AN OPERATIONAL ANALYSIS OF INNOVATIVE TRANSPORTATION TECHNOLOGIES

Neda Mirzaeian

A Dissertation

Submitted to the Tepper School of Business

In Partial Fulfilment of the Requirements for the Degree of

Doctor of Philosophy in Operations Management

Dissertation Committee:

Soo-Haeng Cho (Chair)

Alan Scheller-Wolf

Sean Qian

Joseph Xu

Alexandre Jacquillat

May 2022

Abstract

This dissertation focuses on examining three problems at the intersection of smart city operations and innovative transportation technologies. In particular, the first two chapters study the potential effects of autonomous vehicles (AVs) on highway congestion and downtown parking, respectively. The third chapter of this dissertation studies the effect of ride-hailing passenger drop-offs on downtown rush-hour congestion and parking.

In the first chapter, I investigate the effects of AVs on highway congestion. AVs have the potential to significantly reduce highway congestion because they can maintain smaller intervehicle gaps and travel together in larger platoons than human-driven vehicles (HVs). Various policies have been proposed to regulate AV travel on highways, yet no in-depth comparison of these policies exists. To address this shortcoming, I develop a queueing model for a multilane highway and analyze two policies: the designated-lane policy ("D policy"), under which one lane is designated to AVs, and the integrated policy ("I policy"), under which AVs travel together with HVs in all lanes. I connect the service rate to intervehicle gaps (governed by a Markovian arrival process) and congestion, and measure the performance using mean travel time and throughput. My analysis shows that although the I policy performs at least as well as a benchmark case with no AVs, the D policy outperforms the benchmark only when the highway is heavily congested and AVs constitute the majority of vehicles; in such a case, this policy may outperform the I policy only in terms of throughput. These findings caution against recent industry and government proposals that the D policy should be employed at the beginning of the mass appearance of AVs. Finally, I calibrate the model to data and show that for highly congested highways, a moderate number of AVs can make a substantial improvement (e.g., 22% AVs can improve throughput by 30%), and when all vehicles are AVs, throughput can be increased by over 400%.

In the second chapter, I study how AVs may change the morning commute travel pattern and improve downtown parking. I develop a continuous-time traffic model that takes into account key economic deterrents to driving, such as parking fees and traffic congestion, and characterize the departure time and parking location (downtown or outside downtown parking area) patterns of commuters in equilibrium. To illustrate the results, the model is calibrated to data from Pittsburgh. For the calibrated model, my analysis shows that all AV commuters choose to park outside downtown, increasing both vehicle hours and vehicle miles traveled as compared to the case with all human-driven vehicles. This change increases the total system cost and suggests a potential downtown land-use change (e.g., repurposing downtown parking spots to commercial and residential areas) in Pittsburgh after mass adoption of AVs. To reduce the total system cost, a social planner may be interested in regulating commuters' decisions by adjusting parking fees and/or imposing congestion tolls as a short-term measure, or adjusting infrastructure, e.g., converting downtown parking spaces to curbside drop-off spots for AVs. My results indicate that these measures can reduce the total system cost substantially (e.g., up to 70% in my calibrated model).

In the third chapter, I investigate how ride-hailing may change the morning commute travel pattern and improve downtown parking. Similar to the second chapter, I develop a continuous-time traffic model that takes into account key economic deterrents to driving, such as parking fees and traffic congestion, and characterize the departure time patterns and transportation modes (driving or ride-hailing) of commuters in equilibrium. To illustrate the results, the model is once again calibrated to data from Pittsburgh. For the calibrated model, my analysis shows that as drop-off congestion increases, the number of commuters who switch from driving to using ride-hailing increases, which leads to an increase in vehicle hours traveled as compared to the case when all commuters drive. This change increases the total system cost and presents a potential opportunity for repurposing downtown parking spots to commercial and residential areas in Pittsburgh. To reduce the total system cost, a social planner may be interested in regulating commuters' and the ride-hailing company's decisions by adjusting parking fees and/or imposing drop-off tolls as a short-term measure, or adjusting infrastructure, i.e., increasing the number of curbside drop-off spots for ride-hailing vehicles. My results indicate that these measures can reduce the total system cost substantially (e.g., up to 77% in the calibrated model).

Acknowledgments

I would like to thank my advisor and dissertation chair, Prof. Soo-Haeng Cho, who offered guidance, support and wisdom throughout my PhD journey. I am forever grateful for having the opportunity to learn from you. Thank you for helping me put things into perspective, guiding me through all stages of my PhD, and preparing me for my future career. I wish to also thank my dissertation committee members and co-authors, Prof. Alan Scheller-Wolf and Prof. Sean Qian, whose insights and inputs helped this dissertation come to fruition. I cannot thank you enough for the training and support you have given me. Additional thanks to my dissertation committee members, Prof. Joseph Xu and Prof. Alexandre Jacquillat, who offered helpful feedback and inspiring ideas on this dissertation.

