{i=n_\lambda}^{\infty} i\pi_{m_\lambda,n_\lambda}(i) = 0$. We have

$$\begin{aligned}
&\limsup_{\lambda\to\infty} \frac{1}{\lambda} \sum_{i=n_\lambda}^{\infty} i\pi_{m_\lambda,n_\lambda}(i) \\
&= \limsup_{\lambda\to\infty} \frac{1}{\lambda} \left[\frac{\left[\prod_{k=1}^{n_\lambda} \rho(k)\right] \sum_{i=n_\lambda}^{\infty} i[\rho(n_\lambda)]^{i-n_\lambda}}{1+\sum_{i=1}^{n_\lambda} \prod_{k=1}^{i} \rho(k) + \left[\prod_{k=1}^{n_\lambda} \rho(k)\right] \sum_{i=n_\lambda+1}^{\infty} [\rho(i)]^{i-n_\lambda}} \right] \\
&= \limsup_{\lambda\to\infty} \frac{1}{\lambda} \left[\frac{n_\lambda}{1-\rho(n_\lambda)} + \frac{\rho(n_\lambda)}{(1-\rho(n_\lambda))^2} \right] \frac{\prod_{k=1}^{n_\lambda} \rho(k)}{1+\sum_{i=1}^{n_\lambda} \prod_{k=1}^{i} \rho(k) + \left[\prod_{k=1}^{n_\lambda} \rho(k)\right] \frac{\rho(n_\lambda)}{1-\rho(n_\lambda)}}.
\end{aligned}$$

By the definition of n_λ and because the stability condition in Lemma A.1.1 holds, we have

$$\limsup_{\lambda\to\infty} \frac{1}{\lambda}\left[\frac{n_\lambda}{1-\rho(n_\lambda)} + \frac{\rho(n_\lambda)}{(1-\rho(n_\lambda))^2}\right] < \infty.$$

Moreover, by Lemma A.1.1, we have

$$\begin{aligned}
&\limsup_{\lambda\to\infty} \frac{\prod_{k=1}^{n_\lambda}\rho(k)}{1+\sum_{i=1}^{n_\lambda}\prod_{k=1}^{i}\rho(k) + \left[\prod_{k=1}^{n_\lambda}\rho(k)\right]\frac{\rho(n_\lambda)}{1-\rho(n_\lambda)}} \\
&\le \limsup_{\lambda\to\infty} \frac{\prod_{k=1}^{n_\lambda}\rho(k)}{(n_\lambda - \lfloor q^*\rfloor)\prod_{k=1}^{n_\lambda}\rho(k)} \\
&= \limsup_{\lambda\to\infty} \frac{1}{n_\lambda - \lfloor q^*\rfloor} \\
&= 0.
\end{aligned}$$

It follows that $\lim_{\lambda\to\infty}\frac{1}{\lambda}\sum_{i=n_\lambda+1}^{\infty} i\pi_{m_\lambda,n_\lambda}(i) = 0$.

A.5.4 Proof of Proposition 2.4.4

Recall that we show (A.16) holds for any $\delta > 0$ in the proof of Proposition 2.4.3. Because $\limsup\frac{(1+\delta)q_3}{m_\lambda} \le \frac{1}{\alpha}(1+\delta)$, where q_3 is define in (A.8), and δ can be arbitrarily small, the result follows directly.

A.5.5 Systems with Imperfectly Patient Customers

In this section, we consider the case where customers are imperfectly patient. We assume that customers are willing to wait but only up to a threshold that is exponentially distributed with rate ζ. In this case, the system dynamic is a birth and death process where the birth rate is λ, and the death rate is $\mu(m,q)$ if $q \le n$, and is $\mu(m,n)+(q-n)\zeta$ otherwise (there are $q-n$ customers waiting to be matched). We provide simulation results in Figure A.2 for an example system where $\mu(m,q)$ is given by (2.4). Abusing notation, let

$$\rho(q) = \begin{cases} \frac{\lambda}{q}\left(\frac{s}{\sqrt{m-n+1}} + s\right), & \text{if } q < n, \\ \frac{\lambda}{n(\frac{s}{\sqrt{m-n+1}}+s)^{-1}+(q-n)\zeta}, & \text{otherwise.} \end{cases}$$

The stationary distribution of the system is given by

$$\pi_{m,n}(0) = \left[1 + \sum_{q=1}^{\infty} \prod_{k=1}^{i} \rho(k)\right]^{-1}, \quad \text{and} \quad \pi_{m,n}(i) = \pi_{m,n}(0) \prod_{k=1}^{i} \rho(i).$$

The service level is given by

$$SL(m,n) = 1 - \frac{\gamma}{\lambda} \sum_{q=n+1}^{\infty} (q-n)\pi_{m,n}(q).$$

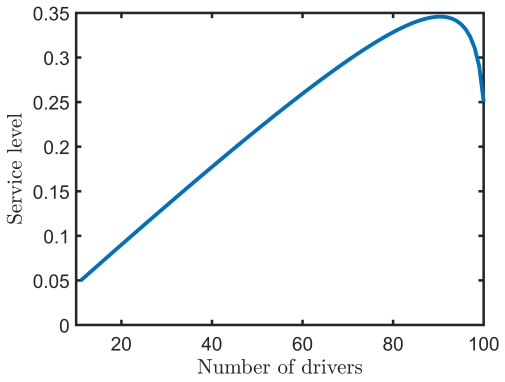
In Figure A.2, we consider an example system with $\lambda = 20$ and $s = 10$ to illustrate the impact of the number of drivers on service level and customer delay. We plot the service level and customer delay for varying numbers of drivers, as shown in Panels (a) and (b) for the supply-limited regime, and Panels (c) and (d) for the supply-rich regime. As we can see, reducing the number of drivers in the supply-limited regime improves service level and reduces customer delay. On the other hand, in the supply-rich regime, it is possible to significantly reduce the number of drivers without significantly affecting service level and customer delay.

A.6 Comparing Systems with Remote Drivers and Systems with in-Vehicle Drivers

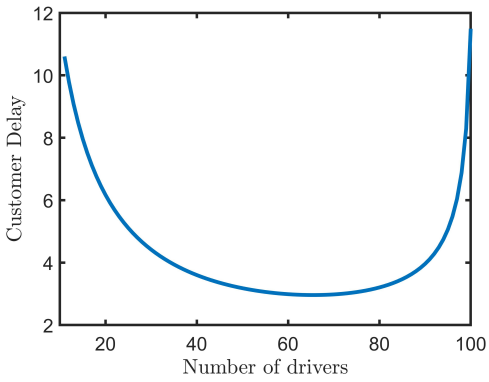
In Section 2.5, we discuss various aspects of the modeling, and in this section, we offer theoretical and numerical evidence to support our claims.

A.6.1 Slower Speed with Remote Drivers

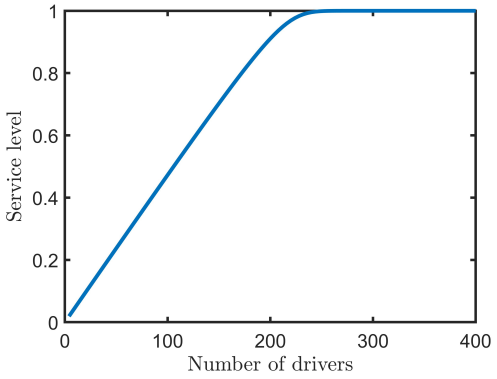
Recall that we define $\gamma_1(\lambda)$, $\gamma_2(\lambda)$, q_1, q_2 and q_3 in Proposition 2.3.1. When the service rate is scaled down by a factor $\zeta \in (0,1]$ (i.e., the service rate in systems with remote drivers is $\zeta\mu(m,q)$), we can show an analogous result to Proposition 2.3.1. In particular, there exist $\gamma_1^\zeta(\lambda)$ and $\gamma_2^\zeta(\lambda)$, such that the system is in the supply-limited regime if $m < \gamma_1^\zeta(\lambda)$, is in the intermediate regime if $\gamma_1^\zeta(\lambda) < m < \gamma_2^\zeta(\lambda)$, and is in the supply-rich regime if $m > \gamma_2^\zeta(\lambda)$. We define q_1^ζ, q_2^ζ, and q_3^ζ similarly (analogous to q_1, q_2, and q_3). We denote



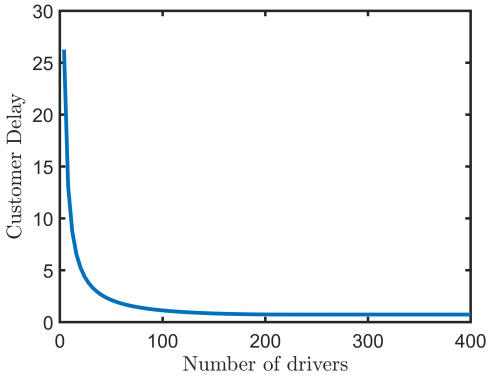
(a) Service level versus number of drivers in the supply-limited regime ($m = 100$)



(b) Customer delay versus number of drivers in the supply-limited regime ($m = 100$)



(c) Service level versus number of drivers in the supply-rich regime ($m = 400$)



(d) Customer delay versus number of drivers in the supply-rich regime ($m = 400$)

Figure A.2: Simulation results for systems with customer reneging (Parameters: $\lambda = 20$, $s = 10$, $\zeta = 2$).

by $SL^{\zeta}(m,n)$ the service level for the system with remote drivers, and let $W^{\zeta}(m,n)$ and $\pi^{\zeta}_{m,n}(q)$ be similarly defined.

In Corollary A.6.1, we present results for systems with impatient customers.

Corollary A.6.1. *If the service rate in systems with remote drivers is $\zeta\mu(m,q)$,*

(i) when $m \leq \gamma_2^{\zeta}(\lambda)$ and Condition (2.5) holds, there exists a threshold $\overline{\zeta}_1(m,\lambda) < 1$ such that $\max_{n\in\{1,\cdots,m\}} SL^{\zeta}(m,n) > SL(m,m)$ if and only if $\zeta > \overline{\zeta}_1(m,\lambda)$;

(ii) when $m > \gamma_2^{\zeta}(\lambda)$, for any $n > q_3^{\zeta}$, $SL^{\zeta}(m,n)$ is lower bounded by $1 - \frac{1}{n-\lfloor q_3^{\zeta}\rfloor}$.

In Corollary A.6.2, we present results for systems with patient customers.

Corollary A.6.2. *When $\gamma_1(\lambda) < m < \gamma_2(\lambda)$, there exists a threshold $\overline{\zeta}_2(m,\lambda) < 1$ such that switching to a tele-driving system with fewer drivers can stabilize an otherwise unstable system if and only if $\gamma > \overline{\zeta}_2(m,\lambda)$.*

Proof of Corollary A.6.1. Case (i). By Theorem 2.3.1.A and Theorem 2.3.1.B, we have $\max_{n\in\{1,\cdots,m\}} SL(m,n) > SL(m,m)$ if and only if (2.5) holds. Therefore, it suffices to show that $\max_{n\in\{1,\cdots,m\}} SL^{\zeta}(m,n)$ is monotonically increasing in ζ. Because

$$\frac{1}{\pi_{m,n}(n)} = 1 + \sum_{i=0}^{n-1} \left(\prod_{k=n-i}^{n} \frac{\rho(q)}{\zeta} \right)^{-1},$$

the result follows naturally.

Case (ii). The proof is the same as that of Theorem 2.3.1.C. ∎

Proof of Corollary A.6.2. By the proof of Lemma A.1.1, $\gamma_1^{\zeta}(\lambda)$ is the unique solution (on m) to $\min_{q\in[1,m]} \frac{\lambda}{q\gamma\mu(m,q)} = 1$, and $\gamma_2^{\zeta}(\lambda)$ is the unique solution to $\frac{\lambda}{m\gamma\mu(m,m)} = 1$. Because $\mu(m,q)$ is increasing in m and $\mu(m,m)$ is invariant in m, it follows that $\gamma_1^{\zeta}(\lambda)$ and $\gamma_2^{\zeta}(\lambda)$ are decreasing in ζ. By Corollary 2.4.1 and Proposition 2.4.1, switching to a tele-driving system can stabilize an otherwise unstable system if the system is unstable with in person-drivers (i.e., $m < \gamma_2(\lambda)$) and it can be stabilized with tele-driving (i.e., $m > \gamma_1^{\zeta}(\lambda)$). The desired result then follows from the monotonicity result on $\gamma_1^{\zeta}(\lambda)$ and $\gamma_2^{\zeta}(\lambda)$. ∎

In the asymptotic regime, we quantify the benefit from switching to a tele-driving system.

Corollary A.6.3. *Assume the state-dependent service rate in systems with remote drivers is given by $\zeta\mu(m,q)$, where $\mu(m,q)$ is defined in (2.4). Let $m_\lambda = \lfloor \frac{\alpha}{\zeta}\lambda s \rfloor$, we have the following results.*

For systems with impatient customers,

(i) if $\frac{\alpha}{\zeta} < 1$, $\liminf_{\lambda\to\infty}\left[\max_{n\in\{1,\cdots,m_\lambda\}}\frac{SL^\zeta(m_\lambda,n)}{SL(m_\lambda,m_\lambda)}\right] > 2\zeta-\alpha$; and

(ii) if $\alpha > 2$, for any $\epsilon > 0$, $\limsup_{\lambda\to\infty}\frac{n_\lambda^{*,\zeta}}{m_\lambda} \le \frac{1}{\alpha}$, where $n_\lambda^{*,\zeta} = \min\{n : SL(m_\lambda,m_\lambda) - SL^\zeta(m_\lambda,n) < \epsilon\}$.

For systems with patient customers, if $\alpha > 2$, for any $\epsilon > 0$, $\limsup_{\lambda\to\infty}\frac{n_\lambda^{*,\zeta}}{m_\lambda} \le \frac{1}{\alpha}$, where (abusing notation)

$$n_\lambda^{*,\zeta} = \min\left\{n : \left|W^\zeta(m_\lambda,n) - W(m_\lambda,m_\lambda)\right| \le \epsilon\right\}.$$

Proof of Corollary A.6.3. We first consider the case where customers are impatient.

Recall that in the proof of Theorem 2.3.2.A, we define $\tilde{\pi}_{m_\lambda,m_\lambda}(m_\lambda)$ in (A.10). We can obtain that $\tilde{\pi}_{m_\lambda,m_\lambda}(m_\lambda) < \pi_{m_\lambda,m_\lambda}(m_\lambda)$ and $\lim_{\lambda\to\infty}\tilde{\pi}_{m_\lambda,m_\lambda}(m_\lambda) = \frac{2\zeta-2\alpha}{2\zeta-\alpha}$. We then define $\hat{\pi}^\zeta_{m_\lambda,\lfloor n^{*,\zeta}\rfloor}(\lfloor n^{*,\zeta}\rfloor)$ and $n^{*,\zeta}$ analogous to $\hat{\pi}_{m_\lambda,m_\lambda}(m_\lambda)$ and n^* (see (A.13)). We can obtain that $\hat{\pi}^\zeta_{m_\lambda,\lfloor n^{*,\zeta}\rfloor}(\lfloor n^{*,\zeta}\rfloor) > \pi^\zeta_{m_\lambda,\lfloor n^{*,\zeta}\rfloor}(\lfloor n^{*,\zeta}\rfloor)$ and $\lim_{\lambda\to\infty}\hat{\pi}^\zeta_{m_\lambda,\lfloor n^{*,\zeta}\rfloor}(\lfloor n^{*,\zeta}\rfloor) = 1-\alpha(1-\epsilon)$. Because $\epsilon > 0$ can be arbitrary small, $\liminf_{\lambda\to\infty}\frac{SL^\zeta(m_\lambda,\lfloor n^{*,\zeta}\rfloor)}{SL(m_\lambda,m_\lambda)} \ge \liminf_{\lambda\to\infty}\frac{1-\hat{\pi}^\zeta_{m_\lambda,\lfloor n^{*,\zeta}\rfloor}(\lfloor n^{*,\zeta}\rfloor)}{1-\tilde{\pi}_{m_\lambda,m_\lambda}(m_\lambda)} = 2\zeta-\alpha$.

(ii) The proof is the same as that of Theorem 2.3.2.C with s being replaced by $\frac{s}{\zeta}$.

We then consider the system with patient customers. By Proposition 2.4.3, when $\frac{\alpha}{\zeta} > 2$, we have $\lim_{\lambda\to\infty}\left|W(m_\lambda,m_\lambda) - \left(\frac{q^*}{\lambda} - s\right)\right| \le \epsilon$ for any $\epsilon > 0$; and when $\alpha > 2$, we have $\lim_{\lambda\to\infty}\left|W^\zeta(m_\lambda,m_\lambda) - \left(\frac{q_3^\zeta}{\lambda} - \frac{s}{\zeta}\right)\right| \le \epsilon$ for any $\epsilon > 0$. Recall from the Proof of Lemma A.1.1, $q^*\mu(m_\lambda,q^*) = \lambda$ and $q_3^\zeta\mu(m_\lambda,q_3^\zeta) = \frac{\lambda}{\zeta}$. It follows that

$$\lim_{\lambda\to\infty}\left(\frac{q^*}{\lambda} - s\right) = \lim_{\lambda\to\infty}\left(\frac{1}{\mu(m_\lambda,q^*)} - s\right) = \lim_{\lambda\to\infty}\left(\frac{1}{\zeta\mu(m_\lambda,q_3^\zeta)} - \frac{s}{\zeta}\right) = \lim_{\lambda\to\infty}\left(\frac{q_3^\zeta}{\lambda} - \frac{s}{\zeta}\right) = 0.$$

The desired result then follows by the same argument as that in the proof of Proposition 2.4.4. ∎

A.6.2 The Economics of Tele-Driving: Numerical Experiments

The current state of knowledge regarding the impact of tele-driving technology on costs remains uncertain. There are predictions that the incorporation of automation technology may lead to an average increase of 20% in vehicle prices (Bosch2018), while labor costs may also rise as tele-drivers are likely to require greater specialized and skilled training than conventional drivers. On the other hand, it has also been suggested that tele-driving may lead to reduced labor costs by enabling the outsourcing of remote operation to regions with lower labor costs (Goodall2020).

To gain a deeper understanding of the economic impact of tele-driving, we compare the expected profit of a ride-hailing platform using both conventional and tele-driving systems. Our numerical simulations are based on a fixed fleet size in both systems, and take into account all three regimes (supply-limited, intermediate, and supply-rich). To align with the main findings of our paper, we assume that in supply-limited and intermediate regimes, the platform optimizes the number of tele-drivers to maximize service level. In the supply-rich regime, the platform selects the smallest number of tele-drivers required to maintain a service level that is less than one percent than that in the conventional system. Note that a more complex optimization model is required to fully understand the economic impact of tele-operation and we leave that as an avenue for future research.

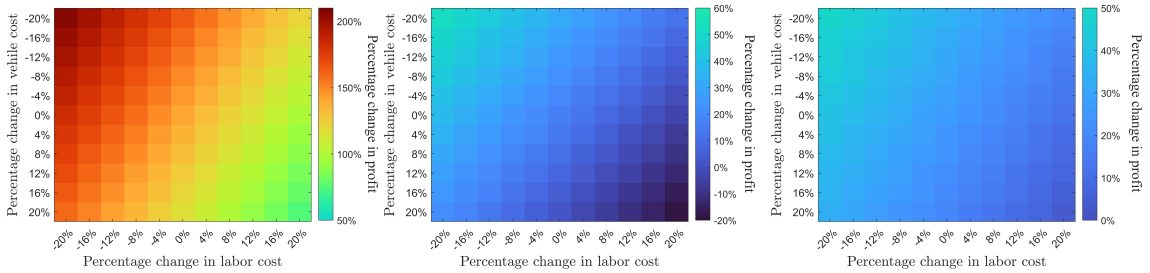
Our numerical simulation assumes an average cost of \$2.50 per mile for ride-hailing services (Terry2019), with in-vehicle drivers earning \$1.875 per mile before accounting for vehicle costs (Uber charges partners 25% fee on all fares)[1]. We set the vehicle cost to be \$0.7 per mile (which we denote by c_v)[2], and use the estimate that the percentage of deadheading miles from ride-hailing is 40.8% (Henao2019). Therefore, the total operational cost of a conventional system can be broken down into a labor cost of \$0.693 ($1.875 - 0.7/(1-0.408)$) per driver per mile in service and a vehicle cost of \$0.7 per vehicle per mile in service and pickup. In contrast, the tele-driving system posits a fixed wage (per unit of time) for tele-drivers. We can compute the average cost of a single driver per unit of time in the conventional system, denoted by c_l. We assume that the wage platform pays tele-drivers is αc_l, where α varies between 0.8 and 1.2. Similarly, we assume that the total cost of

[1] https://www.uber.com/gh/en/drive/basics/tracking-your-earnings

[2] https://newsroom.aaa.com/wp-content/uploads/2022/08/2022-YourDrivingCosts-FactSheet-7-1.pdf

operating a tele-driving vehicle amortized over its expected lifetime is βc_v, where β varies between 0.8 and 1.2.

In the simulation, we calculate the percentage change in expected profit for a ride-hailing platform when switching from a conventional system to a tele-driving system. We use the state-dependent service rate in (2.4) with $\lambda = 10$ and $s = 10$. The results, shown in Figure A.3, indicate that in the supply-limited regime (panel (a)), the platform can enjoy a substantial improvement in profit, even with higher labor and vehicle costs, due to the increase in service level and the reduction in pick-up times (which translates into lower vehicle operation costs). In the supply-rich regime (panel (c)), the platform can also see an improvement in expected profit because of the savings on drivers. However, in the intermediate regime, the impact on profit is more complex and dependent on specific labor and vehicle cost ratios. This is because, in the intermediate regime (panel (b)), the improvement in service level and the savings on drivers are both limited.



(a) Supply-limited regime (b) Intermediate regime (c) Supply-rich regime

Figure A.3: Percentage change in profit for a ride-hailing platform when switching from a conventional system to a tele-driving system.

A.6.3 The Nearest Dispatch Policy

Under the nearest dispatch policy, abusing notation, we let $\rho(q) = \frac{\lambda}{q\mu(m,q)}$ for $q \in \{1, \cdots, n\}$ and $\rho(q) = \frac{\lambda}{n\mu(m,q)}$ for $q > n$, where $\mu(m, q)$ is redefined in Section 2.5 before Lemma 2.5.1. In what follows, we prove the stability condition specified in Lemma 2.5.1.

Proof of Lemma 2.5.1. By following the Proof of Lemma 2.4.1, the system with patient

customers is stable if and only if $\left[1+\sum_{q=1}^{\infty}\prod_{k=1}^{q}\rho(q)\right]<\infty$. Because $\mu(m,q)$ is increasing in q for $q>n$ and $\lim_{q\to\infty}\mu(m,q)=\frac{1}{s}$, there exists $\epsilon>0$ such that $\rho(q)\leq 1-\epsilon$ for q sufficiently large if and only if $\frac{\lambda s}{n}<1$. The desired result then follows immediately. ∎

In Corollary A.6.4, we provide asymptotic analysis for systems under the nearest dispatch policy. We assume that $\mu(m,q)$ is given by (2.4) when $q<n$, and $\mu(m,q)$ is increasing in q for $q\geq n$.

Corollary A.6.4. *Let $m_\lambda=\lfloor\alpha\lambda s\rfloor$ for $\alpha>2$, and define*

$$n_\lambda^*=\min\{n:|W(m_\lambda,n)-W(m_\lambda,m_\lambda)|\leq\epsilon\}$$

for any $\epsilon>0$. Then, $\limsup_{\lambda\to\infty}\frac{n_\lambda^*}{m_\lambda}\leq\frac{1}{\alpha}$.

Proof of Corollary A.6.4. Abusing notation, let $\pi_{m,n}(q)$ denote the stationary probability:

$$\pi_{m,n}(0)=\left[1+\sum_{q=1}^{n}\prod_{k=1}^{q}\rho(k)+\left(\prod_{k=1}^{n}\rho(k)\right)\sum_{q=n+1}^{\infty}\prod_{k=n+1}^{q}\rho(k)\right]^{-1},$$

$$\pi_{m,n}(q)=\pi_{m,n}(0)\prod_{k=1}^{q}\rho(k)\quad\text{for}\quad q\in\{1,\cdots,n\},\quad\text{and}$$

$$\pi_{m,n}(q)=\pi_{m,n}(0)\prod_{k=1}^{n}\rho(k)\prod_{k=n+1}^{q}\rho(k)\quad\text{for}\quad q>n.$$

By observing that $\rho(q)$ is decreasing in q when $q\geq n$, we can establish the same concentration result as displayed in (A.16). The desired result then follows by the same argument as that in the proof of Proposition 2.4.4. ∎

A.7 Numerical Experiments

In this section, we present a detailed description of the data and procedure used to generate the numerical results in Section 2.3.4.

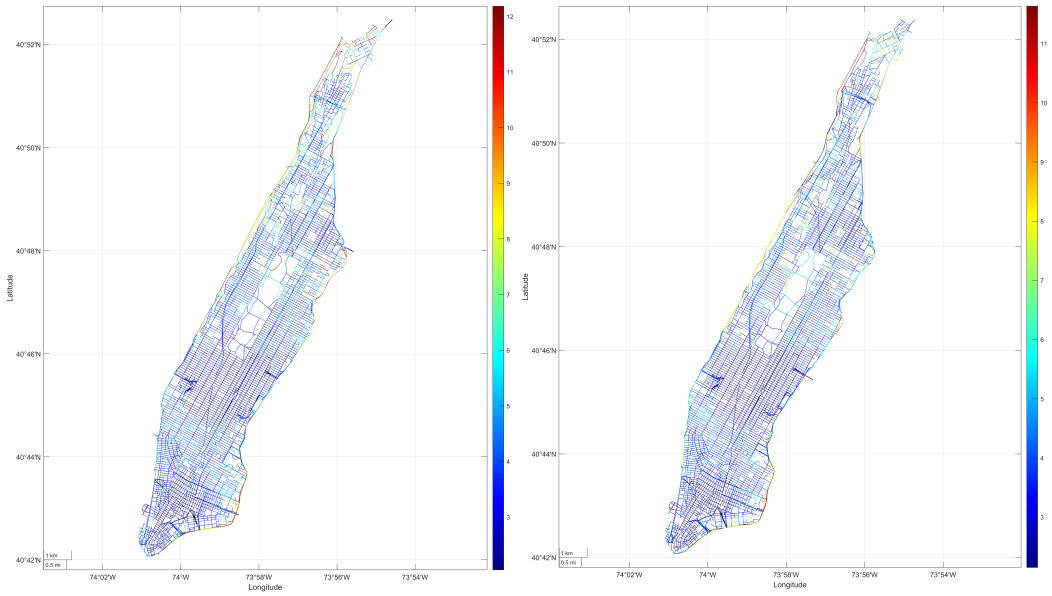
A.7.1 Data Set and Pre-Processing

The data set contains origin-destination data for all 35 million passenger trips by yellow cabs in New York for June, July, and August of 2015. The data set includes the entries: pick-up datetime, drop-off datetime, pick-up longitude, pick-up latitude, drop-off longitude, drop-off latitude. The data set was then filtered by (1) removing trips with pick-up or drop-off locations outside of Manhattan and (2) removing trips with pick-up datetimes on Saturday or Sunday.

We created Manhattan's street network using data from NYC Street Centerline (CSCL) [3], excluding demapped or non-vehicular streets. Following Santi et al. (2014), we extracted the street intersections to construct a network where nodes are the intersections, and directed edges are the streets connecting them. The extracted network contains $6,979$ nodes and $13,786$ directed edges. We mapped the GPS locations from the trip data set to the nearest intersections. During this step, trips with a pick-up or drop-off location more than 100 meters from every street intersection were discarded, resulting in the final data set for the numerical experiments. In the numerical experiments, we assume that passengers are picked up and dropped off at the corresponding street intersections.

The numerical experiments require a travel time estimation for each street (directed edge). With the travel time estimation, we can locate the vehicle that is the closest (in terms of travel time) to a customer who has just arrived. We applied the algorithm proposed in Santi et al. (2014) to obtain the travel time estimation (for each hour of the day) on the directed edges of the street intersection network. The main idea of the algorithm is to minimize (using heuristics) the average relative error between the average trip time and the estimated travel time from the pick-up intersection to the drop-off intersection of each equivalent trip (trips with the same pick-up and drop-off intersections are grouped together). In Figure A.4, we report the estimated travel speed on each street during two different time periods (computed by dividing the estimated travel time of each street by its length).

[3] `https://data.cityofnewyork.us/City-Government/NYC-Street-Centerline-CSCL-/exjm-f27b`



(a) 4:00–5:00 pm.

(b) 5:00–6:00 pm.

Figure A.4: Travel Speed Estimation. The color corresponds to the travel speed (meter/second).

A.7.2 Simulation Procedure

In this section, we describe the simulation procedure (a flowchart can be found in Figure A.5) used to generate the numerical results shown in panel (a) of Figure 2.8. A similar procedure is used to obtain numerical results for different time windows.