I am grateful for the support I received from the Arizona Department of Transportation, especially Reza Karimvand and Vahid Goftar, who kindly shared their valuable data with me. A special thanks to Prof. Sridhar Tayur. You have been one of my most inspiring mentors and I have learned a lot from you both professionally and personally. I wish to also thank Prof. Mor Harchol-Balter, who inspired me to love teaching and encouraged me to apply for this PhD program.

The memories I made with my friends at Tepper are amongst the most valuable things that I will cherish forever. I wish to thank Musa Celdir, Franco Berbeglia, and Siddharth Singh for all the intellectual discussions and also for making grading much more enjoyable. I would like to extend thanks to Thiago Serra and Aleksandr Kazachkov for introducing me to the CMU INFORMS Student Chapter, which not only gave me the opportunity to collaborate with some of the brightest scholars at Tepper, but also opened a door for me to make many other friends around the world. Many thanks to my colleagues and friends Sagnik Das, Melda Korkut, Violet Chen, Yuyan Wang, Ozgun Elci, Savannah Tang, Amin Hosseininasab, Ziye Tang, Cristiana Lopes Lara, David Bernal,

Mehmet Aydemir, Daniel de Roux, Matthew Diabes, Kyra Gan, Anthony Karahalios, Thomas Lavastida, and Neha Neha for your friendship. I have learned so much from all of you. Lawrence Rapp and Laila Lee, I am extremely grateful for you. You make life much easier for Tepper PhD students.

I have so much gratitude and love for the friends I have made in Pittsburgh. Thank you Prof. Anna Svirsko, my first and best American friend, for being there for me from the beginning of graduate school. I am immensely thankful for how you helped me become familiar with American culture and (almost) lose my accent. I wish to also thank my incredible friends Jacquelynn Jones, Zahra Ebrahimi, Faezeh Movahedi and Mahbaneh Eshaghzadeh. You are the best girl friends I could ask for. Dr. Nazanin Esmaili, thank you for taking me under your wing when I moved to Pittsburgh and encouraging me to do a PhD in operations management.

And above all, my family: This thesis is dedicated to you! No words can ever be strong enough to express my gratitude to my parents, Nezhat and Hamid, for their unconditional love. Although I did not get to see you for almost the entire duration of my PhD, you have always been there for me. Thank you for all the sacrifices you have made for me. I am also grateful for my brother, Nima. Thank you for teaching me to break away from the norms and encouraging me to push myself beyond my limits. I would like to extend my gratitude to my future in-laws, especially Simin and Greg Curtis. Thank you for treating me like your own daughter. Finally, a special thanks to my lovely fiancé, Peter Curtis, who joined me mid-way through this PhD journey, but has become my number one supporter. Thank you for your love and thank you for being there for me during the most difficult times. You are my rock.

List of Coauthors

Chapter 1:

Soo-Haeng Cho

Professor of Operations Management and Strategy

Tepper School of Business

Carnegie Mellon University

Alan Scheller-Wolf

Richard M. Cyert Professor of Operations Management Tepper School of Business

Carnegie Mellon University

Chapters 2 & 3:

Soo-Haeng Cho

Professor of Operations Management and Strategy

Tepper School of Business

Carnegie Mellon University

Sean Qian

Henry Posner, Anne Molloy, And Robert And Christine Pietrandrea Associate Professor

Heinz College of Information Systems and Public Policy & Department of Civil and Environmental Engineering

Carnegie Mellon University

Contents

List of Tables

List of Figures

Chapter 0

Introduction

The emergence of innovative transportation technologies can fundamentally change the shape of our cities and improve the efficiency of our transportation networks. However, these new technologies introduce new risks and challenges. This dissertation studies the effect of two innovative technologies – autonomous vehicles (AVs) and ride-hailing services – on different aspects of transportation, and offers practical recommendations for city planners.

The first and second chapters of this dissertation focus on AVs – they are expected to play a determinant role in shaping the future of mobility. AVs are already operating on the roadways of several major cities in the U.S. and other countries. Industry experts predict that, by 2035, a quarter of all vehicles on the road will be autonomous (Bierstedt et al. 2014). As such, many large cities are preparing for AVs in their long-range transportation plans (National League of Cities 2018): This is the window of opportunity for city planners to put in place policies that pave the way for the inevitable mass arrival of AVs.

Chapter 1 investigates how the large-scale adoption of AVs will affect highway congestion. AVs potentially have several advantages over human drivers with respect to highway driving: Since they are able to communicate with each other, AVs move (and brake) more smoothly, and in general have shorter reaction times than human-driven vehicles (HVs). This allows AVs to reduce their inter-vehicle gap on highways, and travel together in larger platoons (or batches) than HVs. Despite the fact that these advantages are known, it is unknown how highway congestion will change if AVs are allowed on highways, or how they will affect HV traffic. Consequently, various policies have been proposed to regulate AVs on the roads, yet no in-depth comparison of these policies exists.

We address these shortcomings.

As a benchmark, we first model a segment of a highway as a queueing model in the absence of AVs. We then propose two policies for a highway with both HVs and AVs: the designated-lane policy ("D policy") and the integrated policy ("I policy"). Under the designated-lane policy, one lane is designated to AVs, to separate them from HVs. Under the integrated policy, AVs travel together with HVs. In our queueing models, the arrival rate to a highway can differ depending on the highway location, but the speed of vehicles primarily depends on the number of vehicles currently driving on the highway, so we focus on estimating the state-dependent speed of vehicles. We estimate this value for HVs based on data from highways in Arizona (Arizona DOT 2017). We then estimate the speed of AVs by modeling the platoon formation process, and calibrating this model to the parameter values derived from field experiments of AVs.

We evaluate our policies both analytically and numerically. Our analysis shows that, in terms of mean travel time of an individual commuter, whereas the I policy always improves the performance of the highway over the benchmark case, the D policy out-performs the benchmark case only when the highway is congested and AVs constitute a significant proportion of the vehicles. In terms of throughput, we show that the performance of the D and I policies depend on the proportion of AVs. In particular, for a highly loaded highway, both policies are capable of increasing the throughput over the benchmark case, and a moderate number of AVs can make a substantial improvement.

Chapter 2 focuses on the effect of AVs on one of the primary issues facing city planners: managing parking and traffic congestion during morning rush hour. To accommodate high parking demands, a city has to surrender large amounts of space to build parking structures; for example, an astounding 14 percent of Los Angeles county land is dedicated to parking. AVs might be able to alleviate this problem, as they have the ability to drop commuters off at their workplaces in a city center and park in suburban areas with less dense businesses and cheaper parking. This ability of AVs not only allows commuters to avoid high parking fees in a city center, but also reduces the need to build or maintain large parking structures there.

This chapter provides guidance to municipal governments on how to adapt infrastructures, including roads and parking facilities, to the special needs and characteristics of AVs. In particular, we investigate the potential of AVs to solve the parking problem that commuters face when traveling to a central business district. We develop a continuous-time traffic model in which commuters decide

when to leave their residences and where to park their cars in one of two available parking areas: central (located in downtown), and external (located outside of downtown). The goal of commuters is to minimize their total transportation and parking costs during their morning commutes.

We show that the commuters' decisions under equilibrium may differ from a social optimum that minimizes the aggregate costs of all commuters. As a way to reduce the gap between individual and social optima, we determine a social planner's decisions on parking fees as well as parking or curbside drop-off capacities. To offer concrete insights into the effect of these two innovative technologies on downtown parking, we use traffic and parking data from the city of Pittsburgh. Results from this research can aid major cities, including the city of Pittsburgh, in the development of short- and long-term transportation and infrastructure plans.

Chapter 3 studies the effect of ride-hailing drop-offs on the morning commute problem. Ride-sharing vehicles might be able to alleviate the parking problem, as they have the ability to drop commuters off at their workplaces in a city center without parking. This allows commuters to avoid high parking fees in a city center, and reduces the need to build or maintain large parking structures there. However, ride-hailing services contribute to traffic congestion as ride-hailing drop-offs may disturb or block the traffic flow of conventional vehicles. This negative externality of ride-hailing could push some commuters to change their mode of transportation from personal vehicles to ride-hailing, creating even more curbside congestion.

Similar to chapter 2, we develop a continuous-time traffic model in which commuters decide when to leave their residences and what transportation mode they choose from one of two available options: personal vehicles and ride-hailing services. The goal of commuters is to minimize their total transportation costs (including parking fees, ride-hailing fares, and imputed costs of early arrivals) during their morning commutes. In addition, the transportation network company (TNC) decides on the ride-hailing fares. The goal of the TNC is to maximize its profit while minimizing the total duration of the ride-hailing trips.

We show that the equilibrium commuters' decisions, which are directly influenced by the TNC's decision on the ride-hailing fares, differ from the decisions made by a social planner. In fact, in an unregulated market, the TNC can control the market to maximize its profit, creating a significant amount of congestion and increasing the total system cost. We provide practical solutions for the social planner to reduce the total system cost. In particular, the social planner is able to

reduce the total system cost as well as the individual commuter cost by imposing dynamic parking fees and drop-off tolls. In addition, to further decrease the cost a social planner can increase the number of dedicated curbside drop-off spots to reduce the drop-off congestion as well as the negative externality this causes for commuters who drive.

Chapter 4 concludes this dissertation, with a summary of our contributions and ideas for future research directions.

Chapter 1

A Queueing Model and Analysis for Autonomous Vehicles on Highways

1.1 Introduction

The autonomous vehicle (AV) industry is growing enormously fast. Today all major automobile manufacturers, as well as many research centers, are running experiments with AVs, and several manufacturers have announced long-term plans to mass-produce AVs (Muoio 2017). In fact, we already see AVs operating on the roads of several major cities in the U.S. and several other countries. As such, more than half of the largest U.S. cities are preparing for autonomous vehicles in their long-range transportation plans (National League of Cities 2018). Industry experts predict that by 2025, fully autonomous vehicles will arrive on highways (The automated driving community 2018), and by 2035, one out of every four vehicles on the road will be an AV (Bierstedt et al. 2014). KPMG predicts that AVs are capable of increasing the capacity of highways by 500%, without building new roads (Albright et al. 2015). But, despite all these claims, very little is actually known regarding the post-AV era.