Data Preparation

- Extract trips with pick-up datetimes between 4:00 and 6:00 pm on June 16, 2015 (12808 trips between 4:00 and 5:00 pm and 17874 trips between 5:00 and 6:00 pm)
- Create a multiset $\mathcal{I}$ containing all the pick-up intersections of trips extracted in the previous step (notice that the multiset $\mathcal{I}$ can have multiple instances for each intersection since some trips may share the same pick-up intersection)
- Load the estimation of travel times for the time period 4:00–5:00 pm and 5:00–6:00 pm

Implementation

- Set the number of drivers (n) and vehicles (m)
- At 4:00 pm, idle all drivers and place vehicles at randomly sampled (without replacement) street intersections from the multiset $\mathcal{I}$
- Use the trip data (obtained in the data preparation stage) to generate the arrival process of customers (the arrival time and location of a customer is set to be the pick-up time and pick-up intersection respectively)
- If there is no idle driver when a customer arrives, the customer is lost. Otherwise, assign the closest vehicle (determined by the estimated travel time) to pick her up
- Once a vehicle-driver pair is assigned to serve a customer, they will be occupied for a period of time which consists the pick-up time (obtained from the travel time estimation) and the trip time (the difference between pick-up time and drop-off time of the trip)

- After a service is completed, the driver becomes idle and the vehicle stays at the street intersection where the service is terminated until it is assigned to pick up another customer
- Record the number of customers serviced between 5:00 pm and 6:00 pm

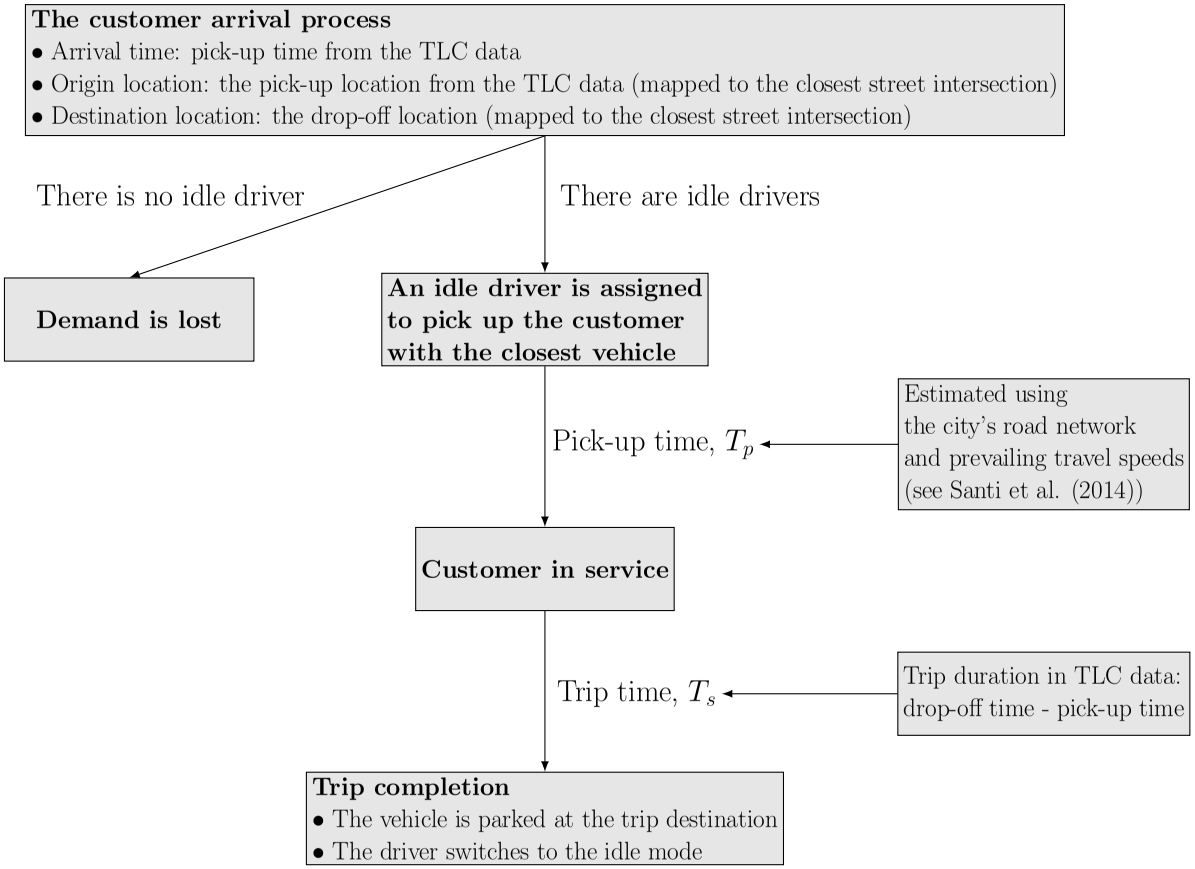
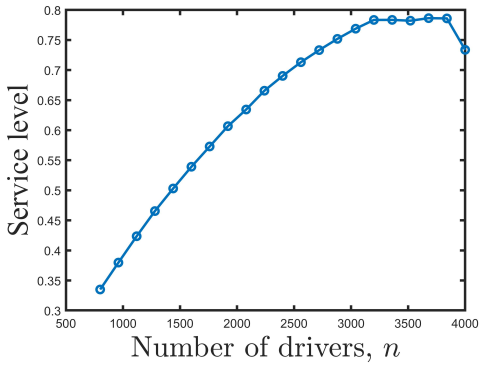


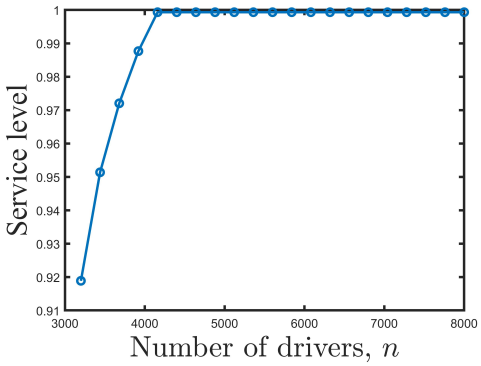
Figure A.5: An illustration of the simulation procedure

A.7.3 Additional Numerical Results Using New York City TLC Data

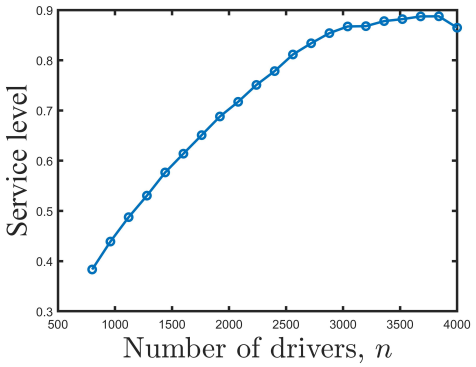
In this section, we provide numerical results using the TLC data for more dates.



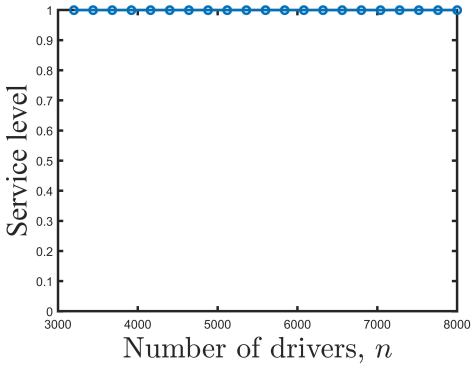
(a) Supply-limited regime (06/22/2015)



(b) Supply-rich regime (06/22/2015)

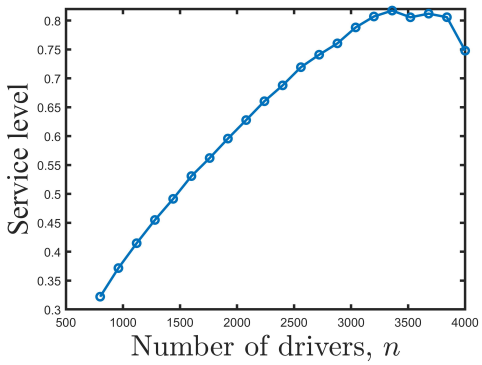


(c) Supply-limited regime (07/06/2015)

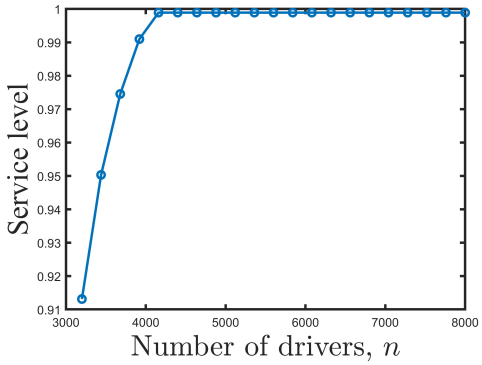


(d) Supply-rich regime (07/06/2015)

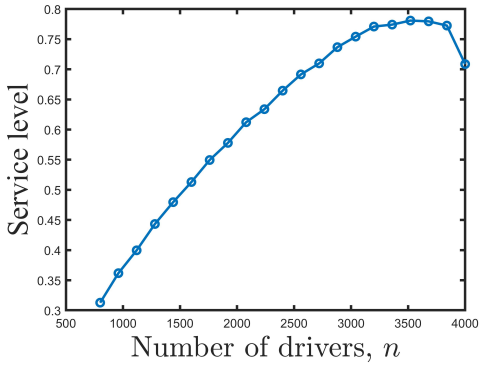
Figure A.6: Additional results from numerical experiments based on TLC data



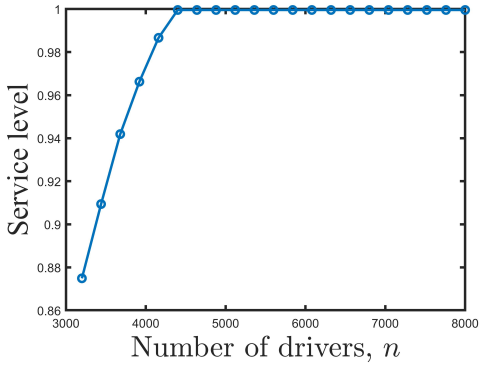
(e) Supply-limited regime (07/10/2015)



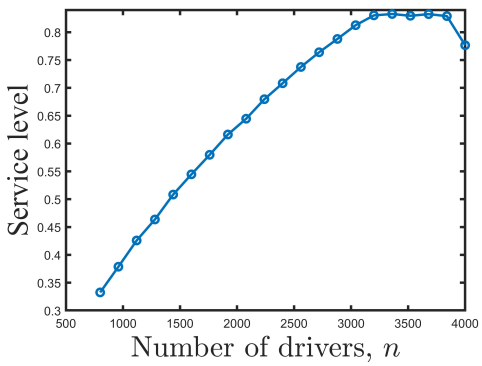
(f) Supply-rich regime (07/10/2015)



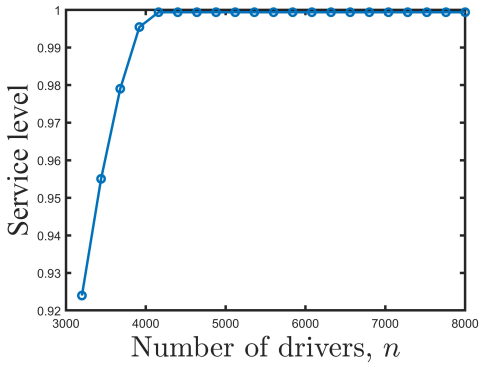
(g) Supply-limited regime (07/29/2015)



(h) Supply-rich regime (07/29/2015)

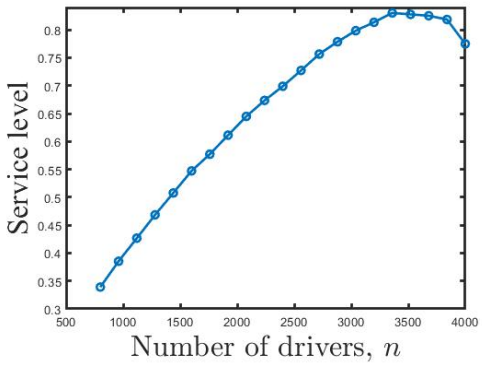


(i) Supply-limited regime (08/04/2015)

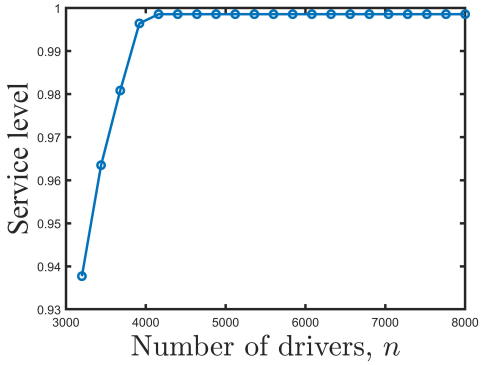


(j) Supply-rich regime (08/04/2015)

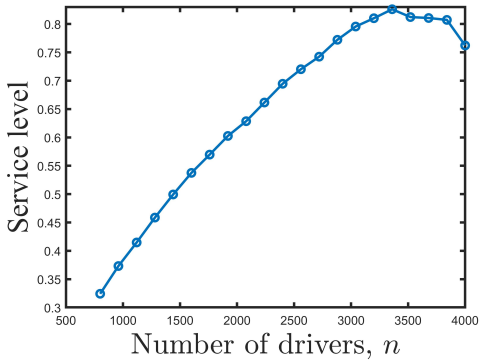
Figure A.6: Additional results from numerical experiments based on TLC data



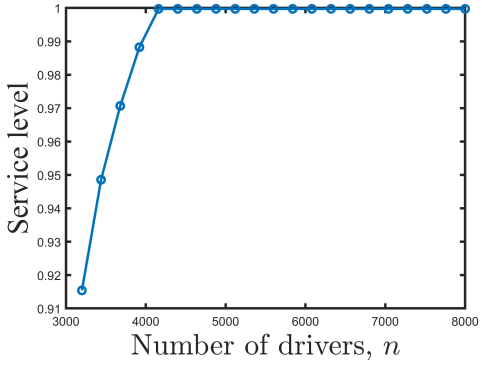
(k) Supply-limited regime (08/18/2015)



(l) Supply-rich regime (08/18/2015)



(m) Supply-limited regime (08/20/2015)



(n) Supply-rich regime (08/20/2015)

Figure A.6: Additional results from numerical experiments based on TLC data

A.8 Support for Assumptions 2.3.1 and 2.3.2

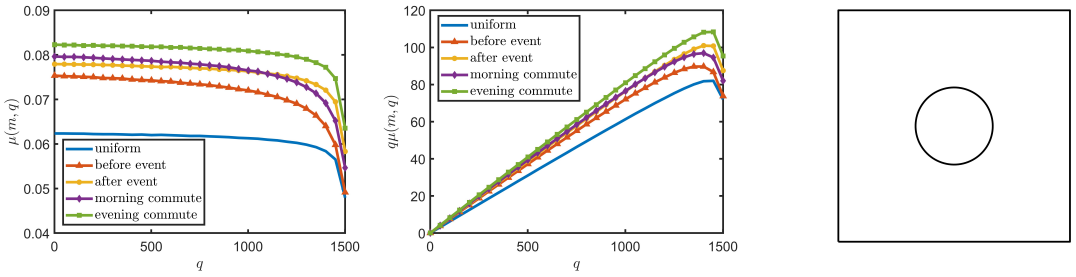
In this section, we provide support for the assumptions we placed on $\mu(m, q)$ (see Assumptions 2.3.1 and 2.3.2). We analyze numerically five different geometries: square, hexagon, disk, grid, and Manhattan road map (as shown in panel (c) in Figure A.7–A.11). In simulations for the first three geometries, namely square, hexagon and disk, we examine five distinct travel patterns, which we categorize as *uniform*, *morning commute*, *evening commute*, *before-event*, and *after-event*. In the uniform pattern, trip origins and destinations are uniformly drawn from the service region. In the morning commute pattern, trip destinations are more likely (with a probability of 80%) to be located in the center inner disk (i.e., downtown area), while the trip origins are more likely (with a probability of 80%) to be situated elsewhere (i.e., suburbs). The reverse holds for the evening commute pattern (i.e., origins are more likely to be located in the inner disk, and destinations are more likely to be situated in the suburbs). In the before-event pattern, trip origins are uniformly drawn from the service region, while the destinations are more likely (with a probability of 80%) to be located in a smaller disk (compared to the disk in the commute pattern). The opposite is true for the after-event pattern (i.e., origins are mostly (with a probability of 80%) generated from a small disk, and the destinations are uniformly distributed across the service region). In addition, we assume that vehicles travel at a constant speed (one unit of distance per unit of time) between any two points using the L^2 norm.

In the simulations for the grid geometry, we examine the same five distinct travel patterns (i.e., uniform, morning commute, evening commute, before-event and after-event), while the origin and destination for each customer can only be drawn from grid lines. In the case of the grid geometry, we assume that vehicles travel at a constant speed (one unit of distance per unit of time) along the grid lines using the L^1 norm.

In the simulations for the Manhattan road map, we examine the travel patterns generated by yellow cab trips in June 2015 (the trip origins, destinations, customer arriving times, and trip times (from origins to destinations) are obtained from the data). The travel speed on each road is estimated following the algorithm described in Santi et al. (2014). We use the travel time along the shortest path to calculate the pickup time between the location of customer and the location of the closest vehicle.

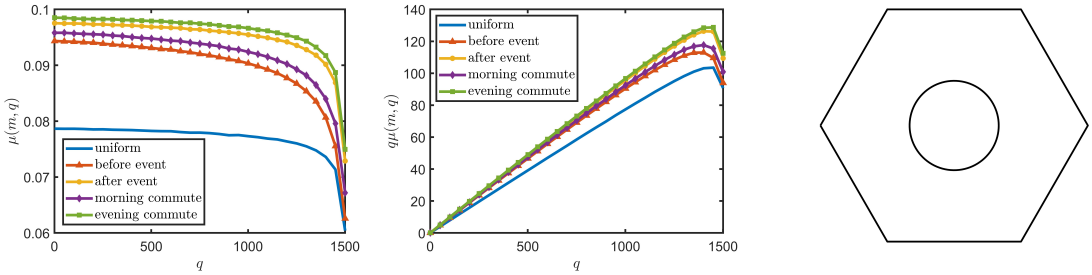
The results obtained from the simulations support the assumptions we made about

$\mu(m,q)$; see A.7–A.11. In particular, we can observe that for the cases considered: (1) $\mu(m,q)$ is strictly concave in q (see panel (a)) and (2) $q\mu(m,q)$ first increases and then decreases in q (panel (b)).



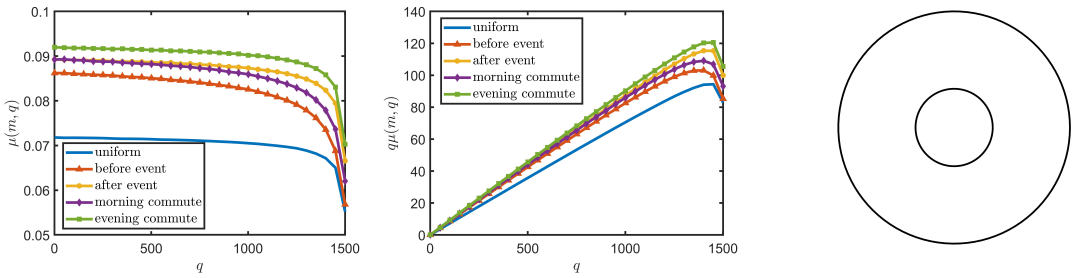
(a) q versus $\mu(m,q)$ (b) q versus $q\mu(m,q)$ (c) Square Geometry

Figure A.7: The case of a square geometry (results are for a square with a side length of 30; the radius of the center inner disk is 5 for morning/evening commute pattern and 3 for before/after-event pattern)



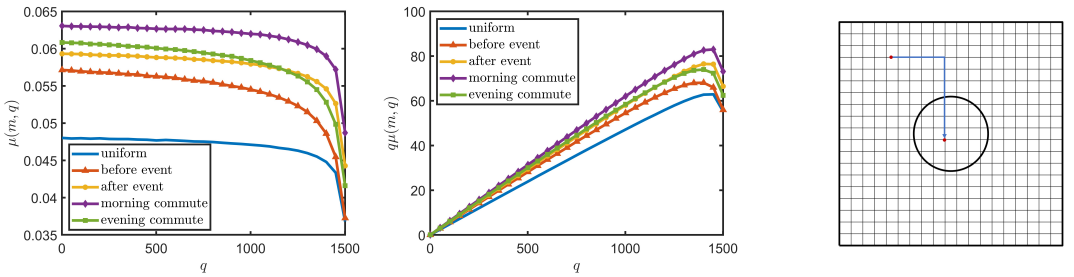
(a) q versus $\mu(m,q)$ (b) q versus $q\mu(m,q)$ (c) Disk Geometry

Figure A.8: The case of a hexagon geometry (the results are for a hexagon with a side length of 15; the radius of the center inner disk is 5 for morning/evening commute pattern and 3 for before/after-event pattern)



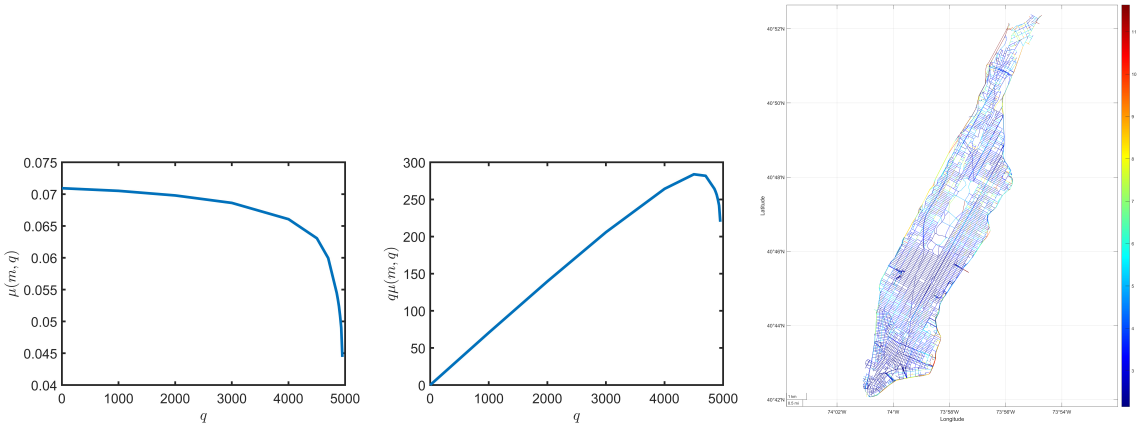
(a) q versus $\mu(m,q)$ (b) q versus $q\mu(m,q)$ (c) Disk Geometry

Figure A.9: The case of a disk geometry (the disk has a radius of 15; the radius of the center inner disk is 5 for morning/evening commute pattern and 3 for before/after-event pattern)



(a) q versus $\mu(m,q)$ (b) q versus $q\mu(m,q)$ (c) Grid Map

Figure A.10: The case of a grid geometry (the results are for a 19×19 grid with side length 30; the radius of the center inner disk is 5 for morning/evening commute pattern and 3 for before/after-event pattern)



(a) q versus $\mu(m,q)$ (b) q versus $q\mu(m,q)$ (c) Manhattan Road Map

Figure A.11: The case of the Manhattan road map with New York City Taxi Data

Appendix B

Appendices for Chapter 3

The Appendix is organized as follows. In Appendix B.1, we provide proofs for systems without AVs (Theorem 3.4.1, Lemma 3.4.1 and Proposition 3.4.1). In Appendix B.2 and B.3, we characterize the optimal strategy and the corresponding outcome under the random/CV-prioritized policy (i.e., Theorem B.2.1 and Proposition B.3.1). In Appendix B.4, we characterize the optimal strategy and the corresponding outcome under the AV-prioritized policy and show that the AV-prioritized policy dominates the other two policies (Theorem 3.5.1). In Appendix B.5, we compare outcomes in systems with and without AVs (Theorem 3.5.2 and Proposition 3.5.1). In Appendix B.6, we provide analysis for the system with location-dependent pricing.

B.1 Proofs for Systems without AVs

We first introduce the concept of a *driver-incentive compatible capacity allocation*, which plays a crucial role in addressing the constraint associated with the CV equilibrium repositioning strategy. A capacity allocation (s, r, q) is called a driver-incentive compatible capacity allocation if no driver has an incentive to change her strategy (which is deduced from (3.4)) under this capacity allocation. In Section B.1.1, we characterize the driver-incentive compatible capacity allocation. In Section B.1.2, we characterize outcomes under the platform's optimal strategy (Theorem 3.4.1). In Section B.1.3, we provide proofs for the centralized system (Lemma 3.4.1 and Proposition 3.4.1).

B.1.1 The Driver-incentive Compatible Capacity Allocation

We first present Lemma B.1.1, in which we characterize the set of optimal repositioning strategies for a single driver given the CV capacity allocation (s^C, r^C, q^C). Using Lemma B.1.1, we obtain the set of driver-incentive compatible capacity allocations in Lemma B.1.2.

Define

$$q_i^* = (1 + \frac{t_{ij}}{t_{ji}})S_{ij}, \quad k_i^* = \frac{S_{ij}}{S_{ji}}, \tag{B.1}$$

$$g_1(s, r, q) = s_{21}q_1 - s_{12}q_2, \quad \text{and} \quad g_2(s, r, q) = s_{12}q_2 - s_{21}q_1. \tag{B.2}$$

Note that the definitions for q_1^* in (B.1) and (3.16) are equivalent. Also note that $g_1(s, r, q) + g_2(s, r, q) = 0$ so that they cannot be both positive.

Lemma B.1.1. *Given the CV capacity allocation (s^C, r^C, q^C), if $g_i(s^C, r^C, q^C) \geq 0$, the set of optimal repositioning strategies $\Omega(s^C, r^C, q^C \mid S_{12}, S_{21})$ for a single driver is:*

(i) $\{(0, 0)\}$ if $q_i^C < q_i^* + k_i^* q_j^C$;

(ii) $\{\eta : \eta_j = 0, \eta_i \in [0, 1]\}$ if $q_i^C = q_i^* + k_i^* q_j^C$; and

(iii) $\{\eta : \eta_j = 0, \eta_i = 1\}$ otherwise.

Proof of Lemma B.1.1. Recall that drivers are paid w per unit time when they are in service. Given the CV capacity allocation (s^C, r^C, q^C), the effective wage (i.e., expected earning per unit time) of a single driver $\hat{w}(\eta_1, \eta_2)$, with respect to her repositioning strategy $\eta = (\eta_1, \eta_2)$, can be obtained via the Renewal Reward Theorem (ross1996stochastic). Without loss of generality, we define the renewal cycle as the time experienced by the driver between completing consecutive services at location 1. Then

$$\hat{w}(\eta_1, \eta_2) = w\frac{T^s(\eta_1, \eta_2)}{T(\eta_1, \eta_2)}, \tag{B.3}$$

where $T(\eta_1, \eta_2)$ is the expected time of a renewal cycle, and $T^s(\eta_1, \eta_2)$ is the expected time the driver is in service within a renewal cycle.

Let x_1 denote the expected time the driver experiences between starting repositioning from location 2 to location 1 and completing a service at location 1. Let x_2 denote the expected time the driver experiences between starting queueing at location 2 and completing

a service at location 1. Recall that we denote by $W_i^C = \frac{q_i^C}{\lambda_{ij}}$ the expected delay experienced by a driver waiting to be matched with customers at location i. Then $T(\eta_1, \eta_2)$, x_1 and x_2 satisfy

$$T(\eta_1,\eta_2) = (1-\eta_1)\big[W_1^C + t_{12} + \eta_2 x_1 + (1-\eta_2)x_2\big] + \eta_1\big[t_{12} + x_2\big],$$
$$x_1 = t_{21} + W_1^C + t_{12} + \eta_2 x_1 + (1-\eta_2)x_2, \quad \text{and} \quad x_2 = W_2^C + t_{21}.$$

Let x_1^s denote the the expected time the driver is in service between starting repositioning from location 2 to location 1 and completing a service at location 1. Let x_2^s denote the expected time the driver is in service between starting queueing at location 2 and completing a service at location 1. Then, $T^s(\eta_1, \eta_2)$, x_1^s and x_2^s satisfy

$$T^s(\eta_1,\eta_2) = (1-\eta_1)[t_{12} + \eta_2 x_1^s + (1-\eta_2)x_2^s)] + \eta_1 x_2^s,$$
$$x_1^s = t_{12} + \eta_2 x_1^s + (1-\eta_2)x_2^s, \quad \text{and} \quad x_2^s = t_{21}.$$

The systems of equations above admit unique solutions for $T(\eta_1, \eta_2)$ and $T^s(\eta_1, \eta_2)$ for all $\lambda_{ij} > 0$. We can obtain that $\frac{\partial \hat{w}(\eta_1,\eta_2)}{\partial \eta_1} = w(\eta_2 - 1)\frac{A_{\eta_1}(\eta_1,\eta_2)}{(B_{\eta_1})^2}$ and $\frac{\partial \hat{w}(\eta_1,\eta_2)}{\partial \eta_2} = w(\eta_1 - 1)\frac{A_{\eta_2}(\eta_1,\eta_2)}{(B_{\eta_2})^2}$, where B_{η_1}, B_{η_2} are some non-zero constants,

$$A_{\eta_1}(\eta_1,\eta_2) = (t_{21}t_{12} + t_{12}^2) + \underbrace{(t_{12}W_2^C - t_{21}W_1^C)}_{b_1} - \eta_2(t_{21}^2 + t_{21}t_{12}), \quad \text{and} \tag{B.4}$$

$$A_{\eta_2}(\eta_1,\eta_2) = (t_{12}t_{21} + t_{21}^2) + \underbrace{(t_{21}W_1^C - t_{12}W_2^C)}_{b_2} - \eta_1(t_{12}^2 + t_{12}t_{21}). \tag{B.5}$$

Because η is defined on a compact set (i.e., $\eta \in [0,1] \times [0,1]$) and $\hat{w}(\eta_1, \eta_2)$ is continuous in η, the maximum of $\hat{w}(\eta_1, \eta_2)$ can be attained by the Extreme Value Theorem (rudin1976principles).

Observe that $A_{\eta_1}(0,0) + A_{\eta_2}(0,0) > 0$. Then there exists $i \in \{1,2\}$ such that $A_{\eta_i}(0,0) > 0$. We first show that for any $\eta^* \in \Omega(s^C, r^C, q^C \mid S_{12}, S_{21})$, which is the set of optimal repositioning strategy for the driver, we must have $\eta_i^* = 0$ if $A_{\eta_i}(0,0) > 0$. Without loss of generality, we assume $A_{\eta_2}(0,0) > 0$ and thus $\frac{\partial \hat{w}(\eta_1,\eta_2)}{\partial \eta_2}_+(0,0) < 0$, where we use a subscript $\cdot_+$ to denote the right-hand derivative as $(0,0)$ is a boundary point. For simplicity, we omit

the subscript $\cdot_+$ in the rest of the proof. We first notice that $\frac{\partial \hat{w}(\eta_1,\eta_2)}{\partial \eta_2}(0,\eta_2) = \frac{\partial \hat{w}(\eta_1,\eta_2)}{\partial \eta_2}(0,0)$ as $\frac{\partial \hat{w}(\eta_1,\eta_2)}{\partial \eta_i}$ is independent of η_i. Therefore, $\hat{w}(0,\eta_2) < \hat{w}(0,0)$ for $\eta_2 \in (0,1]$. Suppose there exists $\tilde{\eta}_1 \in (0,1)$ and $\tilde{\eta}_2 \in (0,1)$ such that $\hat{w}(\tilde{\eta}_1,\tilde{\eta}_2)$ achieves the maximum value. Then we must have $\frac{\partial \hat{w}(\eta_1,\eta_2)}{\partial \eta_1}(\tilde{\eta}_1,\tilde{\eta}_2) = 0$. It follows that $\hat{w}(\tilde{\eta}_1,\tilde{\eta}_2) = \hat{w}(0,\tilde{\eta}_2) < \hat{w}(0,0)$, and thus we reach a contradiction. We then notice that $A_{\eta_1}(1,1) + A_{\eta_2}(1,1) > 0$. Therefore, there exists $i \in \{1,2\}$ such that $\frac{\partial \hat{w}(\eta_1,\eta_2)}{\partial \eta_i}(1,1) < 0$, which implies that $\eta = (1,1)$ is dominated by either $(1,0)$ or $(0,1)$. Moreover, because neither $(0,1)$ nor $(1,1)$ is optimal, $(\eta_1,1)$ cannot be optimal for any $\eta_1 \in (0,1)$ because $\hat{w}(\eta_1,1)$ is monotone in η_1 (recall that $\frac{\partial \hat{w}(\eta_1,\eta_2)}{\partial \eta_i}$ does not depend on η_i). It remains to show that $(1,\eta_2)$ is not optimal for any $\eta_2 \in (0,1)$. This is because $\hat{w}(1,\eta_2)$ is monotone in η_2, $(1,\eta_2)$ is weakly dominated by either $(1,0)$ or $(1,1)$. Therefore, an optimal strategy for the driver $\eta^* = (\eta_1^*,\eta_2^*)$ must satisfy $\eta_2^* = 0$ given $A_{\eta_2}(0,0) > 0$.

Observe that $b_1 + b_2 = 0$, where b_1 and b_2 are defined in (B.4) and (B.5) respectively. Then there exists $i \in (0,1)$ such that $b_i \geq 0$. Note that $b_i \geq 0$ implies that $A_{\eta_i}(0,0) > 0$ and thus $\eta_i^* = 0$ for $i \in \{1,2\}$ according to the previous analysis. Observe that $g_i(s,r,q) \geq 0$ is equivalent to $b_i \leq 0$, where $g_i(s,r,q)$ is defined in (B.2). Without loss of generality, assume that $g_1(s,r,q) \geq 0$ (which implies that $\eta_2^* = 0$). Then it remains to investigate $A_{\eta_1}(0,0)$. Note that $A_{\eta_1}(0,0) \leq 0$ is equivalent to $\frac{\partial \hat{w}(\eta_1,\eta_2)}{\partial \eta_1}(\eta_1,0) \geq 0$ and vise versa. Therefore, we have (i) $\eta_1^* = 0$ if $A_{\eta_1}(0,0) > 0$, (ii) $\eta_1^* \in [0,1]$ if $A_{\eta_1}(0,0) = 0$, and (iii) $\eta_1^* = 1$ otherwise, where $A_{\eta_1}(0,0) = t_{12}t_{21} + t_{12}t_{12} + \frac{t_{12}}{\lambda_{21}}q_2 - \frac{t_{21}}{\lambda_{12}}q_1$. By some algebra, we can obtain Lemma B.1.1. ∎

By Lemma B.1.1, a symmetric strategy which is the best response to (s^C, r^C, q^C) for each driver must satisfy $\eta_1(s^C, r^C, q^C) \geq 0$ and $\eta_2(s^C, r^C, q^C) = 0$ or $\eta_1(s^C, r^C, q^C) = 0$ and $\eta_2(s^C, r^C, q^C) \geq 0$. Because $\Lambda_{12} < \Lambda_{21}$ and by the flow balance constrain (3.6), we must have $\eta_1(s^C, r^C, q^C) \geq 0$ and $\eta_2(s^C, r^C, q^C) = 0$. Then, we can obtain the set of driver-incentive compatible capacity allocations in Lemma B.1.2.