In this paper, we investigate the effects of AVs on highway congestion. Highways are the arteries of nations; for example, highways account for 24% of total travel in the United States (Federal Highway Administration 2011). Congestion has always been a critical issue for urban planners; in the United States, on average 5.5% of commuter time is spent in congestion (INRIX

2017). Owing to their efficient driving performance, AVs have great potential to reduce time spent in congestion, should they be utilized properly. To capitalize on this, the Departments of Transportation in Colorado, Wisconsin and Washington are considering designated lanes for AVs (Aguilar 2018). Bierstedt et al. (2014) prescribe a different evolution. They claim that before 2025 there will be very few AVs on highways and their effect on highway traffic flow will be negligible, obviating the need for a dedicated lane. But they predict that by 2025 there will be enough AVs on highways that designating one separate lane to them will become reasonable. By 2030, after having AVs on designated highway lanes for 5 years, they postulate that HVs and AVs can be integrated on all lanes of highways.

To examine how AVs will affect the congestion of highways and how they should be incorporated onto highways, we develop a queueing model for a multi-lane highway. We analyze two policies for a mixed fleet of human-driven vehicles (HVs) and AVs: the *designated-lane* policy ("D policy") under which one lane is designated to AVs[1], and the *integrated* policy ("I policy") under which AVs travel together with HVs in all lanes. Using this model, we compare the mean travel time of a single vehicle as well as the throughput of the highway under each of the two policies, and also against a benchmark case in which all the vehicles are HVs. Specifically, we answer the following questions: (1) When is it optimal to use the D policy over the I policy?; and (2) How much will AVs improve highway traffic flow under each of these two policies?

We model traffic flow on a highway segment as an $M/G_n/c/c$ queueing system. Vehicles arrive individually to the highway segment, and the service time of a vehicle is defined as the amount of time it takes to traverse the segment. This travel time depends on congestion – the number of vehicles (n) that are simultaneously using the highway segment (i.e., the state of the queueing system) – so it is state-dependent. The queue capacity c of the highway is defined as the number of vehicles that it can accommodate at saturation, i.e., when the traffic forms a jam.

Our queueing model captures the potential benefits of AVs by explicitly modeling platoons and headway. Observing traffic on a highway, one notices that HVs usually move in *platoons* (batches). Within each platoon vehicles follow one another while maintaining a small *intraplatoon headway* - the time gap (in seconds) between two vehicles. The headway between two consecutive

[1]Should large enough numbers of AVs enter the highway, one could assign multiple AV lanes. We consider this case in Appendix A.5.5, while focusing our analysis in this chapter on designating one lane to AVs.

platoons, *interplatoon headway*, is typically significantly greater than the intraplatoon headway. Thus, the overall headway between vehicles depends on the size of platoons and the intraplatoon and interplatoon headways. Platooning is different for AVs and HVs: According to several field experiments (e.g., Bergenhem et al. (2012), Amoozadeh et al. (2015), and Zhao & Sun (2013)), AVs are capable of safely forming larger platoons than HVs, and the intraplatoon headway tends to be smaller for AVs than HVs. These two benefits arise because AVs can communicate with each other, and also move and brake more smoothly than HVs. To capture the platooning process, we use a Markovian Arrival Process (MAP); within a semi-renewal framework, the MAP enables us to model the intraplatoon and interplatoon headways as well as the size of platoons. A platoon consists of only HVs, exclusively AVs or HVs, or a mix of AVs and HVs in the benchmark case, under the D policy, and under the I policy, respectively. The difference in the mix of vehicle fleets leads to different vehicle speeds (hence, different service rates) under different policies.

We calibrate our models to data, and evaluate our policies both analytically and numerically. Our analysis shows that, in terms of mean travel time, while the I policy always improves the performance of the highway over the benchmark case, the D policy outperforms the benchmark case only when the highway is congested and AVs constitute a significant proportion of the vehicles. This calls into question the industry proposals in Bierstedt et al. (2014), as well as the policies being considered in Colorado, Wisconsin and Washington.[2] In terms of throughput, we show that the performance of the D and I policies depends on the proportion of AVs. In particular, for a highly loaded highway, both policies are capable of increasing the throughput over the benchmark case: Under the D policy and the I policy, a 30% increase in throughput is achievable when the AV proportion is 0.24 and 0.22, respectively. This implies that for highly congested highways, a moderate number of AVs can make substantial improvement.