Lemma B.1.2. *In a system without AVs, a CV capacity allocation (s^C, r^C, q^C) is driver-incentive compatible if and only if*

$$r_{21}^C = 0, \text{ and either (i) } q_1^C \leq q_1^*,\ q_2^C = 0,\ r_{12}^C = 0 \text{ or (ii) } q_1^C = q_1^* + k_1^* q_2^C,\ r_{12}^C > 0. \quad \text{(B.6)}$$

B.1.2 Proof of Theorem 3.4.1

By Lemma B.1.2, Problem I can be reformulated as follows:

$$\begin{aligned}\text{(Problem I)} \qquad \max_{w} \quad & \Pi^C = (p-w)(s_{12}^C + s_{21}^C) \\ \text{subject to} \quad & \text{(3.1), (3.2), (3.4), (3.6), (3.7), (3.9), (3.10), (B.6),} \\ & M = 0 \text{ and } F_k = 0 \text{ for } k = 1, 2.\end{aligned}$$

We then solve Problem I via the following 3 steps.

Step (1). We show that any strategy (w) that results in a CV capacity allocation (s^C, r^C, q^C) such that (i) $q_2^C > 0$ or (ii) $0 < q_1^C < q_1^*$ is sub-optimal. By (3.10), we can rewrite the platform's profit as

$$\Pi^C = (p-w)(s_{12}^C + s_{21}^C) = p(s_{12}^C + s_{21}^C) - N\hat{w} = p(s_{12}^C + s_{21}^C) - \frac{N^2\overline{w}}{L}. \tag{B.7}$$

For case (i), we have $q_1^C = q_1^* + k_1^* q_2^C$ by (B.6), $s_{12}^C = S_{12}$ and $s_{21}^C = S_{21}$ by (3.7), which implies that $r_{12}^C = (\Lambda_{21} - \Lambda_{12})t_{12}$ by the flow balance constrain (3.6). Consider another CV capacity allocation $(\tilde{s}^C, \tilde{r}^C, \tilde{q}^C)$ with $\tilde{s}_{ij}^C = s_{ij}^C$, $\tilde{r}_{ij}^C = r_{ij}^C$, $\tilde{q}_1^C = q_1^* < q_1^C$, $\tilde{q}_2^C = 0 < q_2^C$, and the corresponding wage $\tilde{w}$ determined by (3.10). We note that $(\tilde{s}^C, \tilde{r}^C, \tilde{q}^C)$ satisfies constrains (3.1), (3.2), (3.4), (3.6), (3.7), (3.9), (3.10) and (B.6), while the capacity of drivers recruited is smaller. By (B.7), the platform gains a higher profit. For case (ii), we have $r_{12}^C = 0$ by Lemma B.1.2, $s_{12}^C = S_{12}$ by (3.7), $s_{21}^C = \frac{t_{21}}{t_{12}} S_{12}$ by (3.6), and $q_2^C = 0$ by (3.7). Consider another CV capacity allocation $(\tilde{s}^C, \tilde{r}^C, \tilde{q}^C)$ with $\tilde{s}_{ij}^C = s_{ij}^C$, $\tilde{r}_{ij}^C = r_{ij}^C$ and $\tilde{q}_1^C = 0 < q_1^C$, $\tilde{q}_2^C = q_2^C$, and the corresponding wage $\tilde{w}$ determined by (3.9) and (3.10). Then by a similar argument as that for case (i), the platform gains a higher profit.

Step (2). By Lemma B.1.2 and the analysis in step (1), it suffices to consider (s^C, r^C, q^C) with $q_2^C = 0$, and either (i) $r_{12}^C = 0$, $q_1^C = 0$, or (ii) $r_{12}^C > 0$, $q_1^C = q_1^*$. In what follows, we characterize and compare the platform's profits under these two cases.

In case (i), drivers do not queue or reposition and thus their utilization $\rho = 1$. In this case, the platform recruits up to C_1 amount of drivers, where C_1 is the amount of type-1 demand defined in (3.12). Moreover, driver's effective wage $\hat{w} = w$ and thus $N = \frac{Lw}{p}$ by

(3.10). Let $\Pi_1(N)$ denote the platform's profit in this case, we have

$$\Pi_1(N) = N[p-w] = N\left(p - \frac{Np}{L}\right), \quad \text{for} \quad N \in [0, C_1]. \tag{B.8}$$

In case (ii), we first note that $N \in [C_1 + q_1^*, C_1 + C_2 + q_1^*]$, where C_2 is the amount of type-2 demand define in (3.13), by the following observations. By Lemma B.1.2, CVs start to reposition when all type-1 demand is fulfilled and $q_1^C = q_1^*$. Therefore, $N \geq C_1 + q_1^*$. Recall from step (1) that $q_2^C = 0$. It follows that $N \leq C_1 + C_2 + q_1^*$. We then show that drivers' utilization $\rho = \gamma$, where γ is defined in (3.14). Let α denote the fraction of type-2 demand fulfilled. Then the drivers' utilization $\rho = \frac{C_1 + \alpha C_2 \gamma}{C_1 + q_1^* + \alpha C_2} = \gamma$, where the last equality follows from the fact that

$$\gamma = \frac{C_1}{C_1 + q_1^*}. \tag{B.9}$$

It follows that drivers' effective wage $\hat{w} = \gamma w$ and $N = \frac{L\gamma w}{p}$ by (3.10). Let $\Pi_2(N)$ denote the platform's profit in this case. We have

$$\Pi_2(N) = \gamma N(p-w) = \gamma N\left(p - \frac{Np}{\gamma L}\right), \quad \text{for} \quad N \in [C_1 + q_1^*, C_1 + C_2 + q_1^*]. \tag{B.10}$$

Observe that both $\Pi_1(N)$ and $\Pi_2(N)$ are concave and $\Pi_1'(N) = p - \frac{2Np}{L} > \Pi_2'(N) = \gamma p - \frac{2Np}{L}$. Let $\Pi_1^* = \max\limits_{N \in [0, C_1]} \Pi_1(N)$ and $\Pi_2^* = \max\limits_{N \in [C_1 + q_1^*, C_1 + C_2 + q_1^*]} \Pi_2(N)$. To compare Π_1^* and Π_2^*, we consider the following possibilities.

Case (C.i) $\Pi_2'(C_1 + q_1^*) \leq 0$, which is equivalent to $L \leq \frac{2(C_1 + q_1^*)}{\gamma}$. We have $\Pi_1^* - \Pi_2^* = \Pi_1^* - \Pi_2(C_1 + q_1^*) \geq \Pi_1(C_1) - \Pi_2(C_1 + q_1^*) = C_1 p(1 - \frac{C_1}{L}) - \gamma(C_1 + q_1^*)p\left(1 - \frac{C_1 + q_1^*}{\gamma L}\right) = C_1 p(1 - \frac{C_1}{L}) - C_1 p\left(1 - \frac{C_1 + q_1^*}{\gamma L}\right) > 0$, where the last equality is due to (B.9).

Case (C.ii) $\Pi_2'(C_1 + q_1^*) > 0$ and $\Pi_2'(C_1 + C_2 + q_1^*) \leq 0$, which is equivalent to $\frac{2(C_1 + q_1^*)}{\gamma} < L \leq \frac{2(C_1 + C_2 + q_1^*)}{\gamma}$. We have $\Pi_1^* - \Pi_2^* = \Pi_1(C_1) - \Pi_2(\frac{L\gamma}{2}) = C_1 p - \frac{C_1^2 p}{L} - \frac{Lp\gamma^2}{4} \geq 0$ if and only if $L \in \left(\frac{2C_1\left(1 - \sqrt{1-\gamma^2}\right)}{\gamma^2}, \frac{2C_1\left(1 + \sqrt{1-\gamma^2}\right)}{\gamma^2}\right]$.

Case (C.iii) $\Pi_2'(C_1 + C_2 + q_1^*) > 0$, which is equivalent to $L > \frac{2(C_1 + C_2 + q_1^*)}{\gamma}$. We have $\Pi_1^* - \Pi_2^* = \Pi_1(C_1) - \Pi_2(C_1 + C_2 + q_1^*) = \frac{p(C_2 + q_1^*)(C_2 + 2C_1 + q_1^*)}{L} - \gamma p C_2 \geq 0$ if and only if $L \leq \frac{(C_2 + q_1^*)(C_2 + 2C_1 + q_1^*)}{\gamma C_2}$.

Step (3). We characterize the platform's optimal strategy and the corresponding outcomes. Denote by w^C and N^C the optimal wage and the corresponding amount of drivers recruited respectively. By the analysis in Step (2), $w^C = \frac{N^C p}{L}$ if $N^C \in [0, C_1]$ and $w^C = \frac{N^C p}{\gamma L}$ if $N^C \in [C_1 + q_1^*, C_1 + C_2 + q_1^*]$. Therefore, we focus on the characterization of N^C and consider the following two scenarios.

Scenario (C.a): $\frac{2C_1\left(1+\sqrt{1-\gamma^2}\right)}{\gamma^2} < \frac{2(C_1+C_2+q_1^*)}{\gamma}$. By case (C.i), $N^C = \underset{N\in[0,C_1]}{\arg\max}\, \Pi_1(N) = \min(\frac{L}{2}, C_1)$ for $L \le \frac{2(C_1+q_1^*)}{\gamma}$. By case (C.ii), $N^C = \underset{N\in[0,C_1]}{\arg\max}\, \Pi_1(N) = C_1$ for $\frac{2(C_1+q_1^*)}{\gamma} < L < \frac{2C_1\left(1+\sqrt{1-\gamma^2}\right)}{\gamma^2}$. Because $\Pi_1^* - \Pi_2^*$ decreases in L and $\Pi_1^* - \Pi_2^* = 0$ when $L = \frac{2C_1\left(1+\sqrt{1-\gamma^2}\right)}{\gamma^2}$, we have $N^C = \underset{N\in[C_1+q_1^*, C_1+C_2+q_1^*]}{\arg\max}\, \Pi_2(N) = \min(\frac{\gamma L}{2}, C_1 + C_2 + q_1^*)$ for $L > \frac{2C_1\left(1+\sqrt{1-\gamma^2}\right)}{\gamma^2}$ by case (C.ii) and (C.iii).

Scenario (C.b): $\frac{2C_1\left(1+\sqrt{1-\gamma^2}\right)}{\gamma^2} \ge \frac{2(C_1+C_2+q_1^*)}{\gamma}$. By case (C.i), $N^C = \underset{N\in[0,C_1]}{\arg\max}\, \Pi_1(N) = \min(\frac{L}{2}, C_1)$ for $L \le \frac{2(C_1+q_1^*)}{\gamma}$. By case (C.ii), $N^C = \underset{N\in[0,C_1]}{\arg\max}\, \Pi_1(N) = C_1$ for $\frac{2(C_1+q_1^*)}{\gamma} < L \le \frac{2(C_1+C_2+q_1^*)}{\gamma}$. Because $\Pi_1^* - \Pi_2^* \ge 0$ when $N = \frac{2(C_1+C_2+q_1^*)}{\gamma}$, we have $N^C = \underset{N\in[0,C_1]}{\arg\max}\, \Pi_1(N) = C_1$ for $\frac{2(C_1+C_2+q_1^*)}{\gamma} < L < \frac{(C_2+q_1^*)(C_2+2C_1+q_1^*)}{\gamma C_2}$, and $N^C = \underset{N\in[C_1+q_1^*, C_1+C_2+q_1^*]}{\arg\max}\, \Pi_2(N) = C_1 + C_2 + q_1^*$ for $L > \frac{(C_2+q_1^*)(C_2+2C_1+q_1^*)}{\gamma C_2}$ by case (C.iii).

Therefore, there exists a threshold L^C defined in (3.15) such that under the optimal strategy, $q_1^C = 0$ and $r_{12}^C = 0$ if $L \le L^C$, and $q_1 = q_1^*$ and $r_{12}^C > 0$ otherwise. Moreover, the amount of drivers recruited is given by

$$N^C = \begin{cases} \min(\frac{L}{2}, C_1), & \text{if } L \le L^C, \\ \min(\frac{\gamma L}{2}, C_1 + C_2 + q_1^*), & \text{otherwise.} \end{cases} \tag{B.11}$$

Lastly, from the expression of N^C, we can observe that N^C weakly increases in L.

B.1.3 Proofs for Centralized Systems

In this section, we provide proofs for centralized systems where the platform has control over the repositioning of CVs.

Proof of Lemma 3.4.1. First, we claim that in a centralized system, the CV capacity allocation under the platform's optimal strategy must satisfy (i) $q_1^C = q_2^C = 0$, and (ii) $(1 - a_1)r_{12}^C = 0$. Condition (ii) implies that the platform repositions CVs only when all the type-1 demand is fulfilled. In what follows, we show condition (i) and (ii) separately. For condition (i), because the platform can directly control the CV capacity allocation in a centralized system, by (B.7), a CV capacity allocation (s^C, r^C, q^C) with $q_1^C > 0$ or $q_2^C > 0$ is dominated by $(\tilde{s}^C, \tilde{r}^C, \tilde{q}^C)$ with $\tilde{s}_{ij}^C = s_{ij}^C$, $\tilde{r}_{ij}^C = r_{ij}^C$ and $\tilde{q}_i^C = 0$. For condition (ii), by (3.10), whenever $s_{12}^C < S_{12}$ and $r_{12}^C > 0$, it is possible to decease r_{12}^C and increase $s_{12}^C + s_{21}^C$ by the same amount such that (3.1), (3.2), (3.4), (3.6), (3.7), (3.9), (3.10) hold and the amount of drivers recruited remains the same with a lower wage w. Moreover, the platform fulfills more demand (as $s_{12}^C + s_{21}^C$ increases) and thus gains more profit.

It remains to compute the platform's profit with $q_1^C = 0$, $q_2^C = 0$ and either (i) $r_{12}^C = 0$, $s_{12}^C \leq S_{12}$; or (ii) $r_{12}^C > 0$, $s_{12}^C = S_{12}$. In case (i), the platform's profit is given by $\Pi_1(N)$, where $\Pi_1(N)$ is defined in (B.8). In case (ii), we first note that $N \in (C_1, C_1 + C_2]$ under the platform's optimal strategy. Given $N \in (C_1, C_1 + C_2]$, C_1 amount of CVs fulfill all type-1 demand and the remaining $(N - C_2)$ amount of CVs fulfill a fraction of type-2 demand. We can obtain that the drivers' utilization $\rho = \frac{C_1 + \gamma(N - C_1)}{N}$. By (3.10), $w = \frac{N^2 p}{L[C_1 + \gamma(N - C_1)]}$. Let $\hat{\Pi}_2(N)$ denote the platform's profit in case (ii). We have

$$\begin{aligned}\hat{\Pi}_2(N) &= [C_1 + \gamma(N - C_1)](p - w) \\ &= [C_1 + \gamma(N - C_1)] \left(p - \frac{N^2 p}{L[C_1 + \gamma(N - C_1)]} \right), \quad \text{for} \quad N \in (C_1, C_1 + C_2].\end{aligned}$$

Observe that both $\Pi_1(N)$ and $\hat{\Pi}_2(N)$ are concave and $\Pi_1'(N) = p - \frac{2Np}{L} > \hat{\Pi}_2'(N) = \gamma p - \frac{2Np}{L}$. Let N^* denote the amount of drivers recruited in the centralized system. Then with a similar analysis in step (2) and step (3) for the proof of Theorem 3.4.1, $r_{12}^C = 0$ if $L \leq \frac{2C_1}{\gamma}$ and $r_{12}^C \in (0, (\Lambda_{21} - \Lambda_{12})t_{12}]$ otherwise. Moreover,

$$N^* = \begin{cases} \min(\frac{L}{2}, C_1), & \text{if } L \leq \frac{2C_1}{\gamma}, \\ \min(\frac{\gamma L}{2}, C_1 + C_2), & \text{otherwise.} \end{cases}$$

Lastly, from the expression of N^*, we can observe that N^* weakly increases in L. ∎

Proof of Proposition 3.4.1. The result follows by a direct comparison between the outcomes under the optimal strategies of Problem I and II.

■

B.2 Systems under The Random Assignment Policy

In this section, we characterize the platform's optimal strategy and the corresponding outcomes when the platform adopts the random assignment policy. Recall that under the random assignment policy, the platform randomly assigns a vehicle to a customer. By Little's law, the expected delay experienced by AVs and CVs queueing at location i is given by $W_i^A = W_i^C = \frac{q_i}{\Lambda_{ij}}$. Therefore, the amount of AVs and CVs in service from each location must be proportional to the amount of AVs and CVs queueing at that location. That is,

$$\begin{cases} s_{ij}^C = 0 & \text{if } q_i^C = 0 \text{ and } q_i^A > 0, \\ s_{ij}^A = 0 & \text{if } q_i^A = 0 \text{ and } q_i^C > 0, \\ \frac{s_{ij}^A}{s_{ij}^C} = \frac{q_i^A}{q_i^C} & \text{if } q_i^A > 0 \text{ and } q_i^C > 0. \end{cases} \tag{B.12}$$

Given that the platform dispatches vehicles based on the random assignment policy, the platform solves the following problem:

$$\begin{aligned} \text{(Problem R)} \qquad \max_{M,w,\eta^A} \quad & \Pi = p(s_{12}^A + s_{21}^A) + (p-w)(s_{12}^C + s_{21}^C) - M \cdot I, \\ \text{subject to} \quad & \text{(3.1)–(3.10), (B.12) and} \\ & \eta^C \text{ is a CV equilibrium repositioning strategy.} \end{aligned}$$

In Theorem B.2.1, we characterize the optimal strategy under the random assignment policy. Recall that we define L^C in (3.15) and q_1^* in (3.16). Let

$$L^R = \begin{cases} L^C & \text{if } I \geq \gamma p \\ \frac{2pC_1\left(1+\sqrt{1-\gamma^2}\right)}{\gamma I}, & \text{if } I < \gamma p \text{ and } C_2 \geq \frac{\sqrt{1-\gamma^2}}{\gamma} C_1, \\ \frac{p(C_2+q_1^*)(2C_1+C_2+q_1^*)}{IC_2}, & \text{otherwise.} \end{cases} \tag{B.13}$$

Theorem B.2.1. *When the platform adopts the random assignment policy, there exists*

an optimal strategy for the platform under which

(i) $N = \min(\frac{IL}{2p}, C_1)$, $r_{12}^C = r_{21}^C = 0$ *and* $q_1^C = q_2^C = 0$ *if* $L \le L^R$*; and*

(ii) $N = \begin{cases} \min(\frac{\gamma L}{2}, C_1 + C_2 + q_1^*) & \text{if } I \ge \gamma p, \\ \min(\frac{IL}{2p}, C_1 + C_2 + q_1^*) & \text{otherwise,} \end{cases}$ $F_1 = 0$, $r_{12}^C > 0$, $r_{21}^C = 0$, $q_1^C = q_1^*$ *and* $q_2^C = 0$ *otherwise. Moreover, the number of drivers recruited weakly increases in the labor pool size* L*.*

Theorem B.2.1 shows that, when the platform adopts the random assignment policy, there exists a threshold L^R such that when the driver pool size is smaller than L^R, it is optimal for the platform to recruit a limited amount of drivers and do not let them reposition or queue. When the driver pool size is larger than L^R, the platform recruits a large amount of CVs to fulfill all type-1 demand, and reposition with a positive probability to fulfill some type-2 demand. In the latter case, there are q_1^* amount of CVs queuing at the low-demand location.

In what follows, we present the proof for Theorem B.2.1. We first characterize the driver-incentive compatible capacity allocation in Section B.2.1, and then characterize the optimal strategy and the corresponding outcomes in Section B.2.2.

B.2.1 The Driver-incentive Compatible Capacity Allocation

We first obtain the set of optimal strategies for a single driver, given the system capacity allocation (s, r, q) in Lemma B.2.1. We then characterize the driver-incentive compatible capacity allocation in Lemma B.2.2. Recall that we define q_i^* and k_i^* in (B.1), and function $g_i(s, r, q)$ in (B.2).

Lemma B.2.1. *Given the capacity allocation of system* (s, r, q)*, if* $g_i(s, r, q) \ge 0$*, the set of optimal repositioning strategies* $\Omega(s, r, q \mid S_{12}, S_{21})$ *for a single driver is:*

(i) $\{(0, 0)\}$ *if* $q_i < q_i^* + k_i^* q_j$*;*

(ii) $\{\eta : \eta_j = 0, \eta_i \in [0, 1]\}$ *if* $q_i = q_i^* + k_i^* q_j$*; and*

(iii) $\{\eta : \eta_j = 0, \eta_i = 1\}$ *otherwise.*

Proof of Lemma B.2.1. The proof of Lemma B.2.1 is similar to that of Lemma B.1.1 and thus we omit the details here. ∎

By Lemma B.2.1, a symmetric strategy which is the best response to (s, r, q) for each driver satisfies $\eta_i^C(s, r, q) \geq 0$ and $\eta_j^C(s, r, q) = 0$, given that $g_i(s, r, q) \geq 0$. Then we can obtain the set of driver-incentive compatible capacity allocations under the random assignment policy in Lemma B.2.2.

Lemma B.2.2. *Under the random assignment priority policy, if*

$$g_i(s, r, q) \geq 0, \tag{B.14}$$

a capacity allocation is driver-incentive compatible if and only if

$$r_{ji}^C = 0, \text{ and } \begin{cases} \text{(i) } q_i \leq q_i^* + k_i^* q_j, r_{ij}^C = 0, \\ \text{(ii) } q_i = q_i^* + k_i^* q_j, r_{ij}^C \geq 0, \quad \text{or} \\ \text{(iii) } q_i > q_i^* + k_i^* q_j, r_{ij}^C = \frac{t_{ij}}{t_{ji}} s_{ji}^C. \end{cases} \tag{B.15}$$

B.2.2 Proof of Theorem B.2.1

By Lemma B.2.2, Problem R can be reformulated as follows:

$$\begin{aligned} \text{(Problem R)} \qquad & \max_{M, w, (s^A, r^A, q^A)} \quad \Pi = p(s_{12}^A + s_{21}^A) + (p - w)(s_{12}^C + s_{21}^C) - M \cdot I, \\ & \text{subject to} \quad \text{(3.1)–(3.10), (B.12), and (B.14)–(B.15).} \end{aligned}$$

We then solve Problem R via the following 4 steps.

Step (1). We show that under the random assignment policy, any strategy that leads to one of the following three cases is sub-optimal: (i) $r_{21} > 0$, (ii) $q_2 > 0$ and (iii) $q_1 \notin \{0, q_1^*\}$. Note that the platform's profit can be rewritten as

$$\Pi = p(s_{12}^A + s_{21}^A + s_{12}^C + s_{21}^C) - \frac{N^2 \overline{w}}{L} - M \cdot I. \tag{B.16}$$

For case (i), suppose $r_{21} > 0$. By (3.5) – (3.6), $\frac{s_{12} + r_{12}}{t_{12}} = \frac{s_{21} + r_{21}}{t_{21}}$, which implies that $r_{12} > 0$. We further consider 3 subcases: case (i.i) $r_{12}^C = 0$ and $r_{21}^C = 0$, case (i.ii) $r_{12}^C > 0$ and $r_{21}^C = 0$ and case (i.iii) $r_{12}^C = 0$ and $r_{21}^C > 0$. For case (i.i) consider another capacity allocation $(\tilde{s}, \tilde{r}, \tilde{q})$ with $\tilde{r}_{21}^A = 0$ and $\tilde{r}_{12}^A = r_{12}^A - \frac{t_{12}}{t_{21}} r_{21}^A$, and other capacity parameters are identical to those in (s, r, q). Let the corresponding wage $\tilde{w}$ and AV fleet size $\tilde{M}$

determined by (3.10) and (3.8) respectively. Note that $(\tilde{s}, \tilde{r}, \tilde{q})$ satisfies constrains (3.1)–(3.10), (B.12), and (B.14)–(B.15), and fulfills the same amount of demand. Moreover, the platform recruits less drivers and purchases less AVs. By (B.16), the platform gains a higher profit. Similarly, for case (i.ii), another capacity allocation $(\tilde{s}, \tilde{r}, \tilde{q})$ with $\tilde{s}_{ij}^A = s_{ij}^A$, $\tilde{s}_{ij}^C = s_{ij}^C$, $\tilde{r}_{12}^A + \tilde{r}_{12}^C = s_{12}^A + s_{12}^C - \frac{t_{12}}{t_{21}} r_{21}^A$, $\tilde{q}_1^A + \tilde{q}_1^C = q_1^*$ and $\tilde{q}_2^A = \tilde{q}_2^C = 0$ dominates (s, r, q). For case (i.iii), another capacity allocation $(\tilde{s}, \tilde{r}, \tilde{q})$ with $\tilde{s}_{ij}^A = s_{ij}^A$, $\tilde{s}_{ij}^C = s_{ij}^C$, $\tilde{r}_{21}^A = \tilde{r}_{21}^C = 0$, $\tilde{r}_{12}^C = 0$ and $\tilde{r}_{12}^A = r_{12}^A - \frac{t_{12}}{t_{21}}(r_{21}^A + r_{21}^C)$ and $\tilde{q}_i^A = \tilde{q}_i^C = 0$ dominates (s, r, q).

For case (ii), suppose $q_2 > 0$ and we consider 2 subcases: case (ii.i) $r_{12}^C = 0$ and case (ii.ii) $r_{12}^C > 0$. In case (ii.i), we must have $r_{12}^A = (\Lambda_{21} - \Lambda_{12}) t_{12}$ by (3.7) and (3.5)–(3.6), and $q_1 \leq q_1^* + k_1^* q_2$ by Lemma B.2.2. Then another capacity allocation $(\tilde{s}, \tilde{r}, \tilde{q})$ with $\tilde{s}_{ij}^A = s_{ij}^A$, $\tilde{s}_{ij}^C = s_{ij}^C$, $\tilde{r}_{12}^A = r_{12}^A$, $\tilde{r}_{12}^C = r_{12}$ and $\tilde{q}_1^C = \tilde{q}_i^C = 0$ dominates (s, r, q). For case (ii.ii), we must have $q_1 \geq q_1^* + k_1^* q_2$ by Lemma B.2.2. Then another capacity allocation $(\tilde{s}, \tilde{r}, \tilde{q})$ with $\tilde{s}_{ij}^A = s_{ij}^A$, $\tilde{s}_{ij}^C = s_{ij}^C$, $\tilde{q}_1^A + \tilde{q}_1^C = q_1^*$, $\tilde{q}_1^A s_{12}^C = \tilde{q}_1^C s_{12}^A$, $\tilde{r}_{12}^C = r_{12}^C$, $\tilde{r}_{12}^A = r_{12}^A$, and $\tilde{q}_2^C = 0$ dominates (s, r, q).

For case (iii), we consider 2 subcases: case (iii.i) $q_1 < q_1^*$ and case (iii.ii) $q_1 > q_1^*$. For case (iii.i), $r_{12}^C = 0$ by Lemma B.2.2. Then another capacity allocation $(\tilde{s}, \tilde{r}, \tilde{q})$ with $\tilde{q}_1^A = \tilde{q}_1^C = 0$ and other parameters identical to that of (s, r, q) dominates (s, r, q). For case (iii.ii), $r_{12}^C = \frac{t_{12}}{t_{21}} s_{21}^A$ by Lemma B.2.2 and thus $q_1^C = 0$. Then another capacity allocation $(\tilde{s}, \tilde{r}, \tilde{q})$ with $\tilde{q}_1^A = q_1^*$ and other parameters identical to that of (s, r, q) dominates (s, r, q).

Step (2). We show that if a capacity allocation (s, r, q) with $r_{12}^C > 0$ is optimal, then we can increase r_{12}^A and decrease r_{12}^C to achieve the same platform profit until $r_{12}^A = s_{21}^A \frac{t_{12}}{t_{21}}$ (which implies $\eta_{12}^A = 1$) or $r_{12}^C = 0$ (during this process, M and N remain unchanged and all the constrains are satisfied). We first note that given $r_{12}^C > 0$, it suffices to consider the case where $M + N \in [C_1 + q_1^*, C_1 + C_2 + q_1^*]$ (otherwise, the capacity allocation (s, r, q) violates Lemma B.2.2, or it belongs to one of the three cases discussed in Step (1)). By Lemma B.2.2, $q_1 = q_1^*$. Let α denote the fraction of type-2 demand fulfilled. The platform's profit is given by $\Pi = p(C_1 + \gamma\alpha C_2) - \frac{\overline{w} N^2}{L} - M \cdot I$, which does not depend on how type-1 and type-2 demand are divided among AVs and CVs. Therefore, we can increase (decrease) the repositioning capacity of AVs (CVs) and use AVs (CVs) to serve more type-2 (type-1) demand without affecting the platform profit. As a result, we can focus on capacity

allocations with either $r_{12}^C = 0$ or $r_{12}^C > 0$ and $r_{12}^A = s_{21}^A \frac{t_{12}}{t_{21}}$ ($\eta_{12}^A = 1$)[1]. Besides the simplification on proofs, We exclude other possible optimal capacity allocations because when the travel costs of AVs and CVs are considered (fuel and driving effort) and the travel cost of AVs is smaller than that of CVs, any capacity allocation with $r_{12}^C > 0$ and $r_{12}^A < s_{21}^A \frac{t_{12}}{t_{21}}$ ($\eta_{12}^A < 1$) cannot be optimal.

Step (3). By Lemma B.2.2 and the analysis in steps (1) and (2), it suffice to consider capacity allocations with $q_2 = 0$, $r_{21} = 0$, and either $r_{12}^C = 0$ and $q_1 = 0$, or $r_{12}^C > 0$, $q_1 = q_1^*$ and $r_{12}^A = \frac{t_{12}}{t_{21}} s_{21}^A$. Notice that the profit gained by using AVs to fulfill one unit of type-1 demand is $p - I$, and that to fulfill one unit of type-2 demand is $p - \frac{I}{\gamma}$. Hence, in a system with AVs, all type-1 demand must be served as $p - I > 0$, and all type-2 demand must be served if $I < \gamma p$. To summarize, if $I > \gamma p$, it suffices to consider either (i) $r_{12}^C = 0$, $r_{12}^A = 0$ and $q_1 = 0$, or (ii) $r_{12}^C > 0$, $q_1^C = q_1^*$ and $M = 0$. Otherwise, it suffices to consider (iii) $r_{12}^C = 0$, $r_{12}^A > 0$ and $q_1 = 0$, and (iv) $r_{12}^C > 0$, $q_1^C = q_1^*$, $M > 0$ and $\eta_1^A = 1$ (i.e., $r_{12}^A = \frac{t_{12}}{t_{21}} s_{21}^A > 0$). Moreover, in case(iii) and case (iv), $s_{12}^A + s_{12}^C + s_{21}^A + s_{21}^C = C_1 + C_2$. In what follows, we characterize the compare the platform's profit under these four cases.

In case (i), both AVs and CVs do not queue and reposition. The platform recruits up to C_1 amount of CVs and $C_1 - N$ amount of AVs to fulfill all type-1 demand. Let $\Pi_3(N)$ denote the platform's profit in this case, and recall that we define $\Pi_1(N)$ in (B.8). We have

$$\Pi_3(N) = \Pi_1(N) + (C_1 - N)(p - I) = N\left(I - \frac{Np}{L}\right) + C_1(p - I), \quad \text{for} \quad N \in [0, C_1]. \tag{B.17}$$

In case (ii), the platform operates with only CVs, CVs reposition with a positive probability and $q_1^C = q_1^*$. Therefore, the platform's profit is given by $\Pi_2(N)$, where $\Pi_2(N)$ is defined in (B.10).

In case (iii), the platform recruits up to C_1 amount of CVs to fulfill only the type-1 demand. Moreover, the platform recruits $C_1 + C_2 - N$ amount of AVs to fulfill all the type-1 demand that can not be fulfilled by CVs and all the type-2 demand. Let $\Pi_4(N)$ denote the platform's profit in this case, and recall that we define $\Pi_3(N)$ in (B.17). We

[1]We note that as the amount of drivers recruited and the platform profit remain the same, all the comparison results with respect to driver welfare and the platform's profit in this paper do not rely on this simplification.

have

$$\Pi_4(N) = \Pi_3(N) + C_2(\gamma p - I) = N\left(I - \frac{Np}{L}\right) + C_1(p - I) + C_2(\gamma p - I), \quad \text{for} \quad N \in [0, C_1]. \tag{B.18}$$

In case (iv), the platform recruits $N \in [C_1 + q_1^*, C_1 + C_2 + q_1^*]$ amount of CVs such that CVs fulfill all type-1 demand and some type-2 demand, and $q_1^C = q_1^*$. Moreover, the platform recruits $C_1 + C_2 + q_1^* - N$ amount of AVs and let AVs reposition with probability 1 to fulfill all the type-2 demand that can not be fulfilled by CVs. Let $\Pi_5(N)$ denote the platform's profit in this case, and recall that we define $\Pi_2(N)$ in (B.10). We have

$$\begin{aligned}\Pi_5(N) &= \Pi_2(N) + (C_1 + C_2 + q_1^* - N)(\gamma p - I) \\ &= N\left(I - \frac{Np}{L}\right) + (C_1 + C_2 + q_1^*)(\gamma p - I), \quad \text{for} \quad N \in [C_1 + q_1^*, C_1 + C_2 + q_1^*].\end{aligned}$$

Observe that $\Pi_3(N)$, $\Pi_4(N)$ and $\Pi_5(N)$ are concave, and $\Pi_3'(N) = \Pi_4'(N) = \Pi_5'(N) = I - \frac{2Np}{L}$. Let $\Pi_3^* = \max\limits_{N\in[0,C_1]} \Pi_3(N)$, $\Pi_4^* = \max\limits_{N\in[0,C_1]} \Pi_4(N)$ and $\Pi_5^* = \max\limits_{N\in[C_1+q_1^*, C_1+C_2+q_1^*]} \Pi_5(N)$. We then consider the case where $I > \gamma p$ and $I \le \gamma p$ separately.

Case (R.i) $I > \gamma p$. In this case, the platform does not let AVs reposition. It suffices to compare $\Pi_3(N)$ and $\Pi_2(N)$. In particular, we consider the following subcases.

Case (R.i.i) $\Pi_2'(C_1 + q_1^*) < 0$, which implies that $L < \frac{2(C_1+q_1^*)}{\gamma}$. In this case, we have $\Pi_3^* - \Pi_2^* = \Pi_3^* - \Pi_2(C_1 + q_1^*) \ge \Pi_3(C_1) - \Pi_2(C_1 + q_1^*) = \frac{C_1(C_1+q_1^*)p}{\gamma L} - \frac{C_1^2 p}{L} \ge 0$.

Case (R.i.ii) $\Pi_2'(C_1+q_1^*) \ge 0$ and $\Pi_2'(C_1+C_2+q_1^*) < 0$, which is equivalent to $\frac{2(C_1+q_1^*)}{\gamma} < L < \frac{2(C_1+C_2+q_1^*)}{\gamma}$. Because the condition implies that $\Pi_3'(C_1) > 0$, we have $\Pi_3^* - \Pi_2^* = \Pi_3(C_1) - \Pi_2(\frac{\gamma L}{2}) = \Pi_1(C_1) - \Pi_2(\frac{\gamma L}{2}) \ge 0$ if and only if $L \in \left[\frac{2C_1\left(1-\sqrt{1-\gamma^2}\right)}{\gamma^2}, \frac{2C_1\left(1+\sqrt{1-\gamma^2}\right)}{\gamma^2}\right]$.

(R.i.iii) $\Pi_3'(C_1 + C_2 + q_1^*) \ge 0$, which implies that $L \ge \frac{2(C_1+C_2+q_1^*)}{\gamma}$. We have $\Pi_3^* - \Pi_2^* = \Pi_3(C_1) - \Pi_2(C_1 + C_2 + q_1^*) = \Pi_1(C_1) - \Pi_2(C_1 + C_2 + q_1^*) \ge 0$ if and only if $L \le \frac{(C_2+q_1^*)(C_2+2C_1+q_1^*)}{\gamma C_2}$.

Case (R.ii) $I \le \gamma p$. In this case, the platform has an incentive to reposition AVs, and uses AVs to fulfill all the demand that can not be fulfilled by CVs. Therefore, it suffice to compare Π_4^* and Π_5^*. In particular, we consider the following subcases.

Case (R.ii.i) $\Pi_5'(C_1 + q_1^*) < 0$, which is equivalent to $L < \frac{2(C_1+q_1^*)p}{I}$. We have $\Pi_4^* - \Pi_5^* =$

$\Pi_4^* - \Pi_5(C_1 + q_1^*) \geq \Pi_4(C_1) - \Pi_5(C_1 + q_1^*) = C_1(p - \frac{C_1 p}{L}) + \frac{(C_1+q_1^*)^2 p}{L} - \gamma p(C_1 + q_1^*) = \frac{(C_1+q_1^*)^2 p}{L} - \frac{C_1^2 p}{L} \geq 0$, where the last equality follows from (B.9).

Case (R.ii.ii) $\Pi_5'(C_1 + q_1^*) > 0$ and $\Pi_5'(C_1 + C_2 + q_1^*) \leq 0$, which implies that $\frac{2p(C_1+q_1^*)}{I} < L < \frac{2p(C_1+C_2+q_1^*)}{I}$. We have $\Pi_4^* - \Pi_5^* = \Pi_4(C_1) - \Pi_5(\frac{IL}{2p}) = -\frac{I^2 L}{4p} + \frac{C_1 I}{\gamma} - \frac{C_1^2 p}{L} \geq 0$ if and only if $L \in \left[\frac{2pC_1\left(\frac{1}{\gamma} - \sqrt{\frac{1}{\gamma^2}-1}\right)}{I}, \frac{2pC_1\left(\frac{1}{\gamma} + \sqrt{\frac{1}{\gamma^2}-1}\right)}{I}\right]$.

Case (R.ii.iii) $\Pi_5'(C_1 + C_2 + q_1^*) \geq 0$, which implies $L \geq \frac{2p(C_1+C_2+q_1^*)}{I}$. We have $\Pi_4^* - \Pi_5^* = \frac{p(C_2+q_1^*)(2C_1+C_2+q_1^*)}{L} - C_2 I \geq 0$ if and only if $L \leq \frac{p(C_2+q_1^*)(2C_1+C_2+q_1^*)}{C_2 I}$.

Step (4). We characterize the platform's optimal strategy and the corresponding outcomes. By a similar analysis to that in scenarios (C.a) and (C.b) in the step (3) for the proof of Theorem 3.4.1, we can obtain that if $I \geq \gamma p$, there exists a thresholds L^R defined in (3.15) such that under the platform's optimal strategy, $q_1 = 0$, $r_{12}^C = 0$ if $L \leq \hat{L}$, and $r_{12}^C > 0$, $q_1^C = q_1^*$ otherwise. Moreover, the amount of workers recruited is given by

$$N^R = \min(\frac{IL}{2p}, C_1) \text{ if } L \leq L^R, \text{ and } N^R = \begin{cases} \min(\frac{\gamma L}{2}, C_1 + C_2 + q_1^*) & \text{if } I \geq \gamma p, \\ \min(\frac{IL}{2p}, C_1 + C_2 + q_1^*) & \text{otherwise,} \end{cases} \text{ otherwise.} \tag{B.19}$$

Lastly, from the expression of N^R, we can observe that N^R weakly increases in L.

B.3 Systems under The CV-Prioritized Policy

In this section, we characterize the platform's optimal strategy and the corresponding outcomes when the platform adopts the CV-prioritized policy. Under the CV-prioritized policy, if there are both AVs and CVs in the queue at a location, the platform randomly selects a CV to serve an arriving customer. From the driver's perspective, the system is identical to the one without AVs. Given a capacity allocation (s, r, q), if $q_i^C > 0$, the delay experienced by AVs and CVs at location i is given by $W_i^A = +\infty$ and $W_i^C = \frac{q_i^C}{\lambda_{ij}}$ respectively. Otherwise, $W_i^A = \frac{q_i^A}{\lambda_{ij}^A}$ and $W_i^C = 0$. Therefore, the demand at location i is assigned to AVs only if there are no CVs queueing at location i. That is

$$s_{ij}^A q_i^C = 0. \tag{B.20}$$

Given that the platform dispatches vehicles based on the CV-prioritized policy, the platform solves the following problem:

$$\begin{aligned} \text{(Problem C)} \qquad \max_{M,w,\eta^A} \quad & \Pi = p(s_{12}^A + s_{21}^A) + (p-w)(s_{12}^C + s_{21}^C) - M \cdot I, \\ \text{subject to} \quad & \text{(3.1)–(3.10), (B.20) and} \\ & \eta^C \text{ is a CV equilibrium repositioning strategy.} \end{aligned}$$

In Proposition B.3.1, we characterize the optimal strategy under the CV-prioritized policy.

Proposition B.3.1. *Under the CV-prioritized policy, the optimal strategy for the platform and the corresponding outcomes are identical to those described in Theorem B.2.1.*

Proposition B.3.1 indicates, perhaps surprisingly, that the optimal strategy and corresponding outcomes under the CV-prioritized policy are identical to those under the random assignment policy obtained in Theorem B.2.1. This is because, in accordance with Theorem B.2.1, whenever CVs queue at location 1, the platform repositions AVs (if any) with probability 1 to meet otherwise unmet demand by CVs at the high-demand location. This makes the two policies equivalent in terms of outcomes.

Proof of Proposition B.3.1. Because the system is identical to that without AVs from the driver's perspective, by Lemma B.1.2, a capacity allocation (s, r, q) under the random assignment priority policy is driver-incentive compatible if and only if (B.6) holds. Therefore, Problem C can be reformulated as follows:

$$\begin{aligned} \text{(Problem C)} \qquad \max_{M,w,(s^A,r^A,q^A)} \quad & \Pi = p(s_{12}^A + s_{21}^A) + (p-w)(s_{12}^C + s_{21}^C) - M \cdot I, \\ \text{subject to} \quad & \text{(3.1)–(3.10), (B.20), and (B.6).} \end{aligned}$$

To solve Problem C, we first note that the platform has no incentive to let AVs to queue up anywhere. Otherwise, the platform gains a higher profit with another capacity allocation such that the capacity of AVs queued in both locations are removed and the other capacity parameters remains the same. Additionally, pursuant to Lemma B.1.2,

drivers reposition only when $q_1^C \geq q_1^*$. Therefore, the optimal outcome must fall in one of the following cases:

case (i) $r_{12}^C = 0$ and $q_1^C = 0$, or

case (ii) $r_{12}^C > 0$, $q_1^C = q_1^*$ and $\eta_1^A = 1$ (if $M > 0$).

As a result, we can follow the same argument to that used in the proof of Theorem B.2.1 to obtain the same optimal strategy and outcomes under the CV-prioritized policy. ■

B.4 Proofs for Systems under The AV-Prioritized Policy

In this section, we solve the platform's problem under the AV-prioritized policy (i.e., Problem A). We characterize the set of driver-incentive compatible capacity allocation in Section B.4.1, and the outcome under the platform's optimal strategy in Section B.4.2 (i.e., proof of Theorem 3.5.1).

B.4.1 The Driver-incentive Compatible Capacity Allocation

Under the AV-prioritized policy, from the driver's perspective, the system is identical to the one in which the demand fulfilled by AVs is removed. By Lemma B.1.2, we can obtain the set of driver-incentive compatible capacity allocations under the AV-prioritized policy per Lemma B.4.1. Recall that we define q_i^* and k_i^* in (B.1), function $g_i(s, r, q)$ in (B.2), and F_i as the fraction of effective demand fulfilled by AVs.

Lemma B.4.1. *Under the AV-prioritized policy, if*

$$g_i(s^C, r^C, q^C) \geq 0, \tag{B.21}$$

a capacity allocation is driver-incentive compatible if and only if

$$r_{ji}^C = 0, \text{ and } \begin{cases} \text{(i) } q_i^C < (1-F_i)q_i^* + \frac{1-F_i}{1-F_j}k_i^* q_j^C, r_{ij}^C = 0, \\ \text{(ii) } q_i^C = (1-F_i)q_i^* + \frac{1-F_i}{1-F_j}k_i^* q_j^C, r_{ij}^C \geq 0, \quad \text{or} \\ \text{(iii) } q_i^C > (1-F_i)q_i^* + \frac{1-F_i}{1-F_j}k_i^* q_j^C, r_{ij}^C = \frac{t_{ij}}{t_{ji}} s_{ji}^C. \end{cases} \tag{B.22}$$

B.4.2 Proof of Theorem 3.5.1

By Lemma B.4.1, Problem A can be reformulated as follows:

$$\text{(Problem A)} \quad \max_{M,w,(s^A,r^A,q^A)} \quad \Pi = p(s_{12}^A + s_{21}^A) + (p-w)(s_{12}^C + s_{21}^C) - M \cdot I,$$
$$\text{subject to} \quad (3.1)\text{–}(3.10),\ (3.17) \text{ and } (B.21)\text{–}(B.22). \tag{B.23}$$

We then solve Problem A via the following 3 steps.

Step (1). We show that under the AV-prioritized policy, any strategy that results in a capacity allocation (s, r, q) satisfying one of the following cases is sub-optimal: (i) $r_{21} > 0$ (ii) $q_2 > 0$ and (iii) $(1 - F_1)((\Lambda_{21} - \Lambda_{12})t_{12} - r_{12}^C) \neq 0$ if $r_{12}^C > 0$. Recall that the platform's profit can be rewritten as (B.16).

The analysis for case (i) and (ii) is similar to that under the random assignment policy. For case (iii), consider a capacity allocation (s, r, q) with $r_{12}^C > 0$ and $(1 - F_1)((\Lambda_{21} - \Lambda_{12})t_{12} - r_{12}^C) > 0$, which implies that $F_1 < 1$, $r_{12}^C > (\Lambda_{21} - \Lambda_{12})t_{12}$ and $q_1^C = (1 - F_1)$ by Lemma B.4.1. We further consider 2 subcases: (iii.i) $I \leq \gamma p$ and (iii.ii) $I > \gamma p$. For case (iii.i), is suffices to consider the case where (a) all the demand in the system are fulfilled (otherwise the platform is better off using AVs to fulfill otherwise unfulfilled demand), and (b) $r_{12}^A > 0$ (otherwise, $r_{12}^C = (\Lambda_{21} - \Lambda_{12})t_{12}$ as all the demand are fulfilled). Consider another capacity allocation $(\tilde{s}, \tilde{r}, \tilde{q})$ with $\tilde{r}_{12}^A = r_{12}^A - \delta$, $\tilde{s}_{12}^A = s_{12}^A + \delta$, $\tilde{s}_{21}^A = s_{21}^A$, $\tilde{F}_1 = \frac{\tilde{s}_{12}^A}{S_{12}}$, $s_{12}^C = s_{12}^C - \delta$, $\tilde{r}_{12}^C = r_{12}^C + \delta$, $\tilde{s}_{21}^C = s_{21}^C$, and $\tilde{q}_1^C = (1 - \tilde{F}_1)q_1^*$. Let the corresponding wage $\tilde{w}$ and the AV fleet size $\tilde{M}$ determined by (3.10) and (3.8) respectively, and let $\tilde{N}$ denote the corresponding CVs fleet size. Observe that $(\tilde{s}, \tilde{r}, \tilde{q})$ satisfies constrains (3.1)–(3.10), (3.17) and (B.21)–(B.22). Because $\tilde{F}_1 < F_1$, the platform gains a higher profit by (B.16). Therefore, the platform gains a higher profit by keeping increasing δ until either CVs reposition with probability 1 (i.e., $\tilde{F}_1 = 1$), or all the demand at location 2 are fulfilled (i.e., $\tilde{r}_{12}^C = (\Lambda_{21} - \Lambda_{12})t_{12}$). For case (iii.i), it suffice to consider the case where $r_{12}^A = 0$ (otherwise, the platform is better off removing the amount of AVs serving type-2 demand). Consider another capacity allocation $(\tilde{s}, \tilde{r}, \tilde{q})$ with $\tilde{s}_{12}^A = s_{12}^A + \delta$, $\tilde{s}_{21}^A = s_{21}^A + \frac{t_{21}}{t_{12}}\delta$, $\tilde{r}_{12}^A = 0$, $\tilde{q}_1^A = 0$, $\tilde{s}_{12}^C = s_{12}^C - \delta$, $\tilde{s}_{21}^C = s_{21}^C + \frac{t_{21}}{t_{12}}\delta$, $\tilde{r}_{12}^C = r_{12}^C + 2\delta$, $\tilde{q}_1^C = q_1^C - \frac{\delta}{S_{12}}q_1^* = q_1^C - (1 + \frac{t_{21}}{t_{12}})\delta$. Let the corresponding wage $\tilde{w}$ and the AV fleet size $\tilde{M}$ determined by (3.10) and (3.8) respectively, and let $\tilde{N}$ denote the corresponding CVs fleet size. Observe that $(\tilde{s}, \tilde{r}, \tilde{q})$ satisfies constrains

(3.1)–(3.10), (3.17) and (B.21)–(B.22). Because $\tilde{N} = N$ and $\tilde{M} = M + (1 + \frac{t_{21}}{t_{12}})\delta$, the platform gains a higher profit by (B.16) as $p > I$. Therefore, the platform gains a higher profit by keeping increasing δ until either CVs reposition with probability 1 (i.e., $\tilde{F}_1 = 1$), or all the demand at location 2 are fulfilled (i.e., $\tilde{r}^C_{12} = (\Lambda_{21} - \Lambda_{12})t_{12}$).

Step (2). By Lemma B.4.1 and the analysis in step (1), it suffices to consider capacity allocations with $q_2 = 0$, $r_{21} = 0$. Specifically, if $I > \gamma p$, it suffices to consider: (i) $r^C_{12} = 0$, $q_1 = 0$ and $r^A_{12} = 0$, (ii) $r^C_{12} > 0$, $r^A_{12} = 0$, $F_1 = 1$ (which implies that $\eta^C_1 = 1$) and $q_1 = 0$, and (iii) $r^C_{12} = (\Lambda_{21} - \Lambda_{12})t_{12}$, $r^A_{12} = 0$, $q^C_1 = (1 - F_1)q^*_1$. Otherwise, it suffice to consider (iv) $r^C_{12} = 0$, $r^A_{12} = (\Lambda_{21} - \Lambda_{12})t_{12}$ and $q_1 = 0$, (v) $r^C_{12} > 0$, $r^A_{12} = (\Lambda_{21} - \Lambda_{12})t_{12} - r^C_{12}$, $F_1 = 1$ (which implies that $\eta^C_1 = 1$) and $q_1 = 0$; and case (iii). In what follows, we characterize the compare the platform's profit under these cases.

In case (i), both AVs CVs do not queue and reposition. Therefore, the platform's profit is given by $\Pi_3(N)$, where $\Pi_3(N)$ is defined in (B.17).

In case (ii), the platform recruits C_1 amount of AVs to fulfill all the type-1 demand, and recruits up to C_2 amount of CVs and CVs reposition with probability 1 to fulfill a fraction of type-2 demand. In this case, driver's effective wage is given by $\hat{w} = \gamma w$, and thus $w = \frac{Np}{\gamma L}$ by (3.10). Therefore, profit earned by AVs is $(p - I)C_1$ and that earned by CVs is $\gamma N(p - w)$. Let $\Pi_6(N)$ denote the platform's profit in this case, we have

$$\Pi_6(N) = C_1(p - I) + \gamma N \left(p - \frac{Np}{\gamma L}\right), \quad \text{for} \quad N \in [0, C_2].$$

In case (iii), the platform recruits up to C_1 amount of AVs to fulfill a fraction F_1 of type-1 demand, and recruits at least C_2 and up to $C_1 + C_2 + q^*_1$ amount of CVs to fulfill all the type-2 demand and the type-1 demand that are not fulfilled by AVs. By Lemma B.4.1, $q^C_1 = (1 - F_1)q^*_1$, and thus driver's utilization $\rho = \frac{s^C_{12} + s^C_{21}}{N} = \frac{(1-F_1)C_1 + \gamma C_2}{(1-F_1)(C_1 + q^*_1) + C_2} = \gamma$ by (B.9). Therefore, $w = \frac{Np}{\gamma L}$ by (3.10). The profit earned by CVs is $\gamma N(p - w)$ and that earned by AVs is $(C_1 - \gamma(N - C_2))(p - I)$. Let $\Pi_8(N)$ denote the platform's profit in this case, we have

$$\Pi_8(N) = \gamma N \left(p - \frac{Np}{\gamma L}\right) + [C_1 - \gamma(N - C_2)](p - I), \quad \text{for} \quad N \in [C_2, C_1 + C_2 + q^*_1].$$

In case (iv), the platform uses AVs to fulfill all the type-2 demand and a fraction of

type-1 demand, and recruits up to C_1 amount of CVs to fulfill the type-1 demand that are not fulfilled by AVs. Therefore, the platform's profit is given by $\Pi_4(N)$, where $\Pi_4(N)$ is defined in (B.18).

In case (v), the platform uses AVs to fulfill all type-1 demand and a fraction of type-2 demand, and recruits up to C_2 amount of CVs to fulfill the type-2 demand that are not fulfilled by AVs. Let $\Pi_7(N)$ denote the platform's profit in this case, and recall that we define $\Pi_6(N)$ in (B.4.2). We have

$$\begin{aligned}\Pi_7(N) &= \Pi_6(N) + (C_2 - N)(\gamma p - I) \\ &= C_1(p - I) + (C_2 - N)(\gamma p - I) + \gamma N\left(p - \frac{Np}{\gamma L}\right), \quad \text{for} \quad N \in [0, C_2].\end{aligned}$$

Observe that $\Pi_6(N)$, $\Pi_7(N)$ and $\Pi_8(N)$ are concave, and $\Pi_6'(N) = \gamma p - \frac{2Np}{L}$ and $\Pi_7'(N) = \Pi_8'(N) = I - \frac{2Np}{L}$. Let $\Pi_6^* = \max_{N\in[0,C_2]} \Pi_6(N)$, $\Pi_7^* = \max_{N\in[0,C_2]} \Pi_7(N)$, and $\Pi_8^* = \max_{N\in[C_2,C_1+C_2+q_1^*]} \Pi_8(N)$. We consider the following possibilities.

Case (A.i) $I > \gamma p$ and $\Pi_8'(C_2) \le 0$, which implies that $L \le \frac{2C_2 p}{\gamma I}$. In this case, AVs do not reposition, and thus we compare Π_3^*, Π_6^* and Π_8^*. Because $\Pi_6^* \ge \Pi_6(C_2) = \Pi_8(C_2) = \Pi_8^*$, it suffices to compare Π_6^* with Π_3^*. We consider the following subcases.

Case (A.i.i) $\Pi_3'(C_1) < 0$ and $\Pi_6'(C_2) \le 0$, which implies that $L \le \min\{\frac{2C_1 p}{I}, \frac{2C_2}{\gamma}\}$. We have $\Pi_3^* - \Pi_6^* = \Pi_3\left(\frac{IL}{2p}\right) - \Pi_6\left(\frac{\gamma L}{2}\right) = \frac{I^2 L}{4p} - \frac{\gamma^2 pL}{4} > 0$ as $I > \gamma p$.

Case (A.i.ii) $\Pi_3'(C_1) \le 0$ and $\Pi_6'(C_2) \ge 0$, which implies that $L \in [\frac{2C_2}{\gamma}, \frac{2C_1 p}{I}]$. We have $\Pi_3^* - \Pi_6^* = \Pi_3\left(\frac{IL}{2p}\right) - \Pi_6(C_2) \ge \Pi_3\left(\frac{IL}{2p}\right) - \Pi_6\left(\frac{\gamma L}{2}\right) > 0$ by case (A.i.i).

Case (A.i.iii) $\Pi_3'(C_1) \ge 0$ and $\Pi_6'(C_2) \le 0$, which implies that $L \in [\frac{2C_1 p}{I}, \frac{2C_2}{\gamma}]$. We have $\Pi_3^* - \Pi_6^* = \Pi_3(C_1) - \Pi_6\left(\frac{\gamma L}{2}\right) = C_1\left(I - \frac{C_1 p}{L}\right) - \frac{\gamma^2 pL}{4} \ge 0$ if and only if

$$L \in \left[\frac{2C_1\left(I - \sqrt{I^2 - \gamma^2 p^2}\right)}{\gamma^2 p}, \frac{2C_1\left(I + \sqrt{I^2 - \gamma^2 p^2}\right)}{\gamma^2 p}\right].$$

Case (A.i.iv) $\Pi_3'(C_1) \ge 0$ and $\Pi_6'(C_2) \ge 0$, which implies that $L \ge \max\left(\frac{2C_1 p}{I}, \frac{2C_2}{\gamma}\right)$. We have $\Pi_3^* - \Pi_6^* = \Pi_3(C_1) - \Pi_6(C_2) = C_1 I - \frac{C_1^2 p}{L} - \gamma C_2 p + \frac{C_2^2 p}{L}$. Observe that $\Pi_3^* - \Pi_6^*$ is monotone in L, it increases in L if $C_1 > C_2$, and decreases otherwise.

Case (A.ii) $I > \gamma p$ and $\Pi_8'(C_2) \geq 0$, which implies that $L \geq \frac{2C_2 p}{\gamma I}$. In this case, AVs do not reposition, we compare Π_3^*, Π_6^* and Π_8^*. Because $\Pi_6^* = \Pi_6(C_2) = \Pi_8(C_2) < \Pi_8^*$, it suffices to compare Π_3^* and Π_8^*. We consider the following subcases.

Case (A.ii.i) $\Pi_3'(C_1) \leq 0$ and $\Pi_8'(C_1 + C_2 + q_1^*) \leq 0$, which implies that $L \leq \frac{2C_1 p}{I}$. We have $\Pi_8^* - \Pi_3^* = \Pi_8\left(\frac{\gamma IL}{2p}\right) - \Pi_3\left(\frac{IL}{2p}\right) = \gamma C_2(p-I) + \frac{\gamma^2 I^2 L}{4p} - \frac{I^2 L}{4p} \overset{(a)}{\leq} \frac{\gamma^2 IL}{2} - \frac{\gamma^2 I^2 L}{4p} - \frac{I^2 L}{4p} \overset{(b)}{<} 0$, where (a) is due to $L \geq \frac{2C_2 p}{\gamma I}$ and (b) is due to $I > \gamma p$.

Case (A.ii.ii) $\Pi_3'(C_1) \geq 0$ and $\Pi_8'(C_1+C_2+q_1^*) \leq 0$, which implies $L \in [\frac{2C_1 p}{I}, \frac{2(C_1+C_2+q_1^*)p}{\gamma I}]$. We have $\Pi_8^* - \Pi_3^* = \Pi_8\left(\frac{\gamma IL}{2p}\right) - \Pi_3(C_1) = \frac{\gamma^2 I^2 L}{4p} + \gamma C_2(p-I) - C_1 I + \frac{C_1^2 p}{L} \leq 0$ if and only

$$L \in \left[\frac{2p\left[C_1 I - \gamma C_2(p-I) - \sqrt{[C_1 I - \gamma C_2(p-I)]^2 - \gamma^2 I^2 C_1^2}\right]}{\gamma^2 I^2}, \right.$$
$$\left. \frac{2p\left[C_1 I - \gamma C_2(p-I) + \sqrt{[C_1 I - \gamma C_2(p-I)]^2 - \gamma^2 I^2 C_1^2}\right]}{\gamma^2 I^2} \right].$$

Case (A.ii.iii) $\Pi_3'(C_1) \geq 0$ and $\Pi_8'(C_1+C_2+q_1^*) \geq 0$, which implies that $L \geq \frac{2(C_1+C_2+q_1^*)p}{\gamma I}$. We have $\Pi_8^* - \Pi_3^* = \Pi_8(C_1 + C_2 + q_1^*) - \Pi_3(C_1) = \gamma C_2 p - \frac{(C_1+C_2+q_1^*)^2 p}{L} + \frac{C_1^2 p}{L} \geq 0$ only if $L \geq \frac{(C_2+q_1^*)(2C_1+C_2+q_1^*)}{\gamma C_2}$.

Case (A.iii) $I \leq \gamma p$ and $\Pi_8'(C_2) \leq 0$, which implies that $L \leq \frac{2C_2 p}{\gamma I}$. In this case, the platform has an incentive to reposition AVs. Therefore, we compare Π_4^*, Π_7^* and Π_8^*. Because $\Pi_7^* \geq \Pi_7(C_2) = \Pi_8(C_2) = \Pi_8^*$, it suffices to compare Π_4^* and Π_7^*. Observe that $\Pi_4(N)$ and $\Pi_7(N)$ has the same form, but different domains, (i.e., $N \in [0, C_1]$ for $\Pi_4(N)$ and $N \in [0, C_2]$ for $\Pi_7(N)$). We consider the following subcases.

Case (A.iii.i) $\Pi_4'(C_1) \leq 0$ and $\Pi_7'(C_2) \leq 0$, which implies that $L \leq \min\left(\frac{2C_1 p}{I}, \frac{2C_2 p}{I}\right)$. We have $\Pi_4^* - \Pi_7^* = \Pi_4(\frac{IL}{2p}) - \Pi_7(\frac{IL}{2p}) = 0$.

Case (A.iii.ii) $\Pi_4'(C_1) \leq 0$ and $\Pi_7'(C_2) \geq 0$, which implies that $L \in [\frac{2C_2 p}{I}, \frac{2C_1 p}{I}]$. We have $\Pi_4^* - \Pi_7^* = \Pi_4(\frac{IL}{2p}) - \Pi_7(C_2) \geq 0$.

Case (A.iii.iii) $\Pi_4'(C_1) \geq 0$ and $\Pi_7'(C_2) \leq 0$, which implies that $L \in [\frac{2C_1 p}{I}, \frac{2C_2 p}{I}]$. We have $\Pi_4^* - \Pi_7^* = \Pi_4(C_1) - \Pi_7(\frac{IL}{2p}) \leq 0$.

Case (A.iii.iv) $\Pi_4'(C_1) \geq 0$ and $\Pi_7'(C_2) \geq 0$, which implies that $L \geq \max\{\frac{2C_1 p}{I}, \frac{2C_2 p}{I}\}$. We have $\Pi_4^* - \Pi_7^* = \Pi_4(C_1) - \Pi_7(C_2) \leq 0$ if and only if $C_1 \leq C_2$.

Case (A.iv), $I \leq \gamma p$ and $\Pi_8'(C_2) \geq 0$, which implies that $L \geq \frac{2C_2 p}{\gamma I}$. In this case, the platform has an incentive to reposition AVs. Therefore, we compare Π_4^*, Π_7^* and Π_8^*. Because $\Pi_7^* = \Pi_7(C_2) = \Pi_8(C_2) \leq \Pi_8^*$, it suffices to compare Π_4^* and Π_8^*. We consider the following subcases.