We provide suggestions to policy makers about when and under what conditions each of the D and I policies should be utilized. Based on our analysis, if the mean travel time of vehicles is of primary importance, the I policy is advisable. If a policy maker bases a decision primarily

[2]This paper focuses on the potential effects of employing the D and I policies on travel time and throughput. We do not consider other possible benefits (e.g., fuel consumption, and the environment), or behavioral issues (e.g., confusion of human drivers) of AVs. In particular, we do not explicitly consider safety in our model, which might favor the D policy because HVs and AVs potentially may not mix well together under the I policy (Eliot 2019).

on throughput in a congested highway, then for moderate AV proportions, the D policy is recommended; otherwise, the I policy performs better. Specially, in our calibrated model, only when the AV proportion is between 0.25 and 0.55 does the D policy result in a higher throughput than the I policy on a congested highway.

The rest of this chapter is organized as follows. In §1.2, we review the related literature. Our two policies with AVs as well as a benchmark case for highway traffic flow are presented in §1.3. In §1.4 we calibrate the model to data, and in §1.5, we compare the policies with the benchmark case. In §1.6, we present a simulation study that validates our queueing model and analysis. We conclude in §1.7.

1.2 Related Literature

Our work is related to three streams of research: smart city operations, highway traffic flow modeling and platooning of vehicles, and autonomous vehicles.

In the smart city operations literature, ride-sharing (e.g., Benjaafar et al. 2018 and Qi et al. 2018), electric vehicles (e.g., Lim et al. 2014 and Mak et al. 2013), and the intersection of these two (e.g., He et al. 2017) have been studied. For a comprehensive review of ride-sharing and electric vehicles, we refer readers to He et al. (2018) and Pelletier et al. (2016), respectively. Recently, we begin to see some studies on AVs by operations researchers. For example, Baron et al. (2018) study the effect of AVs on social welfare, and Daw et al. (2019) investigate staffing of a remote support center for AVs. Our paper contributes to the expanding literature on smart city operations by examining the effect of AVs on highway congestion.

To model highway traffic flow, a variety of queueing models have been used. Kuwahara & Newell (1987) use a non-stationary queueing network to model traffic flow into a core city. Heidemann (1996) models an uninterrupted traffic flow as the stationary queueing models $M/M/1$ and $M/G/1$. Jain & Smith (1997) use an $M/G_n/c/c$ model in which state n is defined as the number of vehicles simultaneously driving on the highway. In their model, state-dependent service rates account for the effect of congestion on the speed of vehicles, but vehicles never form platoons. Vandaele et al. (2000) study the transient behavior of $M/M/1$, $M/G/1$ and $GI/G/1$ queues with or without state-dependent service times in traffic modeling. The validity of an $M/G_n/c/c$ queueing model is shown

empirically by Van Woensel & Vandaele (2006). Van Woensel & Vandaele (2007) provide a review of different queueing models for traffic on highways. In this paper, we follow Jain & Smith (1997) by using an $M/G_n/c/c$ queueing model, augmented with an incorporation of platooning.

Platoon formation in the case of interrupted traffic flow (e.g., when vehicular motion is interrupted by stoppages such as traffic lights) is studied in Dunne (1967), Lehoczky (1972) and Daganzo (1994). Neuts & Chakravarthy (1981) examine a continuous-time MAP for platoon formation on highways. A MAP is first introduced by Neuts (1979) as a versatile Markovian point process; he describes it as an extension of a Poisson process. Later Lucantoni (1991) provides a more convenient notation for the MAP. Alfa & Neuts (1995) show that the MAP is a valid model for platoon arrivals to a highway, but they do not provide a queueing model to examine traffic flow. Breuer & Alfa (2005) present a procedure for estimating the parameters of the MAP. In this paper, we use the MAP to model the formation of platoons as vehicles drive on highways. We then use the headway between vehicles derived from the MAP to calculate the state-dependent service rates of the $M/G_n/c/c$ queueing system, coupling the two models.

Research on autonomous vehicles is nascent, but growing fast. Prior studies that investigate the effects of AVs on traffic flow are particularly relevant to our paper. Qom et al. (2016) and Talebpour et al. (2017) conduct simulation studies investigating the effect of designating a lane to AVs on throughput. Chen et al. (2017) develop an analytical model to show that segregating AVs and HVs, rather than mixing them, leads to a smaller improvement in the capacity of the highway. Different from our paper, they assume that the number of AVs entering the highway is fixed, and that the headway between vehicles is deterministic, ignoring the effect of congestion on traffic flow.

For a mixed fleet of AVs and HVs, several papers have studied the effects of adding some preliminary autonomous features to vehicles on throughput. For example, adaptive cruise control (ACC) and cooperative ACC (CACC) have been studied in multiple papers. For example, Shladover et al. (2012) use experimental data to study the effects of ACC and CACC on traffic, and find that ACC does not have a remarkable impact on highway capacity; however, if a moderate to high percentage of vehicles adopt CACC, a significant increase in capacity is expected. In contrast, Stern et al. (2018) show through a field experiment that congestion can be eradicated by only a few AVs (1 out of 21 vehicles, or about 5%). Mohajerpoor & Ramezani (2019) characterize the mean headway between AVs by modeling platoon size as a binomial distribution. They calculate delay

for a segmented and mixed fleet of AVs and HVs on a 2-lane highway. In their model the effect of headway on throughput is not investigated, and the traffic flow is not explicitly modeled. There exist several other experimental and simulation analyses of mixed traffic; see Liu et al. (2018) and references therein. In our paper, as in Shladover et al. (2012), AVs are equipped with CACC. We find that although a small proportion of AVs can have a substantial effect on highway performance, the result obtained by Stern et al. (2018) is overly optimistic: to fully eradicate congestion, a substantial number of AVs are typically needed.