Case (A.iv.i) $\Pi_4'(C_1) \leq 0$ and $\Pi_8'(C_1 + C_2 + q^*) \leq 0$, which implies that $L \leq \frac{2C_1 p}{I}$. We have $\Pi_8^* - \Pi_4^* = \Pi_8\left(\frac{\gamma IL}{2p}\right) - \Pi_4\left(\frac{IL}{2p}\right) = (1-\gamma)I\left[C_2 - \frac{IL}{4p}(1+\gamma)\right] < 0$ as $L \geq \frac{2C_2 p}{\gamma I}$.

Case (A.iv.ii) $\Pi_4'(C_1) > 0$ and $\Pi_8(C_1+C_2+q_1^*) < 0$, which implies $L \in [\frac{2C_1 p}{I}, \frac{2(C_1+C_2+q_1^*)p}{\gamma I}]$. We have $\Pi_8^* - \Pi_4^* = \Pi_8\left(\frac{\gamma IL}{2p}\right) - \Pi_4(C_1) = \frac{\gamma^2 I^2 L}{4p} + C_2 I(1-\gamma) - C_1 I + \frac{C_1^2 p}{L} \leq 0$ if and only if

$$L \in \left[\frac{2p\left[C_1 - C_2(1-\gamma) - \sqrt{[C_1-(1-\gamma)C_2]^2 - C_1^2\gamma^2}\right]}{\gamma^2 I}, \right. \\ \left. \frac{2p\left[C_1 - C_2(1-\gamma) + \sqrt{[C_1-(1-\gamma)C_2]^2 - C_1^2\gamma^2}\right]}{\gamma^2 I}\right].$$

Case (A.iv.iii) $\Pi_4'(C_1) \geq 0$ and $\Pi_8'(C_1 + C_2 + q^*) \geq 0$, which implies that $L \geq \frac{2(C_1+C_2+q_1^*)p}{\gamma I}$. We have $\Pi_8^* - \Pi_4^* = C_2 I - \frac{p(C_1+C_2+q_1^*)^2 - C_1^2 p}{L} \geq 0$ if and only if $L \geq \frac{p(C_2+q_1^*)(2C_1+C_2+q_1^*)}{C_2 I}$.

Step (3). We characterize the platform's optimal strategy and the corresponding outcomes. For convenience, we define the following three types of outcomes with respect to the capacity allocation of CVs.

- Type I outcome: CVs do not queue and reposition, and fulfill only type-1 demand.
- Type II outcome: CVs do not queue, and reposition with probability 1 to fulfill only type-2 demand.
- Type III outcome: CVs reposition with a positive probability to fulfill all the type-2 demand and the type-1 demand that are not fulfilled by AVs. Moreover, there is a queue of CVs with $q_1^C = (1 - F_1)q_1^*$.

By the analysis in step (2), it suffices to compare Π_3^*, Π_6^* and Π_8^* if $I > \gamma p$, and it suffices to compare Π_4^*, Π_7^* and Π_8^* otherwise. In the former case, $\Pi_3^* = \max(\Pi_3^*, \Pi_6^*, \Pi_8^*)$

implies a type I outcome, $\Pi_6^* = \max(\Pi_3^*, \Pi_6^*, \Pi_8^*)$ implies a type II outcome, and $\Pi_8^* = \max(\Pi_3^*, \Pi_6^*, \Pi_8^*)$ implies a type III outcome; and in the latter case $\Pi_4^* = \max(\Pi_4^*, \Pi_7^*, \Pi_8^*)$ implies a type I outcome, $\Pi_7^* = \max(\Pi_4^*, \Pi_7^*, \Pi_8^*)$ implies a type II outcome, and $\Pi_8^* = \max(\Pi_4^*, \Pi_7^*, \Pi_8^*)$ implies a type III outcome.

We first consider the case where $I > \gamma p$, and consider the possible relationships among $\frac{2C_2}{\gamma}$, $\frac{2C_2 p}{\gamma I}$ and $\frac{2C_1 p}{I}$. We have the following three scenarios.

Scenario (A.a), $\frac{2C_2}{\gamma} < \frac{2C_2 p}{\gamma I} < \frac{2C_1 p}{I}$. When $L \leq \frac{2C_1 p}{I}$, by case (A.i.i), (A.i.ii) and (A.ii.i), the platform's optimal strategy results in a type I outcome. When $L \geq \frac{2C_1 p}{I}$, we have the following analysis. Because $\Pi_8^* - \Pi_3^* \leq 0$ when $L = \frac{2C_1 p}{I}$, by case (A.ii.ii), we must have

$$\frac{2C_1 p}{I} \in \left[\frac{2p\left[C_1 I - \gamma C_2 (p-I) - \sqrt{[C_1 I - \gamma C_2 (p-I)]^2 - \gamma^2 I^2 C_1^2}\right]}{\gamma^2 I^2}, \right.$$
$$\left. \frac{2p\left[C_1 I - \gamma C_2 (p-I) + \sqrt{[C_1 I - \gamma C_2 (p-I)]^2 - \gamma^2 I^2 C_1^2}\right]}{\gamma^2 I^2} \right].$$

We then consider two possibilities. (i) If $\frac{2p\left[C_1 I - \gamma C_2 (p-I) + \sqrt{[C_1 I - \gamma C_2 (p-I)]^2 - \gamma^2 I^2 C_1^2}\right]}{\gamma^2 I^2} < \frac{2(C_1 + C_2 + q_1^*)p}{\gamma I}$, let $L_1 = \frac{2p\left[C_1 I - \gamma C_2 (p-I) - \sqrt{[C_1 I - \gamma C_2 (p-I)]^2 - \gamma^2 I^2 C_1^2}\right]}{\gamma^2 I^2}$. By case (A.ii.ii) and (A.ii.iii), the platform's optimal strategy results in a type I outcome for $L \in [\frac{2C_1 p}{I}, L_1]$. Because $\Pi_3^* = \Pi_8^*$ when $L = L_1$ and $\Pi_3^* - \Pi_8^*$ decreases in L for $L \geq L_1$, the platform's optimal strategy results in a type III outcome for $L \geq L_1$. (ii) Otherwise, let $L_1 = \frac{(C_2 + q_1^*)(2C_1 + C_2 + q_1^*)}{\gamma C_2}$. By case (A.ii.ii) and (A.ii.iii), the platform's optimal strategy results in a type I outcome if $L < L_1$, and that results in a type III outcome otherwise. To summarize, define

$$L_1 = \begin{cases} \frac{2p\left[C_1 I - \gamma C_2 (p-I) + \sqrt{[C_1 I - \gamma C_2 (p-I)]^2 - \gamma^2 I^2 C_1^2}\right]}{\gamma^2 I^2}, & \text{if } \frac{2p\left[C_1 I - \gamma C_2 (p-I) + \sqrt{[C_1 I - \gamma C_2 (p-I)]^2 - \gamma^2 I^2 C_1^2}\right]}{\gamma^2 I^2} \\ & \quad < \frac{2(C_1 + C_2 + q_1^*)p}{\gamma I} \\ \frac{(C_2 + q_1^*)(2C_1 + C_2 + q_1^*)}{\gamma C_2}, & \text{otherwise.} \end{cases} \tag{B.24}$$

The platform's optimal strategy results in a type I outcome if $L \leq L_1$, and that results in

a type III outcome otherwise. In this case, the amount of workers recruited is

$$N^A = \begin{cases} \min(\frac{IL}{2p}, C_1) & \text{if } L \le L_1, \\ \min(\frac{\gamma IL}{2p}, C_1 + C_2 + q_1^*) & \text{otherwise.} \end{cases} \tag{B.25}$$

Scenario (A.b) $\frac{2C_2}{\gamma} < \frac{2C_1p}{I} < \frac{2C_2p}{\gamma I}$. When $L < \frac{2C_1p}{I}$, by case (A.i.i), (A.i.ii) and (A.ii.i), the platform's optimal strategy results in a type I outcome. When $L \in [\frac{2C_1p}{I}, \frac{2C_2p}{\gamma I}]$, by case (A.i.iv), because $L(\Pi_3^* - \Pi_6^*) = L(C_1I - \gamma C_2p) - p(C_1^2 - C_2^2) \overset{\text{(a)}}{\ge} \frac{2C_1p}{I}(C_1I - \gamma C_2p) - p(C_1^2 - C_2^2) \overset{\text{(b)}}{\ge} pC_1^2 - 2C_2C_1p + pC_2^2 \ge 0$, where (a) is due to $L \ge \frac{2C_1p}{I}$ and (b) is due to $I \ge \gamma p$, the platform's optimal strategy results in a type I outcome. When $L \ge \frac{2C_2p}{\gamma I}$, the analysis is the same as that for scenario (A.a) for the case where $L > \frac{2C_1p}{I}$. To summarize, there exists threshold L_1 defined in (B.24) such that the optimal strategy of the platform results in a type I outcome if $L \le L_1$, and that results in a type III outcome otherwise. Moreover, the amount of workers recruited is given by (B.25).

Scenario (A.c) $\frac{2C_1p}{I} < \frac{2C_2}{\gamma} < \frac{2C_2p}{\gamma I}$. When $L \le \frac{2C_1p}{I}$, by case (A.i.i), the platform's optimal strategy results in a type I outcome. When $L \ge \frac{2C_1p}{I}$, we have the following analysis. Because $\Pi_3^* - \Pi_6^* \ge 0$ when $L = \frac{2C_1p}{I}$, by case (A.i.iii), we must have

$$\frac{2C_1p}{I} \in \left[\frac{2C_1\left(I - \sqrt{I^2 - \gamma^2p^2}\right)}{\gamma^2p}, \frac{2C_1\left(I + \sqrt{I^2 - \gamma^2p^2}\right)}{\gamma^2p} \right].$$

We then consider two possibilities. (i) If $\frac{2C_1\left(I+\sqrt{I^2-\gamma^2p^2}\right)}{\gamma^2p} < \frac{2C_2}{\gamma}$, which implies that $C_1 < \frac{\gamma pC_2}{I+\sqrt{I^2-\gamma^2p^2}}$, let $L_2 = \frac{2C_1\left(I+\sqrt{I^2-\gamma^2p^2}\right)}{\gamma^2p}$. By case (A.i.iii) the platform's optimal strategy results in a type I outcome when $L \in [\frac{2C_1p}{I}, L_2]$, and that results in a type II when $L \in [L_2, \frac{2C_2}{\gamma}]$. When $L \in [\frac{2C_2}{\gamma}, \frac{2C_2p}{\gamma I}]$, by case (A.i.iv), $\Pi_3^* - \Pi_6^*$ decreases in L as $C_2 > C_1$ (implied by $\frac{2C_1\left(I+\sqrt{I^2-\gamma^2p^2}\right)}{\gamma^2p} \le \frac{2C_2}{\gamma}$) and $\Pi_3^* \le \Pi_6^*$ when $L = \frac{2C_2}{\gamma}$, the platform 's optima strategy results in a type II outcome. When $L \ge \frac{2C_2p}{\gamma I}$, the platform's optimal strategy results in a type III outcome by case (A.ii). (ii) Otherwise, by case (A.i.iii), when $L \in [\frac{2C_1p}{I}, \frac{2C_2}{\gamma}]$, the optimal strategy of the platform results in a type I outcome. When $L \in [\frac{2C_2}{\gamma}, \frac{2C_2p}{\gamma I}]$, note that $\Pi_3^* - \Pi_6^* \ge 0$ when $L = \frac{2C_2}{\gamma}$, and $\Pi_3^* - \Pi_6^*$ is monotone in L by case (A.i.iv). Therefore, it suffice to check the value of $\Pi_3^* - \Pi_6^*$ when $L = \frac{2C_2p}{\gamma I}$.

When $L = \frac{2C_2p}{\gamma I}$, if $\Pi_3^* - \Pi_6^* = \frac{2C_2p}{\gamma I}[C_1 I - \gamma C_2 p] - p(C_1^2 - C_2^2) \geq 0$, which implies that $C_1 \in [\frac{IC_2 - \sqrt{I^2C_2^2 + I\gamma(I\gamma - 2p\gamma)C_2^2}}{I\gamma}, \frac{IC_2 + \sqrt{I^2C_2^2 + I\gamma(I\gamma - 2p\gamma)C_2^2}}{I\gamma}]$, the platform's optimal strategy results in a type I equilibrium. Then by following the same analysis that in scenario (A.a), there exists a threshold L_1 define in (B.24) such that the platform's optimal strategy results in a type I outcome for $L \in [\frac{2C_2p}{\gamma I}, L_1]$, and that results in a type III outcome for $L \geq L_1$. Otherwise, combined with $\frac{2C_1p}{I} < \frac{2C_2}{\gamma}$, we have must have $C_1 < \left(\frac{1}{\gamma} - \sqrt{\frac{1}{\gamma^2} - \frac{2p}{I} + 1}\right) C_2$. Therefore, if $\frac{C_2\gamma p}{I + \sqrt{I^2 - \gamma^2 p^2}} < C_1 < \left(\frac{1}{\gamma} - \sqrt{\frac{1}{\gamma^2} - \frac{2p}{I} + 1}\right) C_2$, we redefine $L_2 = \frac{p(C_1^2 - C_2^2)}{C_1 I - \gamma C_2 p}$. Then the platform's optimal strategy results in a type I outcome for $L \in [\frac{2C_2}{\gamma}, L_2]$, and that results in a type II outcome for $L \in [L_2, \frac{2C_2p}{\gamma I}]$. When $L \geq \frac{2C_2p}{\gamma I}$, the platform's optimal strategy results in a type III outcome by case (A.ii). To summarize, define

$$\theta = \max\left(\frac{1}{\gamma} - \sqrt{\frac{1}{\gamma^2} - \frac{2p}{I} + 1}, \frac{\gamma p}{I + \sqrt{I^2 - \gamma^2 p^2}}\right) < 1, \quad \text{and} \tag{B.26}$$

$$L_2 = \begin{cases} \frac{2C_1\left(I + \sqrt{I^2 - \gamma^2 p^2}\right)}{\gamma^2 p}, & \text{if } C_1 \leq \frac{C_2\gamma p}{I + \sqrt{I^2 - \gamma^2 p^2}}, \\ \frac{p(C_1^2 - C_2^2)}{C_1 I - \gamma C_2 p}, & \text{if } \frac{C_2\gamma p}{I + \sqrt{I^2 - \gamma^2 p^2}} < C_1 < \left(\frac{1}{\gamma} - \sqrt{\frac{1}{\gamma^2} - \frac{2p}{I} + 1}\right) C_2. \end{cases} \tag{B.27}$$

If $C_1 \leq \theta C_2$, there exists a threshold L_2 such that the platform's optimal strategy results in a type I outcome for $L \leq L_2$, that results in a type II outcome for $L \in [L_2, \frac{2C_2p}{\gamma I}]$, and that results in a type III outcome for $L \geq \frac{2C_2p}{\gamma I}$. In this case, the amount of drivers recruited is

$$N^A = \begin{cases} \min(\frac{IL}{2p}, C_1) & \text{if } L \leq L_2, \\ \min(\frac{\gamma L}{2}, C_2) & \text{if } L \in [L_2, \frac{2C_2p}{\gamma I}], \\ \min(\frac{\gamma IL}{2p}, C_1 + C_2 + q_1^*) & \text{otherwise.} \end{cases} \tag{B.28}$$

Otherwise, the platform's optimal strategy results in a type I outcome for $L \leq L_1$, and that results in a type III outcome otherwise, where L_1 is define in (B.24). Moreover, the amount of workers recruited is given by (B.25).

We then consider the case where $I \leq \gamma p$, and consider the possible relationships among $\frac{2C_2p}{I}$, $\frac{2C_2p}{\gamma I}$ and $\frac{2C_1p}{I}$. Similar to the analysis under the random assignment policy, we focus on the case where the repositioning of CVs is minimized [2]. In particular, in the case where

[2] All comparison results in Theorem 3.5.2 and Proposition 3.5.1 hold without this assumption.

$\Pi_4^* = \Pi_7^*$, we assume that the platform's optimal strategy results in a type I outcome rather than a type II outcome.

Scenario (A.d) $\frac{2C_2p}{I} \leq \frac{2C_2p}{\gamma I} \leq \frac{2C_1p}{I}$. When $L < \frac{2C_1p}{I}$, by case (A.iii.i), (A.iii.ii) and (A.iv.i), the platform's optimal strategy results in a type I outcome. When $L \geq \frac{2C_1p}{I}$, we have the following analysis. Because $\Pi_4^* \geq \Pi_8^*$ when $L = \frac{2C_1p}{I}$, by case (A.iv.ii),

$$\frac{2C_1p}{I} \in \left[\frac{2p\left[C_1 - C_2(1-\gamma) - \sqrt{[C_1-(1-\gamma)C_2]^2 - C_1^2\gamma^2}\right]}{\gamma^2 I}, \right. \\ \left. \frac{2p\left[C_1 - C_2(1-\gamma) + \sqrt{[C_1-(1-\gamma)C_2]^2 - C_1^2\gamma^2}\right]}{\gamma^2 I} \right].$$

We then consider two possibilities. (i) If $\frac{2p\left[C_1-C_2(1-\gamma)+\sqrt{[C_1-(1-\gamma)C_2]^2-C_1^2\gamma^2}\right]}{\gamma^2 I} < \frac{2(C_1+C_2+q_1^*)p}{\gamma I}$, let $L_3 = \frac{2p\left[C_1-C_2(1-\gamma)+\sqrt{[C_1-(1-\gamma)C_2]^2-C_1^2\gamma^2}\right]}{\gamma^2 I}$, by case (A.iv.ii), when $L \in [\frac{2C_1p}{I}, L_3]$, the platform's optimal strategy results in a type I outcome. Because $\Pi_4^* = \Pi_8^*$ when $L = L_3$, and by case (A.iv.ii) and (A.iv.iii), $\Pi_4^* - \Pi_8^*$ decreases in L for $L \geq L_3$, the platform's optimal strategy results in a type III outcome for $L \geq L_3$. (ii) Otherwise, let $L_3 = \frac{p(C_2+q_1^*)(2C_1+C_2+q_1^*)}{C_2 I}$. By case (A.iv,ii) and (A.iv.iii), the platform's optimal strategy results in a type I outcome if $L \in [\frac{2C_1p}{I}, L_3]$, and that results in a type III outcome if $L \geq L_3$. To summarize, let

$$L_3 = \begin{cases} \frac{2p\left[C_1-C_2(1-\gamma)+\sqrt{[C_1-(1-\gamma)C_2]^2-C_1^2\gamma^2}\right]}{\gamma^2 I}, & \text{if } \frac{2p\left[C_1-C_2(1-\gamma)+\sqrt{[C_1-(1-\gamma)C_2]^2-C_1^2\gamma^2}\right]}{\gamma^2 I} \\ & \qquad < \frac{2(C_1+C_2+q_1^*)p}{\gamma I}, \\ \frac{p(C_2+q_1^*)(2C_1+C_2+q_1^*)}{C_2 I}, & \text{otherwise.} \end{cases} \tag{B.29}$$

The platform's optimal strategy results in a type I outcome for $L \leq L_3$ and that results in a III outcome for $L \geq L_3$. Moreover, the amount of workers recruited

$$N^A = \begin{cases} \min(\frac{IL}{2p}, C_1) & \text{if } L \leq L_3, \\ \min(\frac{\gamma IL}{2p}, C_1 + C_2 + q_1^*) & \text{otherwise.} \end{cases} \tag{B.30}$$

Scenario (A.e) $\frac{2C_2p}{I} < \frac{2C_1p}{I} < \frac{2C_2p}{\gamma I}$. When $L \leq \frac{2C_2p}{\gamma I}$, by case (A.iii.i), (A.iii.ii) and

(A.iii.iv), the platform's optimal strategy results in a type I outcome. When $L \geq \frac{2C_2p}{\gamma I}$, the analysis is similar to that in scenario (A.d) for the case where $L \geq \frac{2C_1p}{I}$. Therefore, the platform's optimal strategy results in a type I outcome if $L < L_3$, and that results in a type III outcome otherwise. Moreover, the amount of drivers recruited is given by (B.30).

Scenario (A.f) $\frac{2C_1p}{I} < \frac{2C_2p}{I} < \frac{2C_2p}{\gamma I}$. When $L \leq \frac{2C_1p}{I}$, by case (A.iv.i), the platform's optimal strategy results in a type I outcome. When $L \in [\frac{2C_1p}{I}, \frac{2C_2p}{\gamma I}]$, by case (A.iii.ii) and (A.iii.iv), the platform's optimal strategy result in a type II outcome. When $L \geq \frac{2C_2p}{\gamma I}$, by case (A.iv), the platform's optimal strategy results in a type III outcome. In this case, the amount of workers recruited is

$$N^A = \begin{cases} \min(\frac{IL}{2p}, C_1) & \text{if } L \leq \frac{2C_1p}{I}, \\ \min(\frac{IL}{2p}, C_2) & \text{if } L \in [\frac{2C_1p}{I}, \frac{2C_2p}{\gamma I}], \\ \min(\frac{\gamma IL}{2p}, C_1 + C_2 + q_1^*) & \text{otherwise.} \end{cases} \tag{B.31}$$

By combining the analysis through scenario (A.a) to (A.f), we can define

$$L^A = \begin{cases} L_1, & \text{if } I > \rho p \text{ and } C_1 \geq \theta C_2, \\ L_2, & \text{if } I > \rho p \text{ and } C_1 < \theta C_2, \\ L_3, & \text{if } I \leq \rho p \text{ and } C_1 \geq C_2, \\ \frac{2C_1p}{I}, & \text{if } I \leq \rho p \text{ and } C_1 < C_2, \end{cases} \tag{B.32}$$

where L_1, L_2 and L_3 are defined respectively in (B.24), (B.27) and (B.29). Under the platform's optimal strategy, CVs do not reposition if $L \leq L^A$, and CVs reposition such that $(1 - F_1)((\Lambda_{21} - \Lambda_{12})t_{12} - r_{12}^C) = 0$ otherwise. Moreover, from the expressions of N^A in (B.25), (B.28), (B.30) and (B.31), we can observe that N^A weakly increases in L.

Finally, we note that the AV-prioritized policy outperforms the other two policies as any incentive compatible capacity allocation that is achieved under the CV-prioritized policy (see Lemma B.1.2) and the random assignment policy (see Lemma B.2.2) can also be achieved under the AV-prioritized policy (see Lemma B.4.1). Therefore, the platform can use the AV-prioritized policy to mimic the outcomes under the other two policies

B.5 Comparison of Systems with and without AVs

B.5.1 Proof of Theorem 3.5.2

By (3.10), it suffices to compare the number of drivers recruited under the optimal policy. The amount of drivers recruited in a system without AVs (N^C) is given by (B.11). The amount of drivers recruited in a system with AVs under the AV-prioritized policy (N^A) is given by (B.25), (B.28), (B.30) or (B.31), depending on the parameters (i.e., I, C_1 and C_2). Therefore, we consider 4 cases.

Case (i) $I \leq \gamma p$ and $C_1 < C_2$. In this case, N^A is given in (B.31). We then consider 3 subcases.

Case (i.i) $L^C < \frac{2C_1 p}{I}$. We illustrate the values of N^C and N^A in Figure B.1.

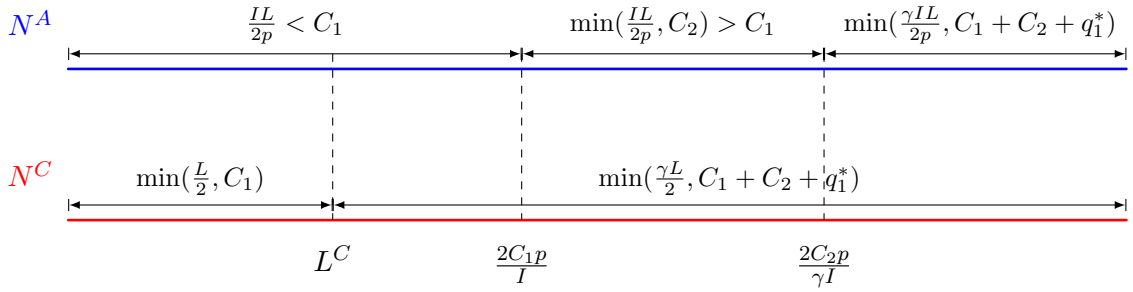


Figure B.1: An illustration of N^A and N^C with respect to L when $I \leq \gamma p$ and $C_1 < C_2$.

By Figure B.1, $N^A = \frac{IL}{2p} \leq N^C = \min(\frac{L}{2}, C_1)$ if $L \leq L^C$, $N^A = \frac{IL}{2p} \leq N^C = \min(\frac{\gamma L}{2}, C_1 + C_2 + q_1^*)$ (as $I \leq \gamma p$) if $L \in [L^C, \frac{2C_1 p}{I}]$, $N^A = \min(\frac{IL}{2p}, C_2) \leq N^C = \min(\frac{\gamma L}{2}, C_1 + C_2 + q_1^*)$ if $L \in [\frac{2C_1 p}{I}, \frac{2C_2 p}{\gamma I}]$, and $N^A = \min(\frac{\gamma IL}{2p}, C_1 + C_2 + q_1^*) \leq N^C = \min(\frac{\gamma IL}{2p}, C_1 + C_2 + q_1^*)$ otherwise.

Case (i.ii) $L^C \in [\frac{2C_1 p}{I}, \frac{2C_2 p}{\gamma I}]$. We illustrate the values of N^A and N^C in Figure B.2.

By Figure B.2, $N^A = \frac{IL}{2p} \leq N^C = \min(\frac{L}{2}, C_1)$ if $L \leq \frac{2C_1 p}{I}$, $N^A = \min(\frac{IL}{2p}, C_2) > C_1 \geq \min(\frac{L}{2}, C_1) = N^C$ if $L \in (\frac{2C_1 p}{I}, L^C)$, $N^A = \min(\frac{IL}{2p}, C_2) \leq N^C = \min(\frac{\gamma L}{2}, C_1 + C_2 + q_1^*)$ (as $I \leq \gamma p$) if $L \in [L^C, \frac{2C_2 p}{\gamma I}]$, and $N^A = \min(\frac{\gamma IL}{2p}, C_1 + C_2 + q_1^*) \leq N^C = \min(\frac{\gamma L}{2}, C_2 + C_2 + q_1^*)$ otherwise.

Case (i.iii) $L^C > \frac{2C_1 p}{I}$. We illustrate the values of N^A and N^C in Figure B.3.

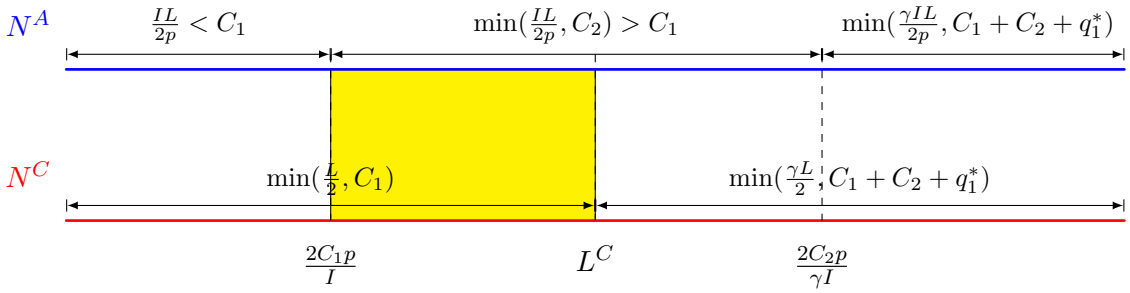


Figure B.2: An illustration of N^A and N^C with respect to L when $I \leq \gamma p$ and $C_1 < C_2$. The region where drivers are better of after the introduction of AVs is highlighted in yellow.

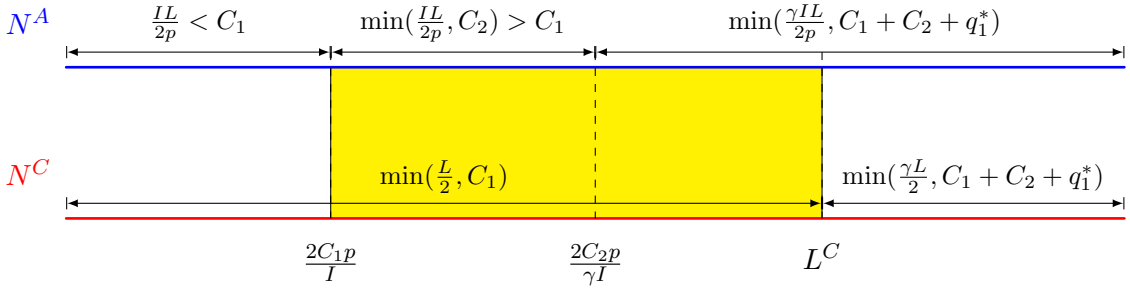


Figure B.3: An illustration of N^A and N^C with respect to L when $I \leq \gamma p$ and $C_1 < C_2$. The region where drivers are better of after the introduction of AVs is highlighted in yellow.

By Figure B.3, $N^A = \frac{IL}{2p} \leq C_1 \leq N^C = \min(\frac{L}{2}, C_1)$ if $L \leq \frac{IL}{2p}$, $N^A = \min(\frac{IL}{2p}, C_2) > C_1 > \min(\frac{L}{2}, C_1) = N^C$ if $L \in (\frac{2C_1p}{I}, \frac{2C_2p}{\gamma I}]$, $N^A = \min(\frac{\gamma IL}{2p}, C_1 + C_2 + q_1^*) > C_2 > C_1 > N^C$ if $L \in (\frac{2C_1p}{\gamma I}, L^C)$, and $N^A = \min(\frac{\gamma IL}{2p}, C_1 + C_2 + q_1^*) \leq N^C = \min(\frac{\gamma L}{2}, C_1 + C_2 + q_1^*)$ otherwise.

Combined the analysis in case (i.i), (i.ii) and (i.iii), $N^A > N^C$ if and only if $\frac{2C_1p}{I} < L < \hat{L}$.

By the same analysis as in case (i), we can show that (ii) when $I \leq \gamma p$, $C_1 \geq C_2$, $N^A > N^C$ if and only if $L_3 < L < \hat{L}$; (iii) when $I > \gamma p$, $C_1 < \theta C_2$, $N^A > N^C$ if and only if $L_2 < L < \hat{L}$, and (iv) when $I > \gamma p$, $C_1 \geq \theta C_2$, $N^A > N^C$ if and only if $L_1 < L < \hat{L}$.

We last consider the wage and the repositioning of drivers. We start by showing that $w^A > w^C$ if and only if $L \in (L^A, L^C)$. For convenience, let ρ^A (ρ^C) and $\hat{w}^A$ ($\hat{w}^C$) denote the driver's utilization and the effective wage in a system with and without AVs respectively. By the previous analysis, $\hat{N}^A > \hat{N}^C$ if and only if $L \in (L^A, L^C)$. By the proof of Theorem 3.4.1 and that of Theorem 3.5.1, $\rho^C = 1$ if $L \leq L^C$ and $\rho^C = \gamma$ otherwise; $\rho^A = 1$ if $L \leq L^A$ and $\rho^A = \gamma$ otherwise. Therefore, if $L \leq \min(L^A, L^C)$, $\hat{w}^A = w^A \leq \hat{w}^C = w^C$; if $L \in (L^A, L^C)$, $\hat{w}^A = \gamma w^A > \hat{w}^C = w^C$ implies that $w^A > w^C$; if $L \in (L^C, L^A)$, $\hat{w}^A = w^A < \hat{w}^C = \gamma w^C$ implies that $w^C > w^A$; and otherwise $\hat{w}^A = \gamma w^A = \hat{w}^C = \gamma w^C$ implies that $w^A \leq w^C$. We then show that $(r_{12}^C)^A > (r_{12}^C)^C$ for $L \in (L^A, L^C)$. Because $\rho^A = \gamma$ and $\rho^C = 1$ for $L \in (L^A, L^C)$, we must have $(r_{12}^C)^C = 0 < (r_{12}^C)^A$.

B.5.2 Proof of Proposition 3.5.1

By (3.15) and (B.13), we observe that $L^C \leq L^R$. Then the comparison results on worker welfare follows directly by comparing the amount of drivers recruited in a system without AV given in (B.11), and that under the random assignment/CV-prioritized policy (recall from Proposition B.3.1 that these two policies induce the same amount of drivers recruited) given in (B.19).

B.6 Location-Dependent Pricing

In this section, we consider a setting where the platform can adjust prices based on the origin and destination of the requested trip. We do so to test the robustness of our main

result (regarding the possibility of driver welfare improving with the introduction of AVs) when the platform has the additional lever of location-dependent pricing at its disposal. Specifically, we assume that the platform charges a base price p per unit of travel time for trips originating from location 1 (the low demand location) and a price $p + \kappa$ for trips originating from location 2 (the high-demand location). Additionally, in contrast to our original model, where the platform pays a fixed wage w per unit of time the driver spends transporting customers, we assume that the platform pays drivers a fixed percentage $(1 - \beta)\%$ of the price it charges customers. This allows us to also let the wage rate be location-dependent.

We assume that customers' valuation of the service, denoted by v, follows a continuous probability distribution with a cumulative distribution function $F(\cdot)$. Specifically, customers originating from location 1 (location 2) choose to seek service from the platform if their valuation, v, is greater than or equal to p $(p + \kappa)$. We let $\overline{\Lambda}_{ij}$ denote the potential demand rate from location i to location j, and we assume that $\overline{\Lambda}_{12} < \overline{\Lambda}_{21}$. Then the realized demand rate from location 1 to location 2 is $\Lambda_{12} = \bar{\Lambda}_{12}(1 - F(p))$, and that from location 2 to location 1 is $\Lambda_{21} = \bar{\Lambda}_{21}(1 - F(p + \kappa))$. Recall that we use s_{ij}^C to denote the volume of CVs in service from location i to location j. Then drivers' expected earnings (i.e., the effective wage) can be expressed as

$$\hat{w} = \frac{(1 - \beta)[ps_{12}^C + (p + \kappa)s_{21}^C]}{N}, \tag{B.33}$$

where N is the supply of CVs which satisfies (3.10).

In a system without AVs, the platform decides on the price adjustment κ and the commission rate β. The platform solves the following problem:

$$\begin{aligned} \text{(Problem E1)} \qquad \max_{\kappa,\beta} \quad & \Pi^C = \beta p(s_{12}^C + s_{21}^C) + \beta\kappa s_{21}^C \\ \text{subject to} \quad & (3.1), (3.2), (3.4), (3.6), (3.7), (3.9), (3.10), (B.33), \\ & \eta^C \text{ is a CV equilibrium repositioning strategy,} \\ & M = 0 \text{ and } F_k = 0 \text{ for } k = 1, 2. \end{aligned}$$

In a system with AVs, the platform decides on the price adjustment κ, the commission

rate β, the AV fleet size M, and the AVs' repositioning strategy η^A. The platform solves the following problem under the AV-prioritized policy.

$$\begin{aligned} \text{(Problem E2)} \qquad & \max_{\kappa,\beta,M,\eta^A} \quad \Pi = ps_{12}^A + (p+\kappa)s_{21}^A + \beta[ps_{12}^C + (p+\kappa)s_{21}^C] - M \cdot I, \\ & \text{subject to} \quad \text{(3.1)–(3.9), (3.10), (3.17), (B.33) and} \\ & \qquad\qquad\quad \eta^C \text{ is a CV equilibrium repositioning strategy.} \end{aligned}$$

Analytical results are difficult to obtain for this case. However, numerical results (see Figure B.4) suggest that introduction of AVs can still result in an improvement in the welfare of drivers.

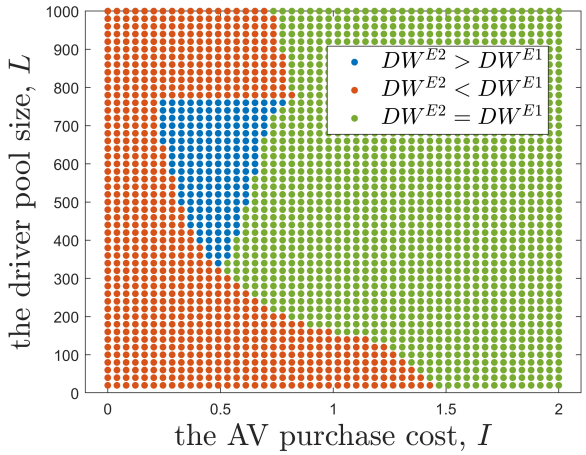


Figure B.4: Driver welfare in systems with and without AVs, where DW^{E1} and DW^{E2} denote the driver welfare under the optimal solutions to Problem E1 and E2 respectively. Model parameters: $\bar{\Lambda}_{12} = 20$, $\bar{\Lambda}_{21} = 80$, $t_{12} = t_{21} = 1$, $p = 1$, $\bar{w} = 2$ and $v \sim U[1, 2]$.

Appendix C

Appendices for Chapter 4

We present the subgame analysis in Appendix C.1, characterize potential duopoly equilibria in Appendix C.2, prove the existence of duopoly equilibria (under some conditions) in Appendix C.3, and compare the equilibrium outcomes before and after the entry of a new platform in Appendix C.4. For convenience, we denote by NTSE non-trivial subgame equilibrium, and denote by TSE trivial subgame equilibrium, we let $S = S_b$ and $\rho = \frac{\lambda_1+\lambda_2}{S}$.

C.1 Subgame Equilibrium Analysis

In this section, we conduct the subgame equilibrium analysis. We characterize additional conditions needed to form an NTSE in Appendix C.1.1, and provide the conditions for the existence of NTSE in Appendix C.1.2. In Appendix C.1.3, we characterize the feasible region of decision variables and propose a change of variable to facilitate the analysis.

C.1.1 No Deviation Conditions

Besides the necessary conditions specified in (4.10)–(4.12) (partially covered market), or (4.13)–(4.15) (fully covered market), the existence of a nontrivial subgame equilibrium also requires that no individual worker or customer benefits from deviation. When either (4.10)–(4.12) or (4.13)–(4.15) are satisfied, it follows directly that an individual customer does not benefit from any deviation, while an individual active worker can potentially benefit from deviating to work for only one platform. Suppose a worker deviates to work

for platform 1 only. The expected workload for the deviating worker is in the range of $\left[\frac{\lambda_1}{S}, \frac{\lambda_1}{S-\lambda_2}\right]$, depending on the assignment rules. For the rest of the paper, we assume that the expected workload for a deviating worker is $\frac{\lambda_1}{S-\lambda_2}$. By doing so, we are able to construct sufficient conditions for the existence of non-trivial subgame equilibrium [1], which is given by $\frac{1}{1+\frac{S-(\lambda_1+\lambda_2)}{\lambda_1}} = \frac{\lambda_1}{S_b-\lambda_2}$. Therefore, to form a non-trivial subgame equilibrium, $\mathbf{A}|_{\mathbf{P}}$ should also satisfy

$$\frac{w_1\lambda_1 + w_2\lambda_2}{S} \geq \max\left\{\frac{w_1\lambda_1}{S-\lambda_2}, \frac{w_2\lambda_2}{S-\lambda_1}\right\} \Leftrightarrow S \leq \min\left\{Mw_1, Mw_2\right\}. \tag{C.1}$$

C.1.2 Subgame Equilibrium

In this section, we examine subgame equilibria for any given strategy profile $\mathbf{P}$ of the two platforms. Without loss of generality, we assume $p_1 \leq p_2$. Motivated by (4.1) and (4.2), we let

$$LHS(\lambda) = t\lambda + c\left(\frac{2\lambda + (p_1-p_2)/t}{\sqrt{M\{w_1\lambda + w_2[\lambda + (p_1-p_2)/t]\}}}\right) + p_1 - 1. \tag{C.2}$$

By (4.10)– (4.12), if there is an NTSE such that the demand market is not fully covered, there must exist a $\lambda' \in \Lambda = \left[\frac{p_2-p_1}{t}, \frac{t+p_2-p_1}{2t}\right]$ such that $LHS(\lambda') = 0$. Denote the values of $LHS(\lambda)$ at the two extreme points by

$$LB(\mathbf{P}) = LHS(\frac{p_2-p_1}{t}) = p_2 + c\left(\sqrt{\frac{p_2-p_1}{tMw_1}}\right) - 1, \quad \text{and} \tag{C.3}$$

$$\begin{aligned} UB(\mathbf{P}) &= LHS(\frac{t+p_2-p_1}{2t}) \\ &= \frac{t+p_2+p_1}{2} + c\left(\frac{1}{\sqrt{M[w_1(t+p_2-p_1)/(2t) + w_2(t+p_1-p_2)/(2t)]}}\right) - 1. \end{aligned} \tag{C.4}$$

By (4.13) – (4.15), if there exists an NTSE under $\mathbf{P}$ such that the demand market is fully-covered, then $UB(\mathbf{P}) \leq 0$. Moreover, an NTSE $\mathbf{A}|_{\mathbf{P}}$ must also satisfy (C.1). We

[1] Our assumption regarding the workload for a deviating worker is consistent with the workload obtained from a fluid model under the following setting: the demand for platform i arrives at constant rate λ_i; there are S workers working for both platforms; service requests for platform i are randomly assigned to idle workers who work for platform i; each service requires one unit of time. The workload of a deviating worker can thus be obtained from the Renewal Reward Theory (Ross (1996))

summarize results for NTSE in lemma C.1.1.

Lemma C.1.1. *Given a strategy profile* $\boldsymbol{P}$ *with* $p_1 \leq p_2$*, we have the following results:*

1. *when* $w_1 \geq w_2/3$*,* $LHS(\lambda)$ *increases in* λ*, and thus:*

   - *if* $LB(\mathbf{P}) \leq 0$ *and* $UB(\mathbf{P}) \geq 0$*, there exists a unique* $\lambda' \in \Lambda$ *such that* $LHS(\lambda') = 0$*. In this case, there exists an NTSE such that* $\lambda_1 = \lambda'$ *and* $\lambda_2 = \frac{t\lambda_1 + p_1 - p_2}{t}$*, given* (C.1) *is satisfied.*
   - *if* $LB(\mathbf{P}) \leq 0$ *and* $UB(\mathbf{P}) < 0$*, there exists an NTSE such that the market is fully covered, and* $\lambda_1 = \frac{t + p_2 - p_1}{2t}$ *and* $\lambda_2 = \frac{t + p_1 - p_2}{2t}$*, given* (C.1) *is satisfied.*
   - *if* $LB(\mathbf{P}) > 0$*, there does not exist an NTSE.*

2. *when* $w_1 < w_2/3$*,*

   - *if* $UB(\mathbf{P}) \leq 0$*, there exists an NTSE such that the market is fully covered, and* $\lambda_1 = \frac{t + p_2 - p_1}{2t}$ *and* $\lambda_2 = \frac{t + p_1 - p_2}{2t}$*, given* (C.1) *is satisfied.*
   - *if* $UB(\mathbf{P}) > 0$ *and* $LB(\mathbf{P}) \leq 0$*, there exists* $\lambda' \in \Lambda$ *such that* $LHS(\lambda') = 0$*. In this case, there exists an NTSE such that* $\lambda_1 = \lambda'$ *and* $\lambda_2 = \frac{t\lambda_1 + p_1 - p_2}{t}$*, given* (C.1) *is satisfied.*
   - *if* $UB(\mathbf{P}) > 0$ *and* $LB(\mathbf{P}) > 0$*, we find the minimum of* $LHS(\lambda)$ *for* $\lambda \in \Lambda$*. (i) If* $\min_{\lambda \in \Lambda} LHS(\lambda) \leq 0$*, there exists* $\lambda' \in \Lambda$ *such that* $LHS(\lambda') = 0$ *and there exists an NTSE such that* $\lambda_1 = \lambda'$ *and* $\lambda_2 = \frac{t\lambda_1 + p_1 - p_2}{t}$*, given* (C.1) *is satisfied. (ii) if* $\min_{\lambda \in \Lambda} LHS(\lambda) > 0$*, there does not exist an NTSE.*

Proof of Lemma C.1.1. We have

$$LHS'(\lambda) = t + c'\left(\frac{2\lambda + (p_1 - p_2)/t}{\sqrt{M\{w_1\lambda + w_2[\lambda + (p_1 - p_2)/t]\}}}\right)\frac{2(w_1 + w_2)\lambda + (w_1 - 3w_2)(p_2 - p_1)/t}{2\sqrt{M}\{w_1\lambda + w_2[\lambda + (p_1 - p_2)/t]\}^{\frac{3}{2}}}.$$

Because $p_1 \leq p_2$ and $\lambda \geq (p_2 - p_1)/t$, $2(w_1 + w_2)\lambda + (w_1 - 3w_2)(p_2 - p_1)/t > (3w_1 - w_2)(p_2 - p_1)/t$. Because $c'(\cdot) > 0$, if $w_1 \geq w_2/3$, then $LHS'(\lambda) > 0$ and the results in part 1 follow naturally. When $w_1 < w_2/3$, the results in part 2 are derived by enumeration. ∎

C.1.3 Preliminary Results

By Lemma C.1.1, a strategy profile $\mathbf{P} = (p_1, p_2, w_1, w_2)$ may not uniquely determine the market allocation $\mathbf{A}\,|_{\mathbf{P}} = (\lambda_1, \lambda_2, S)$. By Lemma C.1.2 below, we show that (λ_1, p_2, S, w_2) can uniquely determine (p_1, w_1, λ_2). Therefore, we shall use (λ_1, S) as the platform 1's decision variables to identify potential duopoly equilibra strategies (a change of decision variables).

Lemma C.1.2. *Gievn the platform 2's strategy (p_2, w_2) with $w_2 \leq p_2 \in (0,1)$, define $FR_{(p_2,w_2)} \coloneqq \{(\lambda_1, S) \mid \lambda_1 \in [0,1], S \in (0, M], 1-p_2-c(\frac{\lambda_1}{S}) \geq 0$ and $1-t-p_2+t\lambda_1-c(\frac{1}{S}) \leq 0\}$. Then for any $(\lambda_1, S) \in FR_{(p_2,w_2)}$, (4.10) and (4.12) uniquely determine (p_1, w_1, λ_2) with $\lambda_2 \in (0, 1-\lambda_1]$.*

Proof of Lemma C.1.2. By (4.10), let $f(\lambda_2) = 1 - p_2 - t\lambda_2 - c(\frac{\lambda_1+\lambda_2}{S})$. Observe that $f(\lambda_2)$ is decreasing in λ_2. Then $f(\lambda_2) = 0$ admits a unique solution for $\lambda_2 \in [0, 1-\lambda_1]$ if and only if $f(0) = 1 - p_2 - c(\frac{\lambda_1}{S}) \geq 0$ and $f(1-\lambda_1) = 1 - t - p_2 + t\lambda_1 - c(\frac{1}{S}) \leq 0$. We can then obtain that $p_1 = t\lambda_2 + p_2 - t\lambda_1$ by (4.10), and $w_1 = \frac{S^2 - Mw_2\lambda_2}{M\lambda_1}$ by (4.12). ∎

Note that, given the strategy (p_2, w_2) of platform 2, $FR_{(p_2,w_2)}$ defined in Lemma C.1.2 and the uniquely determined (p_1, w_1, S) cover all possible strategies of platform 1 (in terms of (p_1, w_1)) such that the resulting market allocation is either PC or KC NTSEs. In the remaining Appendices, we will frequently work on the set $FR_{(p_2,w_2)}$.

C.2 Local Duopoly Equilibra

In this section, we characterize *local (duopoly) equilibra*, which we refer to as the set of NTSEs that satisfy the KKT conditions of a relaxed version of Problem (4.4) (condition (C.1) is relaxed) for each platform. In Appendix C.3, we show that, under some conditions (i.e., M and t being sufficiently large), the local equilibra characterized in this section are indeed global equilibra. Depending on the market coverage outcome, we refer to $\mathbf{A}|_{\mathbf{P}}$ a PC (partial-coverage) NTSE if (4.10)–(4.12) hold; (ii) a KC (Kinked-Coverage) NTSE if (4.13)–(4.15) hold with (4.14) being binding; and (iii) an FC (Full-coverage) NTSE if (4.13)–(4.15) hold with (4.14) being unbinding.

Any FC NTSE can not be an equilibrium. First note that when (4.14) is unbinding, it is equivalent to $UB(\mathbf{P}) < 0$, where $UB(\mathbf{P})$ is define in (C.4). In this case, $\lambda_i = \frac{t+p_j-p_i}{2t}$ for $i \in \{1,2\}$ by (4.13), which are independent of w_i and w_j. Therefore, whenever $UB(\mathbf{P}) < 0$, given platform 2's strategy (p_2, w_2), platform 1 can always increase its profit $\lambda_1(p_1 - w_1)$ by fixing p_1 and lowering w_1.

It suffices to focus on local PC and local KC NTSEs. When (4.14) is binding, $t\lambda_i = 1 - p_i - c(\frac{1}{S})$ and $\lambda_1 + \lambda_2 = 1$. Combined with (4.10)–(4.12), it suffices to consider the following problem for platform i:

$$\begin{aligned} \max_{p_i, w_i} \quad & \lambda_i(p_i - w_i) \\ \text{subject to} \quad & \text{(4.10), (4.12) and } \lambda_1 + \lambda_2 \leq 1. \end{aligned} \tag{C.5}$$

Observe that the solution induces a local PC equilibrium if $\lambda_1 + \lambda_2 < 1$ and a local KC equilibrium otherwise. Based on the discussion in Appendix C.1.3, it is more convenience to analyze (λ_i, S) other than (p_i, w_i) given the platform j's strategy (p_j, w_j). Substituting $p_i = t\lambda_j + p_j - t\lambda_i$ and $w_i = \frac{S^2 - M w_2 \lambda_2}{M\lambda_1}$ by (4.10) and (4.12), we can reformulate the above problem as follows [2]:

$$\begin{aligned} \max_{\lambda_i, S} \quad & \Pi_i(\lambda_i, S) = t\lambda_i\lambda_j + \lambda_i p_j - t\lambda_i^2 - \frac{S^2}{M} + w_j\lambda_j, \\ \text{subject to} \quad & (\lambda_i, S) \in FR_{(p_j, w_j)}. \end{aligned} \tag{C.6}$$

Because we focus on symmetric duopoly equilibrium, it is not hard to show that most of the constraints defined in $FR_{(p_j,w_j)}$ cannot be binding under any symmetric duopoly equilibrium except the constraint that $\lambda_1 + \lambda_2 \leq 1$. Therefore, we study the following

[2] Notice that given $(\lambda_1, S) \in FR_{(p_2,w_2)}$, the unique solution (p_1, w_1, λ_2) induced by (4.10) and (4.12) does not guarantee that $0 \leq w_1 \leq p_1 \leq 1$. This does not cause any problem as the local equilibrium strategy (p^d, w^d) satisfies $0 < w^d < p^d \leq 1$ by the analysis in this section, and the set of (λ_1, S_1) such that (λ_1, p^d, S, w^d) uniquely determines (p_1, w_1, λ_2) with $0 \leq w_1 \leq p_1 \leq 1$ is a subset of $FR_{(p^d,w^d)}$ by Lemma C.1.2.

simplified problem.

$$\begin{aligned} \max_{\lambda_i, S} \quad & \Pi_i(\lambda_i, S) = t\lambda_i\lambda_j + \lambda_i p_j - t\lambda_i^2 - \frac{S^2}{M} + w_j\lambda_j, \\ \text{subject to} \quad & \lambda_1 + \lambda_2 \leq 1. \end{aligned} \tag{C.7}$$

Given the platform 2's strategy (p_2, w_2), by applying the implicit function theorem to (4.10) and (4.12) for $(\lambda_1, S_1) \in FR^{\circ}_{(p_2,w_2)}$, where $FR^{\circ}_{(p_2,w_2)}$is the interior of FR_{p_2,w_2} defined in Lemma C.1.2, we can obtain that

$$\frac{\partial \lambda_2}{\partial \lambda_1} = -\frac{c'(\rho)}{c'(\rho) + St} \quad \text{and} \quad \frac{\partial \lambda_2}{\partial S} = \frac{c'(\rho)\rho}{c'(\rho) + St}. \tag{C.8}$$

Let $\mu \geq 0$ be the KKT multiplier for the constrain $\lambda_1 + \lambda_2 \leq 1$. We can write down the KKT condition for the above problem as follows:

$$t\lambda_2 + p_2 - 2t\lambda_1 - (t\lambda_1 + w_2)\frac{c'(\rho)}{c'(\rho) + St} = \mu\frac{St}{c'(\rho) + St}, \tag{C.9}$$

$$(t\lambda_1 + w_2)\frac{c'(\rho)\rho}{c'(\rho) + St} - \frac{2S}{M} = \mu\frac{c'(\rho)\rho}{c'(\rho) + St}, \text{ and} \tag{C.10}$$

$$\mu(\lambda_1 + \lambda_2 - 1) = 0.$$

In this work, we focus on symmetric equilibra, i.e., $p_1 = p_2 = p$, $w_1 = w_2 = w$ and $\lambda_1 = \lambda_2 = \lambda$. Then, a symmetric local PC equilibrium must satisfy

$$p - t\lambda - (t\lambda + w)\frac{c'(\rho)}{c'(\rho) + St} = \mu\frac{St}{c'(\rho) + St}, \tag{C.11}$$

$$(t\lambda + w)\frac{c'(\rho)\rho}{c'(\rho) + St} - \frac{2S}{M} = \mu\frac{c'(\rho)\rho}{c'(\rho) + St}, \text{ and} \tag{C.12}$$

$$\mu(2\lambda - 1) = 0. \tag{C.13}$$

We then consider two possibilities: case (i) $\mu = 0$ and $2\lambda < 1$, which can induce local PC equilibra (Appendix C.2.1) and case (ii) $\mu \geq 0$ and $2\lambda = 1$, which induces local KC equilibra (Appendix C.2.2). In Appendix C.2.3, we highlight the role of stickiness in determining the type of equilibrium (PC or KC).

C.2.1 Local PC Equilibra

In case (i), $\mu = 0$. By (4.10) and (C.11)–(C.12), $\frac{2S}{M} = \rho(p - t\lambda) = \rho[1 - 2t\lambda - c(\rho)] = \rho[1 - t\rho S - c(\rho)]$, which implies that $S = \frac{M\rho[1-c(\rho)]}{2+Mt\rho^2}$. Moreover, because $S = Mw\rho$, (C.12) implies that $S = \frac{c'(\rho)[M\rho^2 t-2]}{4t}$. It follows that $\frac{M\rho[1-c(\rho)]}{2+Mt\rho^2} = \frac{c'(\rho)[M\rho^2 t-2]}{4t}$, which implies that

$$4 = 4c(\rho) + c'(\rho)[M\rho^3 t - \frac{4}{M\rho t}]. \tag{C.14}$$

We then show that (C.14) adopts a unique solution $\rho^d \in (0, 1)$ when M is sufficiently large. Let $RHS(\rho) = 4c(\rho) + c'(\rho)(M\rho^3 t - \frac{4}{M\rho t})$, then $RHS'(\rho) = 4c'(\rho) + c''(\rho)(M\rho^3 t - \frac{4}{M\rho t}) + c'(\rho)(3M\rho^2 t + \frac{4}{M\rho^2 t})$. When $\rho^2 > \frac{2}{Mt}$, $RHS'(\rho) > 0$ and $RHS(1) > 4$ as $c(1) \geq 1$. If $c(\sqrt{\frac{2}{Mt}}) \leq 1$, that is, $M \geq \frac{2}{t[c^{-1}(1)]^2}$, $RHS(\rho) < 4c(\rho) \leq 4$ when $\rho^2 < \frac{2}{Mt}$. Therefore, there exists a unique $\rho^d \in (0, 1)$ that satisfies (C.14) given $M > \frac{2}{t[c^{-1}(1)]^2}$. Moreover, under symmetric local PC equilibra,

$$\begin{aligned}
S^d &= \frac{M\rho^d[1 - c(\rho^d)]}{2 + Mt(\rho^d)^2} = \frac{c'(\rho^d)[M(\rho^d)^2 t - 2]}{4t}, \\
\lambda^d &= \frac{S^d\rho^d}{2} = \frac{1}{2}\frac{M(\rho^d)^2[1 - c(\rho^d)]}{2 + Mt(\rho^d)^2} = \frac{c'(\rho^d)\rho^d[M(\rho^d)^2 t - 2]}{8t}, \\
w^d &= \frac{S^d}{M\rho^d} = \frac{1 - c(\rho^d)}{2 + Mt(\rho^d)^2} = \frac{c'(\rho^d)[M(\rho^d)^2 t - 2]}{4tM\rho^d}, \quad \text{and} \\
p^d &= 1 - 2t\lambda^d - c(\rho^d) = 1 - \frac{tM(\rho^d)^2[1 - c(\rho^d)]}{2 + Mt(\rho^d)^2} - c(\rho^d) = 1 - c'(\rho^d)\rho^d[M(\rho^d)^2 t - 2]2t - c(\rho^d).
\end{aligned} \tag{C.15}$$

In Lemma C.2.1, we provide some useful results for the symmetric local PC equilibra, which are used throughout the Appendices.

Lemma C.2.1. *Under symmetric local PC equilibra, $w^d < t\lambda^d < p^d$, $p^d - t\lambda^d = 2w^d$, and $\lambda^d < \frac{1}{2t}$.*

Proof of Lemma C.2.1. By (C.11)–(C.12), we have $p^d - t\lambda^d = 2w^d$ which implies (along with (C.11)) that $w^d < t\lambda^d < p^d$. By plugging $p^d = 1 - t\lambda^d - c(\rho^d)$ into (C.11), we have $1 - 2t\lambda^d - c(\rho^d) - (t\lambda^d + w^d)\frac{c'(\rho^d)}{c'(\rho^d)+S^d} = 0$, which implies that $\lambda^d < \frac{1}{2t}$. ∎

In Lemma C.2.2, we provide the limit results on symmetric local PC equilibra. Let $C_P = \left[\frac{4}{tc'(0)}\right]^{1/3}$.

Lemma C.2.2. *Under the symmetric local PC equilibrium, we have* $\lim_{M\to\infty} \frac{\rho^d}{M^{-1/3}} = C_p$, $\lim_{M\to\infty} \frac{S^d}{M^{1/3}} = \frac{1}{tC_P}$, $\lim_{M\to\infty} \frac{w^d}{M^{-1/3}} = \frac{1}{tC_p^2}$, $\lim_{M\to\infty} p^d = \frac{1}{2}$ *and* $\lim_{M\to\infty} \lambda^d = \frac{1}{2t}$.

Proof of Lemma C.2.2. Let $LHS(\rho) = 4c(\rho) + c'(\rho)(M\rho^3 t - \frac{4}{M\rho t})$. We first show that $\lim_{M\to\infty} \rho^d = 0$. Suppose (for contradiction) that $\limsup_{M\to\infty} \rho^d > 0$, then there exists a subsequence of ρ^d indexed by M_k such that $\lim_{k\to\infty} \rho^d > 0$ (we omit the subscript of index for simplicity). Then $\lim_{k\to\infty} LHS(\rho^d) \to \infty$, which contradicts (C.14). Therefore, we must have $\lim_{M\to\infty} \rho^d = \limsup_{M\to\infty} \rho^d = \liminf_{M\to\infty} \rho^d = 0$. We then show that $\lim_{M\to\infty} M\rho^d \to \infty$. Suppose (for contradiction) that $\liminf_{M\to\infty} M\rho^d < \infty$, then there exists a subsequence of ρ^d indexed by M_k such that $\lim_{k\to\infty} M_k\rho^d < \infty$. Then $\lim_{k\to\infty} LHS(\rho^d) < 0$, which contradicts (C.14). Therefore, we must have $\lim_{M\to\infty} M\rho^d = \liminf_{M\to\infty} M\rho^d = \limsup_{M\to\infty} \to \infty$. We can then use (C.14) to obtain that $\lim_{M\to\infty} M(\rho^d)^3 = \liminf_{M\to\infty} M(\rho^d)^3 = \frac{4}{tc'(0)} = \limsup_{M\to\infty} M(\rho^d)^3 = \frac{4}{tc'(0)} = \frac{4}{tc'(0)}$, which is equivalent to $\lim_{M\to\infty} \frac{\rho^d}{M^{-1/3}} = C_P$. It follows that $\lim_{M\to\infty} \lambda^d = \frac{1}{2t}$ by (C.15); $\lim_{M\to\infty} p^d = 1 - 2t\lambda^d - c(\rho^d) = \frac{1}{2}$; $\lim_{M\to\infty} \frac{S^d}{M^{1/3}} = \lim_{M\to\infty} \frac{2\lambda^d}{\rho^d M^{1/3}} = \frac{1}{tC_P}$ and $\lim_{M\to\infty} \frac{w^d}{M^{-1/3}} = \lim_{M\to\infty} \frac{S^d}{M^{2/3}\rho^d} = \lim_{M\to\infty} \frac{S^d}{M^{1/3}} \frac{1}{\rho^d M^{1/3}} = \frac{1}{tC_P^2}$. ∎

C.2.2 Local KC Equilibra

In case (ii), $\mu \geq 0$ and $2\lambda = 1$. By (C.12) $\mu = t\lambda + w - \frac{2S}{M}\frac{c'(\rho)+St}{c'(\rho)\rho}$, and by (C.11) $p - 2t\lambda - w + \frac{2(S)^2 t}{c'(\rho)\rho M} = 0$. Under symmetric equilibra $\rho = \frac{2\lambda}{S} = \frac{1}{S}$. Because $p = 1 - \frac{t}{2} - c(\frac{1}{S})$ by (4.10), and $w = \frac{(S)^2}{M}$ by (4.12), we can obtain that $c'(\frac{1}{S})\big[1 - \frac{3t}{2} - c(\frac{1}{S}) - \frac{(S)^2}{M}\big] + \frac{2t(S)^3}{M} = 0$, which is equivalent to

$$c'(\rho)\big[1 - \frac{3t}{2} - c(\rho) - \frac{1}{M(\rho)^2}\big] + \frac{2t}{M(\rho)^3} = 0. \tag{C.16}$$

We then show that for $t < 1$, there exists a unique $\rho^d \in (\rho_{r1}, \rho_{r2})$ such that (C.16) is satisfied when M is sufficiently large, where $\rho_{r1} < \rho_{r2}$ are the two roots for the equation $0 = 1 - \frac{t}{2} - c(\rho) - \frac{1}{M\rho^2} = RHS(\rho)$[3]. Let $LHS(\rho) = c'(\rho)\big[1 - \frac{3t}{2} - c(\rho) - \frac{1}{M\rho^2}\big] + \frac{2t}{M\rho^3}$. When

[3]We focus on $\rho^d \in (\rho_{r1}, \rho_{r2})$ as it is equivalent to $RHS(\rho^d) \geq 0$, which is further equivalent to $p^d \geq w^d$ by (4.10)–(4.12). When M is sufficiently large, the equation $RHS(\rho) = 0$ has exactly two roots (note that $RHS(\rho)$ is a concave function).