Ghiasi et al. (2017) is the closest work to our paper. They model the platoon structure of a mixed fleet of AVs and HVs driving on a one-lane highway segment, using a Markov chain. In their model, the arrival process to the highway is a vehicle stream of a fixed length, where any number of consecutive AVs in this stream can form a platoon, but no HV can be a part of a platoon. They show that, when the mean headway between an HV and an AV is lower than that between two HVs and higher than that between two AVs, the throughput of the highway increases with the AV proportion. Our model is more general than Ghiasi et al. (2017) in several respects. First, the arrival of vehicles to the highway follows a stochastic process, so the number of vehicles on the highway is not necessarily fixed. Second, the headway between two vehicles in our model not only depends on their vehicle types (i.e., HV-HV, HV-AV, AV-HV, or AV-AV), but also on the number of vehicles simultaneously present on the highway (i.e., state of our queueing system). As a result, the speed of a vehicle on the highway is impacted by all other vehicles. Third, we consider a multi-lane highway, which takes into account the effect of number of lanes on the state-dependent speed of vehicles. In addition, our multi-lane traffic model enables us to compare the performance of the D and I policies. Finally, our richer model yields different results than Ghiasi et al. (2017): For example, our result indicates that an integrated fleet of AVs and HVs can improve the throughput of the highway under more general conditions than those indicated by Ghiasi et al. (2017).

In summary, this chapter presents the first queueing model for a multi-lane highway with AVs. Our model captures several realistic features of highways, such as stochastic headway between vehicles, state-dependent speed of vehicles, stochastic arrival of vehicles to the highway, and mixed platoons of AVs and HVs. Whereas most prior papers focus on either the D policy or on the I policy, we compare these two policies to provide a guideline for policy makers. Although prior studies measure the impact of AVs by throughput, our results suggest that a policy which results

in a higher throughput does not necessarily have a lower mean travel time.

1.3 Model

In §1.3.1 we present a general model for highway traffic flow. In §1.3.2, we tailor this model to represent a benchmark as well as two policies with AVs. Table A1 in the Appendix summarizes our notation.

1.3.1 The Highway Traffic Flow Model

To model traffic flow on a highway, we consider an $M/G/c/c$ queueing system with state-dependent service times (e.g., Jain & Smith 1997), and adapt it to a mixed flow of HVs and AVs forming platoons. This queueing system is also known as an $M/G_n/c/c$ queue, where n is the state of the system, defined as the number of vehicles simultaneously driving on a segment of a long highway. In our model, as in Jain and Smith, vehicles arrive individually to the highway according to a Poisson process with rate λ. But unlike Jain & Smith (1997), our vehicles form platoons while traveling on the highway – the formation of these platoons enables us to capture the different travel dynamics of AVs and HVs. In order to analytically model this platooning behavior, we extend the discrete MAP of Alfa & Neuts (1995) to continuous distributions, which enable us to effectively model different service rates within the $M/G_n/c/c$ framework under different policies. For example, under the I policy, in order to capture the effect of vehicle type (AV or HV) on service rate, we use hyperexponential distributions to model interplatoon and intraplatoon headways. Under the D policy, we divide the highway into an AV lane and HV lane(s), modeling headways in these homogeneous lanes as exponential distributions. This enables us to vary the proportion of AVs, and observe the impact on the aggregate performance of the highway. Thus, a key modeling contribution of our paper is bringing the work of Alfa & Neuts (1995) to the work of Jain & Smith (1997), creating a more powerful and flexible hybrid model capable of capturing the flow of heterogeneous vehicles within a highway setting. Although different vehicle types (e.g., trucks, sedans, SUVs, etc.) travel on a highway, we assume for simplicity that all vehicle types are identical. In §1.3.1.1, we first describe the queueing system, and in §1.3.1.2 we explain how we use our MAP to model the formation of platoons. In §1.3.1.3, the impact of platooning on service time is illustrated.