M is sufficiently large such that ρ_{r1} and ρ_{r2} exist, due to the concavity of $RHS(\rho)$, we have the following observations: (1) $RHS'(\rho_{r1}) = -c'(\rho_{r1}) + \frac{2}{M\rho_{r1}^3} > 0$, which implies that $LHS(\rho_{r1}) > c'(\rho_{r1})[1 - \frac{t}{2} - c(\rho_{r1}) - \frac{1}{M\rho_{r1}^2}] = 0$; and (2) $RHS'(\rho_{r2}) = -c'(\rho_{r2}) + \frac{2}{M\rho_{r2}^3} < 0$, which implies that $LHS(\rho_{r2}) < c'(\rho_{r2})[1 - \frac{t}{2} - c(\rho_{r2}) - \frac{1}{M\rho_{r2}^2}] = 0$. Therefore, there exists a $\rho^d \in (\rho_{r1}, \rho_{r2})$ such that $LHS(\rho^d) = 0$. We then show that ρ^d is the unique solution. Suppose (for contradiction) that there exists $\tilde{\rho}^d \in (\rho^d, \rho_{r2})$ such that $LHS(\tilde{\rho}^d) = 0$ and $LHS(\rho) < 0$ for $\rho \in (\rho^d, \tilde{\rho}^d)$. Then we must have (i) $LHS'(\tilde{\rho}^d) = c''(\tilde{\rho}^d)[1 - \frac{3t}{2} - c(\tilde{\rho}^d) - \frac{1}{M(\tilde{\rho}^d)^2}] + c'(\tilde{\rho}^d)[-c'(\tilde{\rho}^d) + \frac{2}{M(\tilde{\rho}^d)^3}] - \frac{6t}{M(\tilde{\rho}^d)^4} > 0$ (as $LHS(\rho^d) = 0$, $LHS(\tilde{\rho}^d) = 0$, and $LHS(\rho) < 0$ for $\rho \in (\rho^d, \tilde{\rho}^d)$), and (ii) $c'(\rho)\rho^3 > \frac{2}{M}$ and $1 - \frac{3t}{2} - c(\rho) - \frac{1}{M\rho^2} < 0$ for $\rho \in [\rho^d, \tilde{\rho}^d]$ (as $RHS(\rho) > 0$ for $\rho \in [\rho^d, \tilde{\rho}^d] \in (\rho_{r1}, \rho_{r2})$). Observe that these two conditions contradict to each other, and the desired result follows.

In Lemma C.2.3, we provide limit results on the symmetric KC equilibra. Let $C_K = \left[\frac{2t}{(\frac{3}{2}t-1)c'(0)}\right]^{1/3}$.

Lemma C.2.3. *For the symmetric local KC equilibrium, we have the following results:*

(i) when $t \in (\frac{2}{3}, 1)$, $\lim_{M\to\infty} \frac{\rho^d}{M^{-1/3}} = C_K$, $\lim_{M\to\infty} \frac{S^d}{M^{1/3}} = \frac{1}{C_K}$, $\lim_{M\to\infty} \frac{w^d}{M^{-1/3}} = \frac{1}{C_K^2}$, $\lim_{M\to\infty} p^d = 1 - \frac{t}{2}$ *and* $\lambda^d = \frac{1}{2}$*; and*

(ii) when $t < \frac{2}{3}$, $\lim_{M\to\infty} S^d = \frac{1}{c^{-1}(1-\frac{3}{2}t)}$, $\lim_{M\to\infty} \rho^d = c^{-1}(1-\frac{3}{2}t)$, $\lim_{M\to\infty} \frac{w^d}{M^{-1}} = \left[\frac{1}{c^{-1}(1-\frac{3}{2}t)}\right]^2$, $\lim_{M\to\infty} p^d = t$ *and* $\lambda^d = \frac{1}{2}$.

Proof of Lemma C.2.3. We prove Lemma C.2.3 by considering the following two cases separately: case (i) $t \in (\frac{2}{3}, 1)$, and case (ii) $t \in (0, \frac{2}{3})$. Let $LHS(\rho^d) = c'(\rho^d)\big[1 - \frac{3t}{2} - c(\rho^d) - \frac{1}{M(\rho^d)^2}\big] + \frac{2t}{M(\rho^d)^3}$.

Case (i). We first show that $\lim_{M\to\infty} \rho^d = 0$. Suppose (for contradiction) that $\limsup_{M\to\infty} \rho^d > 0$. Then there exists a subsequence of ρ^d indexed by M_k such that $\lim_{k\to\infty} \rho^d > 0$. Then $\lim_{k\to\infty} LHS(\rho^d) < 0$, which contradicts (C.16). Therefore, we must have $\lim_{M\to\infty} \rho^d = \liminf_{M\to\infty} \rho^d = \limsup_{M\to\infty} \rho^d = 0$. We then show that $\lim_{M\to\infty} M\rho^d \to \infty$. Suppose (for contradiction) that $\liminf_{M\to\infty} M\rho^d < \infty$. Then there exists a subsequence of ρ^d indexed by M_k such that $\lim_{k\to\infty} M_k\rho^d < \infty$. Then $\liminf_{k\to\infty} LHS(\rho^d) = \infty$, contradicting (C.16). Therefore, we must have $\lim_{M\to\infty} M\rho^d = \liminf_{M\to\infty} M\rho^d = \limsup_{M\to\infty} M\rho^d \to \infty$. By the same argument, $\lim_{M\to\infty} M(\rho^d)^2 \to \infty$. We can then use (C.16) to obtain that $\lim_{M\to\infty} M(\rho^d)^3 = \liminf_{M\to\infty} M(\rho^d)^3 = \limsup_{M\to\infty} M(\rho^d)^3 =$

C_K^3, which is equivalent to $\limsup_{M\to\infty} \frac{\rho^d}{M^{-1/3}} = C_K$. It follows that $\lim_{M\to\infty} \frac{S^d}{M^{1/3}} = \lim_{M\to\infty} \frac{1}{\rho^d/M^{-1/3}} = \frac{1}{C_K}$; $\lim_{M\to\infty} \frac{w^d}{M^{-1/3}} = \lim_{M\to\infty} \frac{(S^d)^2}{M^{2/3}} = \frac{1}{C_K^2}$, $\lim_{M\to\infty} p^d = \lim_{M\to\infty} [1 - t\lambda^d - c(\rho^d)] = 1 - \frac{t}{2}$.

Case (ii). We first show that $\lim_{M\to\infty} M(\rho^d)^2 \to \infty$. Suppose (for contradiction) that $\liminf_{M\to\infty} M(\rho^d)^2 < \infty$. Then there exists a subsequence of ρ^d indexed by M_k such that $\lim_{k\to\infty} M_k(\rho^d)^2 < \infty$. Then $\lim_{k\to\infty} M_k(\rho^d)^3 = 0$ and $\lim_{k\to\infty} \rho^d = 0$, which contradicts (C.16). Therefore, we must have $\lim_{M\to\infty} M(\rho^d)^2 = \liminf_{M\to\infty} M(\rho^d)^2 = \limsup_{M\to\infty} M(\rho^d)^2 \to \infty$. We then show that $\lim_{M\to\infty} M(\rho^d)^3 \to \infty$. Suppose (for contradiction) that $\liminf_{M\to\infty} M(\rho^d)^3 < \infty$. Then there exists a subsequence of ρ^d indexed by M_k such that $\lim_{k\to\infty} M_k(\rho^d)^3 < \infty$. Then $\lim_{k\to\infty} = c'(0)[1 - \frac{3t}{2}] + \frac{2t}{M_k(\rho^d)^3} > 0$ as $t < \frac{2}{3}$, which contradicts (C.16). Therefore, we must have $\lim_{M\to\infty} M(\rho^d)^3 = \liminf_{M\to\infty} M(\rho^d)^3 = \limsup_{M\to\infty} M(\rho^d)^3 \to \infty$. We can then use (C.16) to obtain that $\lim_{M\to\infty} \rho^d = \liminf_{M\to\infty} \rho^d = \limsup_{M\to\infty} \rho^d = c^{-1}(1 - \frac{3t}{2})$. It follows that $\lim_{M\to\infty} S^d = \frac{1}{\rho^d} = \lim_{M\to\infty} \frac{1}{c^{-1}(1-\frac{3t}{2})}$, $\lim_{M\to\infty} \frac{w^d}{M^{-1}} = \lim_{M^\to\infty} (S^d)^2 = \left[\frac{1}{c^{-1}(1-\frac{3t}{2})}\right]^2$, and $\lim_{M\to\infty} p^d = \lim_{M\to\infty} 1 - \frac{t}{2} - c(\rho^d) = t$. ∎

C.2.3 Role of Stickiness

Finally, we note that when M is sufficiently large, the local PC equilibrium cannot be a global equilibrium when $t < 1$; and a local KC equilibrium cannot be a global equilibrium when $t > 1$. We first note that when $t < 1$, a local PC equilibrium can not be a global equilibrium as $\lim_{M\to\infty} \lambda^d = \frac{1}{2t}$ by Lemma C.2.2, contradicting to the fact that $\lambda_1 + \lambda_2 \leq 1$. We then show that a local KC equilibrium can not be a global equilibrium when $t > 1$. By a similar analysis to that in Appendix C.2.2, we can obtain that (i) when $t \geq 2$, (C.16) does not admit a solution ρ^d such that $p^d \geq w^d$, and thus a symmetric KC equilibrium can not be formed; and (ii) when $t \in (1, 2)$, the solution to (C.16) shares the same form as that for the case when $t \in (\frac{2}{3}, 1)$. We then show the existence of a profitable deviation for platform 1 for $t \in (1, 2)$, given platform 2 adopts the local KC equilibrium strategy (p^d, w^d). Consider the strategy $(\lambda_1, S_1) = (\frac{1}{2} - \epsilon, S^d)$ for platform 1, where $\epsilon > 0$ is sufficiently small. Then $\lim_{M\to\infty} \frac{\lambda_1+\lambda_2}{S^d} = 0$ by Lemma C.2.3. By (4.12) and Lemma C.2.3, $\lim_{M\to\infty} w_1 = \frac{(S^d)^2 - Mw^d\lambda_2}{M(\frac{1}{2}-\delta)} = 0$. By (4.10), $\lim_{M\to\infty} p_1 = 1 - t(\frac{1}{2} - \epsilon) - c(\frac{\lambda_1+\lambda_2}{S})$. It follows that $\lim_{M\to\infty} \lambda_1(p_1 - w_1) = (\frac{1}{2} - \epsilon)[1 - t(\frac{1}{2} - \epsilon)] > \frac{t}{2}(1 - \frac{t}{2}) = \lim_{M\to\infty} \Pi_1(\lambda^d, S^d)$ when $t > 1$

for sufficiently small $\epsilon > 0$. Therefore, $(\lambda^d - \epsilon, S^d)$ is a profitable deviation for platform 1 when M is sufficiently large.

C.3 Global Equilibra

In this section, we show that when M is sufficiently large: (1) the local PC equilibrium is an equilibrium if $t > 1$; (2) the local KC equilibra is an equilibrium if $t \in (\frac{2}{3}, 1)$; and (3) there does not exist a symmetric equilibrium if $t \in (0, \frac{2}{3})$. Specifically, we let platform 2 adopts the local PC equilibrium strategy for $t > 1$ and the local KC equilibrium strategy for $t < 1$, and we consider every possible deviating strategy (p_1, w_1) for platform 1. Throughout this section, let (p^d, w^d) denote the local PC equilibrium strategy for $t > 1$ and local KC equilibrium strategy for $t < 1$. If there exists an NTSE under $\mathbf{P} = (p_1, p^d, w_1, w^d)$, we refer to such (p_1, w_1) as a small deviation; otherwise, we refer to it as a large deviation. We consider large deviations in Section C.3.1, and small deviations in Section C.3.2.

C.3.1 Large Deviations

By the refinement rule introduced in Section 4.4, when platform 2 adopts the local equilibrium strategy (p^d, w^d), (p_1, w_1) is a profitable large deviation only if it satisfies the following *Profitable Large Deviation Conditions*.

- Condition I: the strategy profile $\mathbf{P} = (p_1, p^d, w_1, w^d)$ does not admit an NTSE.
- Condition II: the worker welfare under the TSE associated with platform 1 is greater than that under the TSE associated with platform 2. Specifically, let $\hat{\lambda}_i$ be the unique solution to
$$\lambda = \frac{1}{t}\left[1 - p_i - c\left(\sqrt{\frac{\lambda}{M w_i}}\right)\right], \tag{C.17}$$
and let $\hat{S}_i = \sqrt{M w_i \hat{\lambda}_i}$ denote the customer arrival rate and service supply respectively under the TSE associated with platform i. Let $\hat{\rho}_i = \frac{\hat{\lambda}_i}{\hat{S}_i}$. This condition implies that $w_1\hat{\rho}_1 > w_2\hat{\rho}_2$, which is equivalent to $w_1\hat{\lambda}_1 > w_2\hat{\lambda}_2$.
- Condition III: the platform 1's profit under the TSE associated with platform 1 is higher than that under the local duopoly equilibrium, i.e., $\hat{\lambda}_1(p_1 - w_1) > \lambda^d(p^d - w^d)$.

Because the outcomes of local equilibria under different values of t are different, we provide proofs for different values of t separately. We first provide the limit result of $\hat{\lambda}_2$ per Lemma C.3.1 below.

Lemma C.3.1. *Given that platform 2 adopts (p^d, w^d), which is the local PC equilibrium strategy for $t > 1$, and the local KC equilibrium strategy for $t \in (\frac{2}{3}, 1)$, $\lim_{M\to\infty} \hat{\lambda}_2 = \frac{1}{2t}$.*

Proof of Lemma C.3.1. The result follows directly from the definition of $\hat{\lambda}_2$ in (C.17), Lemma C.2.2 and C.2.3. ∎

The Case in Which $t > 1$

In this section, we show that when $t > 1$ and M is sufficiently large, there does not exist a profitable large deviation. We do so by showing that there does not exist a (p_1, w_1) that satisfies the Profitable Large Deviation Conditions, given that platform 2 adopts the local PC equilibrium strategy (p^d, w^d). Specifically, we show that any (p_1, w_1) which satisfies Condition II and III does not satisfy Condition I of the Profitable Large Deviation Conditions when M is sufficiently large. Recall from Lemma C.1.1, (p_1, w_1) is a small deviation if (p_1, w_1) satisfies $LB(\mathbf{P}) < 0$ and $S \leq \min\{Mw_1, Mw^d\}$, where the second condition comes from (C.1). Note that $LB(\mathbf{P}) = p^d + c(\sqrt{\frac{p^d - p_1}{tMw_1}}) - 1$ if $p_1 < p^d$ and $LB(\mathbf{P}) = p_1 + c(\sqrt{\frac{p_1 - p^d}{tMw^d}}) - 1$ otherwise. For simplicity, we use $LB(p_1, w_1)$ to denote $LB(\mathbf{P})$ (as $(p_2, w_2) = (p^d, w^d)$ is fixed). In what follows, we examine two cases: (i) $\hat{\lambda}_1 \geq \hat{\lambda}_2$ and (ii) $\hat{\lambda}_1 < \hat{\lambda}_2$, where $\hat{\lambda}_i$ is define in (C.17).

Case (i): $\hat{\lambda}_1 \geq \hat{\lambda}_2$.

Step (1): (p_1, w_1) satisfies $LB(p_1, w_1) < 0$. We prove this by considering scenario (i.i) $p_1 \leq p^d$ and scenario (i.ii) $p_1 > p^d$.

Scenario (i.i) $p_1 \leq p^d$. In this case, we need to show that $LB(p_1, w_1) = p^d + c(\sqrt{\frac{p^d - p_1}{tMw_1}}) - 1 < 0$. Observe that $LB(p_1, w_1)$ decreases in p_1 and w_1, Let p_{lb} and w_{lb} be some lower bounds (which we shall specify later) for p_1 and w_1 respectively. It suffices to show that $LB(p_{lb}, w_{lb}) < 0$. By Condition III, we have $p_1 \geq \hat{\lambda}_1(p_1 - w_1) > \lambda^d(p^d - w^d) = \pi^d$. So we let $p_{lb} = \pi^d$. Recall that $\hat{\lambda}_i$ is the unique solution to (C.17). Then $\hat{\lambda}_1 \geq \hat{\lambda}_2$ and $p_1 > \pi^d$ implies that the unique solution (in terms of w) to $t\hat{\lambda}_2 + c(\sqrt{\frac{\hat{\lambda}_2}{Mw}}) + \pi^d = 1$ is a lower bound for w_1, which we denote by w_{lb}. Hence $LB(p_{lb}, w_{lb}) = p^d + c(\sqrt{\frac{p^d - \pi^d}{tMw_{lb}}}) - 1 =$

$p^d-\pi^d+c(\sqrt{\frac{p^d-\pi^d}{tMw_{lb}}})+\pi^d-1$. To prove $LB(p_{lb},w_{lb})<0$, it suffices to show that $p^d-\pi^d<t\hat{\lambda}_2$ as $\hat{\lambda}_2$ satisfies $t\hat{\lambda}_2+c(\sqrt{\frac{\hat{\lambda}_2}{Mw_{lb}}})+\pi^d=1$. Recall from Lemma C.3.1 that $\lim_{M\to\infty}\hat{\lambda}_2=\frac{1}{2t}$. Then by Lemma C.2.2, $\lim_{M\to\infty}(\pi^d+t\hat{\lambda}_2-p^d)=\frac{1}{4t}>0$. Therefore, $LB(p_1,w_1)<0$ when M is sufficiently large.

Scenario (i.ii) $p_1>p^d$**.** In this case, we need to prove $LB(p_1,w_1)=p_1+c(\sqrt{\frac{p_1-p^d}{tMw^d}})-1<0$. Observe that $LB(p_1,w_1)$ is independent of w_1 and increasing in p_1. For simplicity, let $LB(p_1)$ denote $LB(p_1,w_1)$. Then it suffices to show that $LB(p_{ub})<0$ for some $p_{ub}\geq p_1$ which we shall specify later. Note that $t\hat{\lambda}_1+c(\sqrt{\frac{\hat{\lambda}_1}{Mw_1}})+p_1=1$ implies $p_1<1-t\hat{\lambda}_1\leq 1-t\hat{\lambda}_2$ as $\hat{\lambda}_1\geq\hat{\lambda}_2$. Thus, we let $p_{ub}=1-t\hat{\lambda}_2$. Then it suffices to show that $LB(p_{ub})=c(\sqrt{\frac{1-t\hat{\lambda}_2-p^d}{tMw^d}})-t\hat{\lambda}_2<0$ when M is sufficiently large, which is true as $\lim_{M\to\infty}[c(\sqrt{\frac{1-\hat{\lambda}_2-p^d}{tMw^d}})-t\hat{\lambda}_2]=-\frac{1}{2}<0$ by Lemma C.2.2 and Lemma C.3.1. Hence, we have $LB(p_1)<0$ when M is sufficiently large.

Step (2): (p_1,w_1) **satisfies** $S<\min\{Mw_1,Mw^d\}$**.** We prove this by examining two scenarios (a) $w_1\geq w^d$ and (b) $w_1<w^d$.

Scenario (ii.i) $w_1\geq w^d$**.** In this case, it suffices to show that $S<Mw^d$, which is equivalent to $w_1\lambda_1+w^d\lambda_2<M(w^d)^2$. Since $w_1\lambda_1+w^d\lambda_2<w_1<1$, it suffices to show that $1<M(w^d)^2$ when M is sufficiently large. This holds as $\lim_{M\to+\infty}M(w^d)^2=\infty$ by Lemma C.2.2.

Scenario (ii.ii) $w_1<w^d$**.** In this case, it suffices to show that $S<Mw_1$, which is equivalent to $w_1\lambda_1+w^d\lambda_2<Mw_1^2$. In what follows, we show that $w^d<Mw_1^2$. By Condition II, we have $w_1>\frac{w^d\hat{\lambda}_2}{\hat{\lambda}_1}$. As $t\hat{\lambda}_1<1-p_1<1-\pi^d$ and $\hat{\lambda}_2>\lambda^d$ (as λ^d is the unique solution to $t\lambda+p^d+c(\sqrt{\frac{2\lambda}{Mw^d}})=1$), we have $w_1>\frac{w^d\hat{\lambda}_2}{\hat{\lambda}_1}>\frac{w^d\lambda^d t}{1-\pi^d}$. Then, it suffices to show that $M\frac{w^d(\lambda^d)^2t^2}{(1-\pi^d)^2}>1$, which holds when M is sufficiently large as $\lim_{M\to\infty}M\frac{w^d(\lambda^d)^2t^2}{(1-\pi^d)^2}=\infty$ by Lemma C.2.2.

Case (ii): $\hat{\lambda}_1<\hat{\lambda}_2$. In this case, we must have $w_1>w^d$ by Condition II and $p_1>p^d$ by Condition III. The proof of $LB(p_1,w_1)<0$ and that of $S<Mw^d$ are similar to Scenario (i.ii) and (ii.i) respectively. We omit the details for simplicity.

The Case in Which $t\in(\frac{2}{3},1)$

The analysis is the same as that for the case where $t>1$. We omit the details for simplicity.

The Case in Which $t \in (0, \frac{2}{3})$

In this section, we show that a profitable large deviation exists when $t \in (0, \frac{2}{3})$ and M is sufficiently large, given that platform 2 adopts the local KC equilibrium strategy (p^d, w^d) (for the case where $t \in (0, \frac{2}{3})$). Let $p_1 = p^d - \theta t$ and $w_1 = w^d + \alpha t$, where $\alpha, \theta \in (0, 1)$ are some constants. We require $p_1 - w_1 > 0$, which implies $p^d - w^d > (\theta + \alpha)t$. In what follows, we show that there exists a pair of (θ, α) such that (p_1, w_1) satisfies the Profitable Large Deviation Conditions.

First, we note that $UB(p^d, p^d, w^d, w^d) = 0$. Because $p_1 < p^d$ and $w_1 > w^d$, $UB(p_1, p^d, w_1, w^d) < 0$. Therefore, we have (1) $\lambda_1 + \lambda_2 = 1$, and (2) $\hat{\lambda}_1 \geq \lambda_1 > \lambda^d = \frac{1}{2}$, where $\hat{\lambda}_1$ is defined in (C.17). In what follows, we consider each condition separately.

Condition I. By Lemma C.1.1, if $S > Mw^d$, then $\mathbf{P} = (p_1, p^d, w_1, w^d)$ does not admit an NTSE and thus Condition I holds. Note that $S > Mw^d$ is equivalent to $w_1\lambda_1 + w^d\lambda_2 > M(w^d)^2$, which is equivalent to $\frac{\alpha(1+\theta)t}{2} > (Mw^d - 1)w^d$. Because $\lim_{M\to\infty} \frac{w^d}{M^{-1}} = \left[\frac{1}{c^{-1}(1-\frac{3}{2}t)}\right]^2$, the desired condition holds as α and θ are constants.

Condition II. Because $w_1 > w^d$ and $p_1 < p^d$, we have $\hat{\lambda}_1 > \hat{\lambda}_2$, and thus $\hat{\lambda}_1 w_1 > \hat{\lambda}_2 w^d$ holds.

Condition III. Because $t\hat{\lambda}_1 = 1 - p_1 - c(\sqrt{\frac{\hat{\lambda}_1}{Mw_1}}) = 1 - (p^d - t\theta) - c(\sqrt{\frac{\hat{\lambda}_1}{M(w^d+\alpha t)}})$. We have $\lim_{M\to\infty} \hat{\lambda}_1 = \lim_{M\to\infty} \frac{1-p^d+\theta t}{t} = \frac{1-t+\theta t}{t}$. Then $\lim_{M\to\infty} [\hat{\lambda}_1(p_1 - w_1) - \lambda^d(p^d - w^d)] > 0$ implies that

$$2(1 - t + \theta t)(1 - \theta - \alpha) > t. \tag{C.18}$$

It is not difficult to find two constants $\alpha, \theta \in (0, 1)$ such that (C.18) holds. Therefore, a profitable large deviation exists.

C.3.2 No Profitable Small Deviations

In this section, we show that a profitable small deviation for platform 1 does not exist given platform 2 adopts (p^d, w^d), which is the local PC equilibrium strategy for $t > 1$ and the local KC equilibrium strategy for $t \in (\frac{2}{3}, 1)$. Recall that we define $FR_{(p_2,w_2)}$ in Lemma C.1.2, which covers all possible strategy of platform 1 such that $\mathbf{P} = (p_1, p^d, w_1, w^d)$ admits either a PC or KC market allocation. In addition, we show in the beginning of

Appendix C.2 that any FC NTSE can not be an equilibrium. Therefore, it suffice to consider $(\lambda_1, S_1) \in FR_{(p^d,w^d)}$.

The Case in Which $t > 1$

In this section, we assume platform 2 adopts (p^d, w^d), which is the local PC equilibrium strategy. We show that platform 1 has no incentive to deviate from (λ^d, S^d) to any $(\lambda_1, S_1) \in FR_{(p^d,w^d)}$. The proof consists of the following 3 steps.

Step (1). Existence of an optimal strategy. As shown in Figure C.1, $FR_{(p^d,w^d)}$ is the shaded region enclosed by L_1, L_2, L_3, and $\lambda_1 = 0$ (except the point $(0,0)$), where $L_1 : \lambda_1 = c^{-1}(1-p^d)S$, $L_2 : \lambda_1 = \frac{1}{t}[p^d + c(\frac{1}{S}) + (t-1)]$ and $L_3 : S = M$ are characterized by the definition of $FR_{(p^d,w^d)}$. Moreover, we denote the line $\lambda_1 = c^{-1}(1-p^d-t\lambda^d) - \lambda^d$ by L. Note that $PS = (\lambda^d, S^d)$ lies on L. Let $P_1 = (1, \frac{1}{c^{-1}(1-p^d)})$ be the point where L_1 and L_2 intercept; $P_2 = (\frac{1}{t}[p^d + c(\frac{1}{M}) + (t-1)], M)$ be the point where L_2 and L_3 intercept.

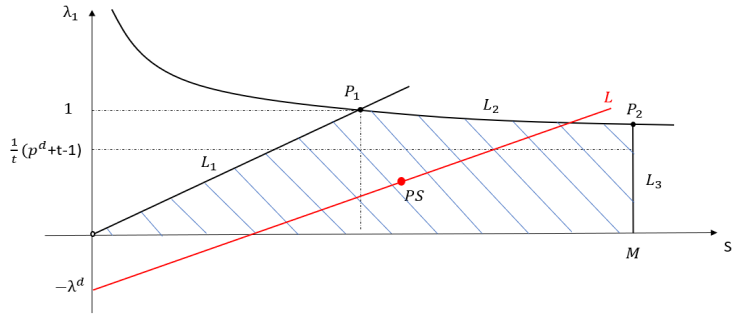


Figure C.1: Illustration of $FR_{(p^d,w^d)}$

Notice that $FR_{(p^d,w^d)}$ is not a compact set. Define $\Omega := \left\{(\lambda_1, S) \in FR \middle| \sqrt{\lambda_1^2 + S^2} < \epsilon\right\}$. Give $\delta > 0$, let $\epsilon < \min\left\{\frac{\delta}{2}, \sqrt{\frac{M}{2}\delta}\right\}$. Therefore, for any $(\lambda_1, S) \in \Omega$, $\lambda_1(p_1 - w_1) \leq \lambda_1 p_1 + |\lambda_1 w_1| = \lambda_1 p_1 + \left|\frac{S^2}{M} - w^d\lambda_2\right| \leq \epsilon + \left|\frac{S^2}{M}\right| + |w^d\lambda_2| \leq \epsilon + \frac{\epsilon^2}{M} + w^d\lambda_2 \leq \delta + w^d\lambda_2$. By (C.8), λ_2 decreases in λ_1. Hence for any point that lies above L, the corresponding $\lambda_2 \leq \lambda^d$. It follows that $\lambda_2 w^d < \lambda^d w^d \leq \lambda^d(p^d - w^d)$, where the last inequality is due to lemma C.2.1. Therefore, there exists a sufficiently small $\delta > 0$ such that $\lambda_1(p_1 - w_1) \leq \lambda^d(p^d - w^d)$ for $(\lambda_1, S_1) \in \Omega$. Then because $FR_{(p^d,w^d)} \backslash \Omega$ is a compact set and $\Pi_1(\lambda_1, S)$ defined in (C.7)

is continuous, by the Extreme Value Theorem (Rudin (1976)), there exists (λ_1^*, S_1^*) that maximizes $\Pi_1(\lambda_1, S_1)$.

Step (2). $(\lambda^{\mathbf{d}}, \mathbf{S^d})$ is the unique interior local maximum point in $\mathbf{FR_{(p^d, w^d)}}$. We first show that (λ^d, S^d) is a local maximum point. For convenience, let $A = c'(\rho)$ and $B = c''(\rho)$. By (C.8), the Hessian matrix of $\Pi_1(\lambda_1, S)$ is given by $H = \begin{pmatrix} \frac{\partial^2 \Pi_1(\lambda_1,S)}{\partial \lambda_1^2} & \frac{\partial^2 \Pi_1(\lambda_1,S)}{\partial \lambda_1 \partial S} \\ \frac{\partial^2 \Pi_1(\lambda_1,S)}{\partial S \partial \lambda_1} & \frac{\partial^2 \Pi_1(\lambda_1,S)}{\partial S^2} \end{pmatrix}$, where

$$\frac{\partial^2 \Pi_1(\lambda_1, S)}{\partial \lambda_1^2} = -2t - \frac{2At}{A+St} - \frac{Bt(1-\frac{A}{A+St})(\lambda_1 t + w^d)}{(A+St)^2} < 0,$$

$$\frac{\partial^2 \Pi_1(\lambda_1, S)}{\partial \lambda_1 \partial S} = \frac{\partial^2 \Pi_1(\lambda_1, S)}{\partial S \partial \lambda_1}$$

$$= \frac{t[A^3\rho + B\rho St(\lambda_1 t + w^d) + ASt(\lambda_1 t + \rho St + w^d) + A^2(\lambda_1 t + 2\rho St + w^d)]}{(A+St)^3} > 0, \text{ and}$$

$$\frac{\partial^2 \Pi_1(\lambda_1, S)}{\partial S^2} = -\frac{2}{M} - \frac{[2A^2 S(\lambda_1 + \lambda_2)t + S^2(\lambda_1 + \lambda_2)(2A + B\rho)t^2](\lambda_1 t + w^d)}{S^2(A+St)^3} < 0.$$

Therefore, $\Pi_1(\lambda_1, S)$ is supermodular and component wise concave. When platform 2 adopts (p^d, w^d), (λ^d, S^d) satisfies (C.9)–(C.10) with $\mu = 0$. It follows that (λ^d, S^d) is a local maximum point.

We then show that any $(\lambda_1, S) \in FR^{\circ}_{(p^d, w^d)}$ and $(\lambda_1, S) \neq (\lambda^d, S^d)$ is not a local maximum point, where $FR^{\circ}_{(p^d, w^d)}$ denote the interior of $FR_{(p^d, w^d)}$. By supermodularity, any (λ_1, S) with $(\lambda_1 - \lambda^d)(S - S^d) \leq 0$ does not satisfy (C.9)–(C.10). We then prove that any (λ_1, S) with $(\lambda_1 - \lambda^d)(S - S^d) > 0$ does not satisfy (C.9)–(C.10) either. Suppose for contradiction that there exists such a (λ_1, S), we consider the following 4 cases.