1.3.1.1 The Queueing System

We focus our attention on a segment of length L in a highway with N lanes. The capacity per mile of each lane is called the jam density J, defined as the maximum number of vehicles per mile per lane of the highway; once J is reached flow comes to a jam, at which point vehicles travel at minimal speed. The capacity of the entire highway is $c = J \times L \times N$. We assume a vehicle that finds c other vehicles on the segment upon its arrival turns away, possibly taking an alternative route. In prior literature (e.g., Jain & Smith (1997), Cheah & Smith (1994), Van Woensel & Vandaele (2007)), this assumption is justified by approximating the $M/G_n/c/c$ queueing model using the expansion method. In this method, the blocked vehicles are rerouted to an $M/M/\infty$ queue, where they wait until the $M/G_n/c/c$ has an available server. Jain & Smith (1997) and Cheah & Smith (1994) show that this approximation method yields very close results to those of the $M/G_n/c/c$ model. This assumption also reflects today's reality that drivers may take alternative routes suggested by navigation systems or apps when highways are extremely congested; this is particularly true for congested highways in urban areas.[3]

Service is defined as the travel time of a single vehicle from the beginning of the highway segment to the end of the highway segment. The speed of a vehicle, and hence its travel time, depends on the number of vehicles present on the highway (i.e., the state of the queueing system): a vehicle travels freely on the highway in the absence of other vehicles, but as the highway becomes more crowded, a vehicle tends to drive at a lower speed. We use the state-dependent speed to define the service rate. Let V_n be the mean speed when there are n vehicles on the highway. We assume V_n is decreasing in n, and $V_n = 0$ for $n \geq c + 1$. When only a single vehicle drives on the highway, the mean speed V_1 is called the *free-flow* speed.[4]

As a vehicle enters the highway, it immediately occupies a server and starts receiving service (i.e., there is no waiting time). As a result, the number of servers is equal to the maximum possible number of vehicles that are traveling on the highway, i.e., the capacity c of the highway.

[3] Our model is well-suited to most areas, except rural areas, where alternative routes may not exist. However, there is typically little need for an alternative route in such areas, because rural highways are usually not congested (except when road work or accidents occur).

[4] Ideally, the free-flow speed should be equal to the speed limit. Yet, our data indicate that the free-flow speed is close to, but not equal to the speed limit. This is also observed in prior literature.

1.3.1.2 The Platooning Process

To capture the platooning effect, we use a MAP. Since we analyze a steady state queueing system, platooning is modeled also in steady state. A MAP, which is a Markovian process with an absorbing state, is defined by two $m \times m$ matrices $\boldsymbol{C}_n^0$ and $\boldsymbol{C}_n^1$. The matrix $\boldsymbol{C}_n^0$ (resp., $\boldsymbol{C}_n^1$) is associated with the rate of transitions to non-absorbing states (resp., the absorbing state) when there are n vehicles on the highway. The matrix $\boldsymbol{C}_n = \boldsymbol{C}_n^1 + \boldsymbol{C}_n^0$ is the irreducible generator matrix of the MAP. The steady state distribution of the MAP, $\tilde{\boldsymbol{\pi}}_n$, satisfies $\tilde{\boldsymbol{\pi}}_n \boldsymbol{C}_n = 0$ and $\tilde{\boldsymbol{\pi}}_n \mathbf{1} = 1$, where $\mathbf{1}$ is a vector of ones. The mean of a MAP is calculated as

$$h_n = \frac{1}{\tilde{\boldsymbol{\pi}}_n \boldsymbol{C}_n^1 \mathbf{1}} \text{ (in units of time)}, \tag{1.1}$$

which also represents the mean headway (i.e., the mean *time* gap between the front bumpers of any two consecutive vehicles on the highway) when there are n vehicles on the highway.

To model platooning using a MAP, one needs to specify the distributions of the following three elements: (1) the size of each platoon, (2) the time gap between two consecutive vehicles traveling in the same platoon ("intraplatoon headway"), and (3) the time gap between the last vehicle of one platoon and the first vehicle of the following platoon ("interplatoon headway"). In Appendix A.2, we specify each of these three elements to characterize the matrices $\boldsymbol{C}_n^1$ and $\boldsymbol{C}_n^0$ of the MAP.

1.3.1.3 The Effect of Platooning on a Service Rate

Platooning of vehicles affects the service rate of the queueing system through mean headway. We now derive the relationship between platooning, headway and service rate (vehicle speed).

A highway traffic stream is characterized by three factors: speed, density and flow. Speed or velocity V_n is in miles per unit time, traffic density $k = \frac{n}{NL}$ is defined as the number of vehicles per unit distance, and traffic flow q is defined as the rate (in vehicles per unit time) at which vehicles travel through some designated roadway point. These three measures are related according to $q = V_n k = \frac{nV_n}{NL}$, and the mean headway h_n is equal to the inverse of flow q. Thus,

$$\frac{1}{h_n} = \frac{nV_n}{NL}. \tag{1.2}$$

For each n, once we compute h_n from (1.1) (which depends on platoon characteristics such as platoon size, and interplatoon and intraplatoon headways), we can derive the speed V_n from (1.2).