Case (i) $\lambda_1 > \lambda^d$, $S > S^d$ and $\rho < \rho^d$. By (C.8), $\lambda_2 > \lambda^d$. By (C.10), $S > S^d$ implies $(t\lambda_1 + w^d)\frac{c'(\rho)}{c'(\rho)+St} > (t\lambda^d + w^d)\frac{c'(\rho^d)}{c'(\rho^d)+S^d t}$, which together with (C.9) implies that $t\lambda_2 + p^d - 2t\lambda_1 > t\lambda^d + p^d - 2t\lambda^d \Leftrightarrow \lambda_2 > 2\lambda_1 - \lambda^d$. It follows that $\lambda_2 > \lambda_1$ as $\lambda_1 > \lambda^d$. By (C.10), (λ_1, S) and (λ^d, S^d) satisfy respectively $\frac{t\lambda^d + w^d}{S^d t} = \frac{2}{M}\frac{c'(\rho^d)+S^d t}{c'(\rho^d)}\frac{1}{\rho^d}$ and $\frac{t\lambda_1 + w^d}{St} = \frac{2}{M}\frac{c'(\rho)+St}{c'(\rho)}\frac{1}{\rho}$. Because $\rho < \rho^d$, and $S > S^d$, we have $\frac{c'(\rho)+St}{c'(\rho)} > \frac{c'(\rho^d)+S^d t}{c'(\rho^d)} \Rightarrow \frac{t\lambda_1 + w^d}{S} > \frac{t\lambda^d + w^d}{S^d} \Leftrightarrow \frac{t\lambda_1 + w^d}{t\lambda^d + w^d} > \frac{S}{S^d}$, and $\frac{S}{S^d} > \frac{\lambda_1 + \lambda_2}{2\lambda^d}$. It follows that $\frac{t\lambda_1 + w^d}{t\lambda^d + w^d} > \frac{\lambda_1 + \lambda_2}{2\lambda^d} \Leftrightarrow t\lambda^d(\lambda_1 - \lambda_2) + w^d(2\lambda^d - \lambda_1 - \lambda_2) > 0$, which contradicts $\lambda_2 > \lambda_1 > \lambda^d$ as $t\lambda^d \geq w^d$ by Lemma C.2.1.

Case (ii) $\lambda_1 > \lambda^d$, $S > S^d$ and $\rho \geq \rho^d$. As $\rho \geq \rho^d$, we have $\lambda_2 \leq \lambda^d < \lambda_1$ which implies

that $t\lambda^d + p^d - 2t\lambda^d > t\lambda_2 + p^d - 2t\lambda_1$. Then by (C.9), $\frac{t\lambda^d+w^d}{t\lambda_1+w^d} > \frac{c'(\rho)}{c'(\rho)+St}/\frac{c'(\rho^d)}{c'(\rho^d)+S^d t}$. By Lemma C.2.1, $2\lambda^d(t\lambda_1 + w^d) - (\lambda_1+\lambda_2)(t\lambda^d + w^d) = t\lambda^d(\lambda_1 - \lambda_2) + w^d(2\lambda^d - \lambda_1 - \lambda_2) \geq w^d(2\lambda^d - 2\lambda_2) \geq 0$, which is equivalent to $\frac{t\lambda_1+w^d}{t\lambda^d+w^d} \geq \frac{\lambda_1+\lambda_2}{2\lambda^d}$. Because $c'(\rho) \geq c'(\rho^d)$, we have $\frac{c'(\rho)}{c'(\rho)+St} \geq \frac{c'(\rho^d)}{c'(\rho^d)+St}$. It follows that $\frac{c'(\rho^d)}{c'(\rho^d)+S^d t}/\frac{c'(\rho)}{c'(\rho)+St} \leq \frac{c'(\rho^d)}{c'(\rho^d)+S^d t}/\frac{c'(\rho^d)}{c'(\rho^d)+St} = \frac{c'(\rho^d)+St}{c'(\rho^d)+S^d t} < \frac{S}{S^d}$, where the last inequality is due to $S > S^d$. Therefore, $\frac{\lambda_1+\lambda_2}{2\lambda^d} \leq \frac{t\lambda_1+w^d}{t\lambda^d+w^d} < \frac{c'(\rho^d)}{c'(\rho^d)+S^d t}/\frac{c'(\rho)}{c'(\rho)+St} < \frac{S}{S^d} \Rightarrow \rho < \rho^d$, which leads to a contradiction.

Case (iii) $\lambda_1 < \lambda^d$, $S < S^d$ and $\rho < \rho^d$. The proof is similar to that of Case (ii).

Case (iv) $\lambda_1 < \lambda^d$, $S < S^d$ and $\rho \geq \rho^d$. The proof is similar to that of Case (i).

Step (3). For any $(\lambda_1, \mathbf{S}) \in \mathbf{FR}_{(\mathbf{p^d}, \mathbf{w^d})}/\mathbf{FR}^{\circ}_{(\mathbf{p^d}, \mathbf{w^d})}$, it either does not satisfy the KKT conditions, or $\mathbf{\Pi_1}(\lambda_1, \mathbf{S}) \leq \mathbf{\Pi_1}(\lambda^{\mathbf{d}}, \mathbf{S^d})$. We consider points on L_1, L_2 and L_3 separately, where L_i for $i \in \{1, 2, 3\}$ are define in Step (1) and illustrated in Figure C.1.

[$\mathbf{L_1}$]. Let $(\lambda_1', S') \in L_1$. We show that: (a) (λ_1', S') does not satisfy the KKT conditions for $S' \leq S^d$, and (b) $\Pi_1(\lambda_1, S) \leq \Pi_1(\lambda^d, S^d)$ for $S' > S^d$.

Scenario (a) $S' \leq S^d$. Let $\mu_1 > 0$. Suppose (for contradiction) that there exists $(\lambda_1', S') \in L_1$ with $S' \leq S^d$ that satisfies the following KKT conditions:

$$t\lambda_2' + p^d - 2t\lambda_1' - (t\lambda_1' + w^d)\frac{c'(\rho')}{c'(\rho') + S't} = \mu_1 c'(\frac{\lambda_1'}{S'})\frac{1}{S'}, \quad \text{and} \tag{C.19}$$

$$(t\lambda_1' + w^d)\frac{c'(\rho')}{c'(\rho') + S't}\frac{\lambda_1' + \lambda_2'}{S'} - \frac{2S'}{M} = -\mu_1 c'(\frac{\lambda_1'}{S'})\frac{\lambda_1'}{S'^2}. \tag{C.20}$$

Then (C.20) implies $(t\lambda_1' + w^d)\frac{c'(\rho')}{c'(\rho')+S't}\frac{\lambda_1'}{S'} \leq \frac{2S'}{M}$, and recall that $(t\lambda^d + w^d)\frac{c'(\rho^d)}{c'(\rho^d)+S^d t}\frac{2\lambda^d}{S^d} = \frac{2S^d}{M}$ by (C.10). It follows that

$$(t\lambda_1' + w^d)\frac{c'(\rho')}{c'(\rho') + S't}\frac{\lambda_1'}{S'} \Big/ \left[(t\lambda^d + w^d)\frac{c'(\rho^d)}{c'(\rho^d) + S^d t}\frac{2\lambda^d}{S^d}\right] \leq \frac{S'}{S^d} \leq 1. \tag{C.21}$$

Note that for $(\lambda_1', S') \in L_1$, the corresponding $\lambda_2' = 0$. As $t\lambda_2 = 1 - p^d - c(\rho)$ by (4.10), $\rho^d < \rho'$, which implies that $\frac{S'}{S^d} < \frac{t\lambda_1'}{2t\lambda^d} < \frac{t\lambda_1'+w^d}{t\lambda^d+w^d}$. Moreover, because $S^d \geq S'$, we have $\frac{c'(\rho')}{c'(\rho')+S't} > \frac{c'(\rho^d)}{c'(\rho^d)+S^d t}$. Therefore, (C.21) can not hold.

Scenario (b) $S' > S^d$. We first introduce a facilitating point $(\hat{\lambda}_1, S^d) \in L_1$ as shown in Figure C.2, i.e., the intersection of L_1 and $S = S^d$. Then for any $(\lambda_1', S') \in L_1$ and $S' > S^d$, we must have $\lambda_1' > \hat{\lambda}_1$. Because $\frac{\partial^2 \Pi_1(\lambda_1, S)}{\partial S^2} < 0$ and $\frac{\partial \Pi_1(\lambda_1, S)}{\partial \lambda_1}|_{(\lambda^d, S^d)} = 0$, we

must have $\Pi_1(\hat{\lambda}_1, S^d) < \Pi_1(\lambda^d, S^d)$. Then it suffices to show that $\Pi_1(\lambda'_1, S') < \Pi_1(\hat{\lambda}_1, S^d)$. Let $\hat{w}_1 = \frac{(S^d)^2 - Mw^d\hat{\lambda}_2}{M\hat{\lambda}_1}$ and $w'_1 = \frac{(S')^2 - Mw^d\lambda'_2}{M\lambda'_1}$ be the corresponding wage under $(\hat{\lambda}_1, S^d)$ and (λ'_1, S') respectively. We first show that $\hat{w}_1 < w'_1$. Because $\hat{\lambda}_2 = 0$ and $\lambda'_2 = 0$ (as $(\hat{\lambda}_1, S^d) \in L_1$ and $(\lambda'_1, S') \in L_1$) and $\hat{\rho} = \frac{\hat{\lambda}_1}{S^d} = \rho' = \frac{\lambda'_1}{S'}$ (as L_1 is linear), then $S' = M\rho' w'_1 > S^d = M\hat{\rho}\hat{w}_1$ implies that $\hat{w}_1 < w'_1$. Therefore, $\hat{w}_1\hat{\lambda}_1 < w'_1\lambda'_1$. We then consider two sub-scenarios: scenario (b.1) $\hat{\lambda}_1 \geq 2\lambda^d$; and scenario (b.2) $\lambda^d < \hat{\lambda}_1 < 2\lambda^d$.

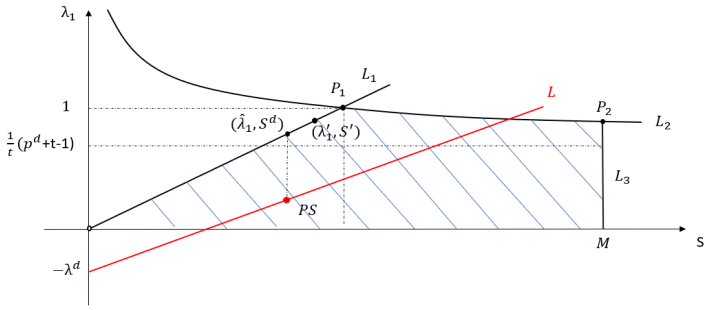


Figure C.2: Illustration of points $(\hat{\lambda}_1, S^d)$ and (λ'_1, S')

Scenario (b.1) $\hat{\lambda}_1 \geq 2\lambda^d$. Because $\hat{\rho} = \rho'$, $t\hat{\lambda}_1 + \hat{p}_1 = t\lambda'_1 + p'_1$. Let $\Delta = t\lambda'_1 - t\hat{\lambda}_1 = \hat{p}_1 - p'_1 > 0$, then $\lambda'_1(p'_1 - w'_1) < (\hat{\lambda}_1 + \frac{\Delta}{t})(\hat{p}_1 - \Delta) - \hat{\lambda}_1\hat{w}_1 = \hat{\lambda}_1(\hat{p}_1 - \hat{w}_1) + \frac{\Delta}{t}(\hat{p}_1 - t\hat{\lambda}_1) - \frac{\Delta^2}{t} < \hat{\lambda}_1(\hat{p}_1 - \hat{w}_1) + \frac{\Delta}{t}(\hat{p}_1 - t\hat{\lambda}_1)$. It suffices to show that $\hat{p}_1 - t\hat{\lambda}_1 < 0$. Because $t\hat{\lambda}_1 + \hat{p}_1 = t\hat{\lambda}_2 + p^d$ by (4.10), $t\hat{\lambda}_1 = p^d - \hat{p}_1 \geq 2t\lambda^d$, which implies that $\hat{p}_1 \leq p^d - 2t\lambda^d \leq 2w^d - t\lambda^d < t\hat{\lambda}_1$, where the last inequality is due to Lemma C.2.1.

Scenario (b.2) $\lambda^d < \hat{\lambda}_1 < 2\lambda^d$. Similar to the analysis in Scenario (b.1), let $\Delta = t\lambda'_1 - t\hat{\lambda}_1 = \hat{p}_1 - p'_1 > 0$. Then $\lambda'_1(p'_1 - w'_1) < \lambda'_1(p'_1 - \hat{w}_1) = (\hat{\lambda}_1 + \frac{\Delta}{t})(\hat{p}_1 - \Delta - \hat{w}_1) = \hat{\lambda}_1(\hat{p}_1 - \hat{w}_1) + \frac{\Delta}{t}(\hat{p}_1 - \hat{w}_1 - t\lambda'_1)$. It suffices to show that $\hat{p}_1 - \hat{w}_1 < t\lambda'_1$. Because $t\hat{\lambda}_1 = p^d - \hat{p}_1$ as $\hat{\lambda}_2 = 0$, and $\lambda^d < \hat{\lambda}_1$, we can obtain that $\hat{p}_1 \leq p^d - t\lambda^d$. Note that $(S^d)^2 = 2M\lambda^d w^d = M\hat{\lambda}_1\hat{w}_1$, we have $\frac{\hat{w}_1}{w^d} = \frac{2\lambda_d}{\hat{\lambda}_1} > 1$ and thus $\hat{w}_1 > w^d$. It follows that $\hat{p}_1 - \hat{w}_1 < p^d - t\lambda^d - \hat{w}_1 < p^d - t\lambda^d - w^d = w^d < t\lambda^d < t\hat{\lambda}_1 < t\lambda'_1$.

[$\mathbf{L_2}$]. Let $(\lambda'_1, S') \in L_2$. Let $\mu_2 > 0$. Suppose (for contradiction) that there exists $(\lambda'_1, S'_1) \in$

L_2 that satisfies the following KKT conditions:

$$t\lambda_2' + p^d - 2t\lambda_1' - (t\lambda_1' + w^d)\frac{c'(\rho')}{c'(\rho') + S't} = \mu_2 t, \text{ and} \tag{C.22}$$

$$(t\lambda_1' + w^d)\frac{c'(\rho')}{c'(\rho') + S't}\frac{\lambda_1' + \lambda_2'}{S'} - \frac{2S'}{M} = \mu_2 c'(\frac{1}{S'})\frac{1}{S'^2}. \tag{C.23}$$

We first consider the case where $p^d \geq \frac{1}{2}$. Because $(\lambda_1', S') \in L_2$, we have $\lambda_2' + \lambda_1' = 1$ and thus $t\lambda_1' = p^d + c(\frac{1}{S'}) + (t-1) \geq \frac{1}{2}$. Then because $t\lambda_2' + p^d = 1 - c(\frac{1}{S'}) < 1$, we have $t\lambda_2' + p^d - 2t\lambda_1' < 0$ which contradicts (C.22).

We then consider the case where $p^d < \frac{1}{2}$. As $\lambda_1 + \lambda_2 = 1$ for $(\lambda_1, S) \in L_2$, if $\lambda_1' > \lambda_2'$, we have $\lambda_1' > \frac{1}{2}$. It follows that $t\lambda_2' + p^d - 2t\lambda_1' \leq p^d - t\lambda_1' < 0$, which contradicts (C.22). If $\lambda_1' \leq \frac{1}{2}$, the proof consists of two parts.

Part (a). (λ', S') does not satisfy (C.22)–(C.23) when $\lambda_1' = \frac{1}{2}$. In this case, $\lambda_2' = \lambda_1' = \frac{1}{2}$. By (C.10), (λ^d, S^d) satisfies $(t\lambda^d + w^d)\frac{c'(\rho^d)}{c'(\rho^d) + S^d t}\rho^d = \frac{2S^d}{M}$. By (C.23), we have $(t\lambda_1' + w^d)\frac{c'(\rho')}{c'(\rho') + S't}\rho' \geq \frac{2S'}{M}$. It follows that

$$(t\lambda_1' + w^d)\frac{c'(\rho')}{c'(\rho') + S't}\rho' \Big/ \Big[(t\lambda^d + w^d)\frac{c'(\rho^d)}{c'(\rho^d) + S^d t}\rho^d\Big] \geq \frac{S'}{S^d}. \tag{C.24}$$

Because $\hat{\lambda}_2 = \frac{1}{2} > \lambda^d$ for $t > 1$ (as $\lim_{M\to\infty} \lambda^d = \frac{1}{2t}$ and one can show that $\frac{\partial \lambda^d}{\partial M} < 0$ by taking derivative with respect to (C.15)), $\rho' = \frac{\lambda_1' + \lambda_2'}{S'} < \rho^d$ by (4.10). It follows that $\frac{\lambda_1'}{\lambda^d} < \frac{S'}{S^d}$ and $S^d < \hat{S}$, which implies that $\frac{c'(\rho')\rho'}{c'(\rho') + S't} < \frac{c'(\rho^d)\rho^d}{c'(\rho^d) + S^d t}$. Moreover, $\lambda_1' = \frac{1}{2} > \lambda^d$ implies $\frac{t\lambda_1' + w^d}{t\lambda^d + w^d} < \frac{\lambda_1'}{\lambda^d} < \frac{S'}{S^d}$. Therefore, we reach a contradiction.

Part (b). Any $(\lambda_1', S') \in L_2$ with $\lambda_1' \leq \frac{1}{2}$ does not satisfy (C.22)–(C.23). Observe that for $(\lambda_1', S') \in L_2$, S' and λ_2' decrease in λ_1, and thus ρ' decreases in λ_1'. Therefore, (C.24) can not hold.

[$\mathbf{L_3}$]. Let $(\lambda_1', S') \in L_3$. In this case, $S' = M = M\frac{\lambda_1' w_1'}{S'} + M\frac{\lambda_2' w^d}{S'} \leq M\frac{\lambda_1' + \lambda_2'}{S'}\max\{w_1', w^d\}$, which implies that $w_1' \geq 1$. Therefore, $\Pi_1(\lambda_1', S') \leq \Pi_1(\lambda^d, S^d)$.

The Case in Which $t \in (\frac{2}{3}, 1)$

In this section, we assume that platform 2 adopts (p^d, w^d), which is the local KC equilibrium strategy. We show that platform 1 has no incentive to deviate from (λ^d, S^d) to any

$(\lambda_1, S_1) \in FR_{(p^d, w^d)}$. By following the same analysis as that for the case where $t > 1$, we can show that for any $(\lambda_1, S) \in FR_{(p^d, w^d)}$ and $(\lambda_1, S) \neq (\lambda^d, S^d)$, it either does not satisfy the KKT conditions, or $\Pi_1(\lambda_1, S) < \Pi_1(\lambda^d, S^d)$. We omit the detailed proof here.

C.4 Compare to the System without Competition

Recall that in the system without competition, the incumbent solves Problem (4.16). We characterize the monopoly equilibrium outcomes in Section C.4.1, and compare outcomes in systems with and without competition in Section C.4.2.

C.4.1 Proof of Theorem 4.5.1

We first show that given (p, w) with $p, w \in (0, 1)$, there exists a unique (λ, S) satisfying (4.17)–(4.18). By (4.17)–(4.18), define $LHS(\lambda) = 1 - t\lambda - p - c(\frac{\lambda}{S}) = 1 - t\lambda - p - c(\sqrt{\frac{\lambda}{Mw}})$. Observe that $LHS(\lambda)$ is continuous and strictly decreasing in λ. Moreover, $LHS(0) = 1 - p > 0$ and $LHS(1) = 1 - t - p - c(\frac{1}{Mw})$. If $LHS(1) \leq 0$, then there exists a unique $\lambda^* \in (0, 1]$ such that $LHS(\lambda^*) = 0$. Otherwise, $\lambda^* = 1$.

When $LHS(1) > 0$, i.e., $\lambda^* = 1$, the profit for the monopolist is $p - w$. In this case, the monopolist can gain more profit by increasing p or decreasing w so that $LHS(1) > 0$ still holds. Therefore, any (p, w) that leads to $LHS(1) > 0$ is suboptimal, and it suffices to focus on strategies such that (4.17)–(4.18) hold. By (4.17) and (4.18), we rewrite the optimization problem for the monopolist as:

$$\begin{aligned} \max_{\lambda, S} \quad & \Pi(\lambda, S) = \lambda - t\lambda^2 - \lambda c(\frac{\lambda}{S}) - \frac{S^2}{M}, \\ \text{subject to} \quad & \lambda \leq 1. \end{aligned}$$

The Hessian matrix for $\Pi(\lambda, S)$ is given by $H = \begin{pmatrix} \frac{\partial^2\Pi(\lambda,S)}{\partial\lambda^2} & \frac{\partial^2\Pi(\lambda,S)}{\partial\lambda\partial S} \\ \frac{\partial^2\Pi(\lambda,S)}{\partial S\partial\lambda} & \frac{\partial^2\Pi(\lambda,S)}{\partial S^2} \end{pmatrix}$, where

$$\begin{aligned}
&\frac{\partial^2\Pi(\lambda,S)}{\partial\lambda^2} = -2t - \frac{2}{S}c'(\frac{\lambda}{S}) - \frac{\lambda}{S^2}c''(\frac{\lambda}{S} < 0) < 0, \\
&\frac{\partial^2\Pi(\lambda,S)}{\partial\lambda\partial S} = \frac{\partial^2\Pi(\lambda,S)}{\partial S\partial\lambda} = \frac{2\lambda}{S^2}c'(\frac{\lambda}{S}) + \frac{\lambda^2}{S^3}c''(\frac{\lambda}{S}) > 0, \text{ and} \\
&\frac{\partial^2\Pi(\lambda,S)}{\partial S^2} = -c''(\frac{\lambda}{S})\frac{\lambda^3}{S^4} - c'(\frac{\lambda}{S})\frac{2\lambda^2}{S^3} - \frac{2}{M} < 0.
\end{aligned}$$

One can check that the determinant of H, $Det(H) = \frac{\partial^2\Pi(\lambda,S)}{\partial\lambda^2}\frac{\partial^2\Pi(\lambda,S)}{\partial S^2} - \frac{\partial\Pi(\lambda,S)}{\partial\lambda\partial S}\frac{\partial^2\Pi(\lambda,S)}{\partial S\partial\lambda} > 0$. Therefore, $\Pi(\lambda, S)$ is concave. Let $\mu \geq 0$, then it suffices to find (λ, S) that satisfies the following KKT conditions:

$$\frac{\partial\Pi(\lambda,S)}{\partial\lambda} = 1 - 2t\lambda - c(\frac{\lambda}{S}) - \frac{\lambda}{S}c'(\frac{\lambda}{S}) = \mu, \tag{C.25}$$

$$\frac{\partial\Pi(\lambda,S)}{\partial S} = c'(\frac{\lambda}{S})\frac{\lambda^2}{S^2} - \frac{2S}{M} = 0, \tag{C.26}$$

$$\mu(\lambda - 1) = 0. \tag{C.27}$$

We then consider the following two cases. Case (i) $\mu = 0$. We have $\lambda = \frac{1}{2t}[1 - c(\rho) - \rho c'(\rho)]$ by (C.25), and $S = \frac{1}{2}M\rho^2c'(\rho)$. Moreover, $\rho = \frac{\lambda}{S}$ implies that

$$tM\rho^3c'(\rho) + c(\rho) + \rho c'(\rho) = 1. \tag{C.28}$$

In this case, (C.28) admits a unique solution $\rho^m \in (0, 1)$. Moreover, $\lambda^m = \frac{1}{2t}[1 - c(\rho^m) - \rho^m c'(\rho^m)] = \frac{1}{2}M(\rho^m)^3c'(\rho^m)$, $S^m = \frac{1}{2}M(\rho^m)^2c'(\rho^m)$, $p^m = 1 - 2t\lambda^m - c(\rho^m) = 1 - \frac{t}{2}M(\rho^m)^3c'(\rho^m)$ and $w^d = \frac{S^m}{M\rho^m} = \frac{1}{2}\rho^m c'(\rho^m)$. In Lemma C.4.1, we provide the limit results on the monopoly equilibrium outcome given that the market is partially covered.

Lemma C.4.1. *For the monopoly equilibrium such that the market is partially covered, let* $C_m = \left[\frac{1}{tc'(0)}\right]^{1/3}$. *We have* $\lim_{M\to\infty}\frac{\rho^m}{M^{-1/3}} = C_m$, $\lim_{M\to\infty}\frac{S^m}{M^{1/3}} = \frac{1}{2tC_m}$, $\lim_{M\to\infty}\frac{w^m}{M^{-1/3}} = \frac{1}{2tC_m^2}$, $\lim_{M\to\infty} p^m = \frac{1}{2}$ *and* $\lim_{M\to\infty}\lambda^m = \frac{1}{2t}$.

Proof of Lemma C.4.1. Because ρ^m is the unique solution to (C.28), the analysis is similar to that of Lemma C.2.2. ∎

Case (ii) $\mu > 0$. By (C.13) $\lambda = 1$, and by (C.27) S^m is the unique solution to $c'(\frac{1}{S})\frac{1}{S^2} - \frac{2S}{M} = 0$, which is equivalent to

$$\frac{1}{2}M\rho^3 c'(\rho) = 1. \tag{C.29}$$

In this case, (C.29) admits a unique solution $\rho^m \in (0,1)$ to (C.29). Moreover, $\lambda^m = 1$, $S^m = \frac{1}{\rho^m}$, $w^d = \frac{1}{M(\rho^m)^2}$ and $p^d = 1 - t - c(\rho^m)$. In Lemma C.4.2, we provide the limit results on the monopoly equilibrium given that the demand market is fully covered.

Lemma C.4.2. *For the monopoly equilibrium such that the demand market is fully covered, we have* $\lim_{M\to\infty} \frac{\rho^m}{M^{-1/3}} = \left[\frac{2}{c'(0)}\right]^{1/3}$, $\lim_{M\to\infty} \frac{S^m}{M^{1/3}} = \left[\frac{2}{c'(0)}\right]^{-1/3}$, $\lim_{M\to\infty} \frac{w^m}{M^{-1/3}} = \left[\frac{2}{c'(0)}\right]^{-2/3}$, $\lim_{M\to\infty} p^m = 1 - t$ *and* $\lambda^m = 1$.

Proof of Lemma C.4.2. Because ρ^m is the unique solution to (C.29), the analysis is similar to that of Lemma C.2.2. ∎

Finally, observe that the solution ρ^m to (C.28) is is increasing in m and $\lambda^m = \frac{1}{2t}[1 - c(\rho^m) - \rho^m c'(\rho^m)]$ is increasing to $\frac{1}{2t}$ as $M \to \infty$. Therefore, if $t \geq \frac{1}{2}$, the optimal strategy of the incumbent induces an equilibrium such that the market is not fully covered. Otherwise, there exists a threshold $\overline{M}^c$ such that the demand market is partially covered if $M < \overline{M}^c$ and the demand market is fully covered otherwise.

C.4.2 Proof of Theorem 4.5.2

Recall from Theorem 4.4.1, a symmetric duopoly equilibrium does not exists when $t \in (0, \frac{2}{3})$. Therefore, we first prove the case in which $t \geq 1$ and then the case in which $t \in (\frac{2}{3}, 1)$.

Case (i). t > 1.

Compare ρ^d and ρ^m. Recall that ρ^m is the unique solution to (C.28), and ρ^d is the unique solution to (C.14) given $M \geq \frac{2}{t[c^{-1}(1)]^2}$. Observe that both $c'(\rho)\rho(tM\rho^2 + 1) + c(\rho)$ and $c'(\rho)\rho(\frac{tM\rho^2}{4} - \frac{1}{tM\rho^2}) + c(\rho)$ increase in ρ. Moreover, $c'(\rho)\rho(M\rho^2 + 1) + c(\rho) > c'(\rho)\rho(\frac{tM\rho^2}{4} - \frac{1}{tM\rho^2}) + c(\rho)$, as $tM\rho^2 + 1 > \frac{tM\rho^2}{4} - \frac{1}{tM\rho^2}$ for any $\rho > 0$. Therefore, $\rho^d > \rho^m$.

Compare w^d and w^m. By Lemma C.2.2 and Lemma C.4.1, we have $\lim_{M\to\infty} \frac{w^d}{w^m} = \frac{2tC_m^2}{tC_P^2} = \left(\frac{1}{2}\right)^{1/3} < 1$. Therefore, $w^m > w^d$ when M is sufficiently large.

Compare LW^d and LW^m. By (4.6) and the fact that $S = M\hat{w}$ it is equivalent to compare S^d and S^m. By Lemma C.2.2 and Lemma C.4.1, $\lim_{M\to\infty} \frac{S^d}{S^m} = \frac{2tC_m}{C_P} = 2^{1/3} > 1$. It follows that $S^d > S^m$ when M is sufficiently large.

Compare p^d and p^m. According to the monopoly equilibrium outcomes characterized in Appendix C.4.1, we can obtain that $t\lambda^m + w^m = p^m - w^m$. Moreover, $t\lambda^m + w^m = 1 - (p^m - w^m) - c(\rho^m)$, we thus have $p^m = w^m + \frac{1}{2}(1 - c(\rho^m))$. Similarly, we have $p^d = w^d + \frac{1}{2}(1 - c(\rho^d))$. Because $\rho^d > \rho^m$ and $w^d < w^m$ when M is sufficiently large, we have $p^d < p^m$ when M is sufficiently large.

Compare CS^d and CS^m. By (4.5), $CS^d = 2\int_0^{\lambda^d} = (1 - p^d - c(\rho^d) - tx)dx = 2\int_0^{\lambda^d} t(\lambda^d - x)dx = t(\lambda^d)^2$ and $CS^m = \frac{t}{2}(\lambda^m)^2$. By Lemma C.2.2 and Lemma C.4.1, $\lim_{M\to+\infty} \lambda^d = \lim_{M\to+\infty} \lambda^m = \frac{1}{2t}$. It follows that $(\lambda^d)^2 \geq \frac{1}{2}(\lambda^m)^2$ and thus $CS^d \geq CS^m$ when M is sufficiently large.

Case (ii): $\mathbf{t} \in (\frac{2}{3}, 1)$.

Compare ρ^d and ρ^m. By (C.16), when M is sufficiently large such that $1 - \frac{t}{2} - c(\rho) - \frac{1}{M\rho^2} = 0$ admits two roots with respect to ρ, we have $0 \geq 1 - \frac{3}{2}t - c(\rho^d) - \frac{1}{M(\rho^d)^2} = p^d - w^d - t > -t$. Thus, $c'(\rho^d) > \frac{2}{M(\rho^d)^3}$. Recall that, ρ^m satisfies (C.28), which implies that $c'(\rho^m) = \frac{1-c(\rho^m)}{tM(\rho^m)^3+\rho^m} < \frac{1-c(\rho^m)}{tM(\rho^m)^3}$. To show that $\rho^d > \rho^m$, it suffices to show that $\frac{1-c(\rho^m)}{t} < 2$, which holds as $t > \frac{2}{3}$.

Compare LW^d and LW^m. It is equivalent to compare S^d and S^m. By Lemma C.2.3 and Lemma C.4.1, $\lim_{M\to\infty} \frac{S^d}{S^m} = \frac{2tC_m}{C_K} = \left[4(\frac{3}{2}t - 1)t\right]^{1/3}$. Therefore, when M is sufficiently large, $S^d > S^m$ if $t \in (\frac{2+\sqrt{10}}{6}, 1)$, and $S^d < S^m$ if $t \in (\frac{2}{3}, \frac{2+\sqrt{10}}{6})$.

Compare w^d and w^m. By Lemma C.2.3 and Lemma C.4.1, $\lim_{M\to\infty} \frac{w^d}{w^m} = \frac{2tC_m^2}{C_K^2} = \left[\frac{3t-2}{t^2}\right] < 1$ for $t \in (\frac{2}{3}, 1)$. It follows that $w^d < w^m$ when M is sufficiently large.

Compare CS^d and CS^m. Because $CS^d = 2\int_0^{\lambda^d}(1 - p^d - c(\frac{1}{S^d}) - tx)dx = \frac{t}{4}$ and $CS^m = \int_0^{\lambda^m}(1 - p^m - c(\rho^m) - tx)dx = \frac{t}{2}(\lambda^m)^2$, when M is sufficiently large, we have $CS^d < CS^m$ if $t \in (\frac{2}{3}, \frac{\sqrt{2}}{2})$ and $CS^d > SC^m$ if $t \in (\frac{\sqrt{2}}{2}, 1)$.