1.3.2 Models with a Specific Fleet Composition

We adapt the model described in §1.3.1 to a benchmark case and two policies with AVs. In §1.3.2.1 we specify the benchmark case which depicts the current situation where all vehicles on highways are HVs. In §1.3.2.2 we present a model that assigns a specific lane to AVs, the designated-lane (D) policy. In §1.3.2.3 we develop a model in which AVs and HVs are allowed to use any lanes, the integrated (I) policy. We let p denote the proportion of AVs in the latter two models. In the rest of this paper, we use superscripts B, I, D, DA and DH to represent the benchmark case, the I policy, the D policy, and the AV queue and the HV queue of the D policy, respectively.

1.3.2.1 The Benchmark Case

To characterize the platooning process for the benchmark case, we specify the platoon size, the intraplatoon headway, and the interplatoon headway: The platoon size follows a geometric distribution with mean $1/\delta^B$, the intraplatoon headway follows an exponential distribution with mean $1/\xi_n^B$, and the interplatoon headway follows an exponential distribution with mean $1/\psi_n^B$.

By forming the matrices $\boldsymbol{C}_n^0$ and $\boldsymbol{C}_n^1$, as described in Appendix A.2, the mean headway of vehicles in the benchmark case is determined by (1.1) as $h_n^B = \delta^B/\psi_n^B + (1-\delta^B)/\xi_n^B$. Thus, according to (1.2), V_n^B can be represented as a function of the platoon characteristics as follows:

$$V_n^B = \frac{NL}{n} \frac{\xi_n^B \psi_n^B}{\delta^B \xi_n^B + (1-\delta^B)\psi_n^B}. \tag{1.3}$$

1.3.2.2 The D Policy

In this model, a vehicle entering a highway segment is an AV with probability p, and must use the designated lane. HVs use all other lanes. As a result, we can consider two independent queueing systems: an AV queueing system and an HV queueing system.

The HV queue is similar to that of the benchmark case, except it has one fewer lane. The capacity of this queue is equal to $JL(N-1)$, and the arrival rate is $(1-p)\lambda$. As in the benchmark

case, the platoon size follows a geometric distribution with mean $1/\delta^{DH}$, the intraplatoon headway follows an exponential distribution with mean $1/\xi_n^{DH}$, and the interplatoon headway follows an exponential distribution with mean $1/\psi_n^{DH}$. These parameters may be different from those in the benchmark case, since the mean speed in an N-lane highway is usually higher than the mean speed in an (N-1)-lane highway for the same number of vehicles in one mile of the highway.

The AV queue is also modeled as an $M/G_n/c/c$ queueing system described in §1.3.1. In this system, the number of lanes is one, and the capacity is equal to JL, and the arrival rate is $p\lambda$. The platoon size follows a geometric distribution with mean $1/\delta^{DA}$, and the intraplatoon and interplatoon headways follow exponential distributions with mean $1/\xi_n^{DA}$ and $1/\psi_n^{DA}$, respectively.

1.3.2.3 The I Policy

We consider an $M/G_n/c/c$ queueing system with arrival rate λ, where a proportion p of the vehicles are AVs. Under the I policy, a mixed fleet of AVs and HVs form platoons. We characterize the platooning process as follows. First, the platoon size follows a geometric distribution with mean δ^I. Second, the intraplatoon headway follows a hyperexponential distribution with mean $1/\xi_n^I$; the intraplatoon headway depends on the types of vehicles following each other. For example, if an AV follows another AV, since they can communicate with each other and also move/brake more smoothly than HVs, the intraplatoon headway between them is lower than the intraplatoon headway between two HVs. There exist four possible pairs of vehicles: AV-AV (denoted by $i = IAA$), AV-HV ($i = IAH$), HV-AV ($i = IHA$), and HV-HV ($i = IHH$); where X-Y means vehicle X is followed by vehicle Y. Since AVs constitute a proportion p of vehicles, the probability of observing each of those four pairs is p^2, $p(1-p)$, $(1-p)p$, and $(1-p)^2$, respectively. Assuming the intraplatoon headway follows an exponential distribution with rate ξ_n^i for the pair $i \in \{\text{IAA, IAH, IHA, IHH}\}$, the overall intraplatoon headway follows a hyperexponential distribution. By linearity of expectation, the mean intraplatoon headway is calculated as

$$1/\xi_n^I = p^2/\xi_n^{IAA} + p(1-p)/\xi_n^{IAH} + (1-p)p/\xi_n^{IHA} + (1-p)^2/\xi_n^{IHH} \text{ for } n = 1, 2, \cdots, c. \quad (1.4)$$

Lastly, similar to the intraplatoon headway, the interplatoon headway also follows a hyperexponential distribution with parameters ψ_n^i, where $i \in \{\text{IAA, IAH, IHA, IHH}\}$. The mean interplatoon

headway is as follows

$$1/\psi_n^I = p^2/\psi_n^{IAA} + p(1-p)/\psi_n^{IAH} + (1-p)p/\psi_n^{IHA} + (1-p)^2/\psi_n^{IHH} \text{ for } n = 1, 2, \cdots, c. \quad (1.5)$$

Having the platoon characteristics specified, similar to the benchmark case, we are able to calculate the mean headway h_n^I, and then represent V_n^I as a function of h_n^I.

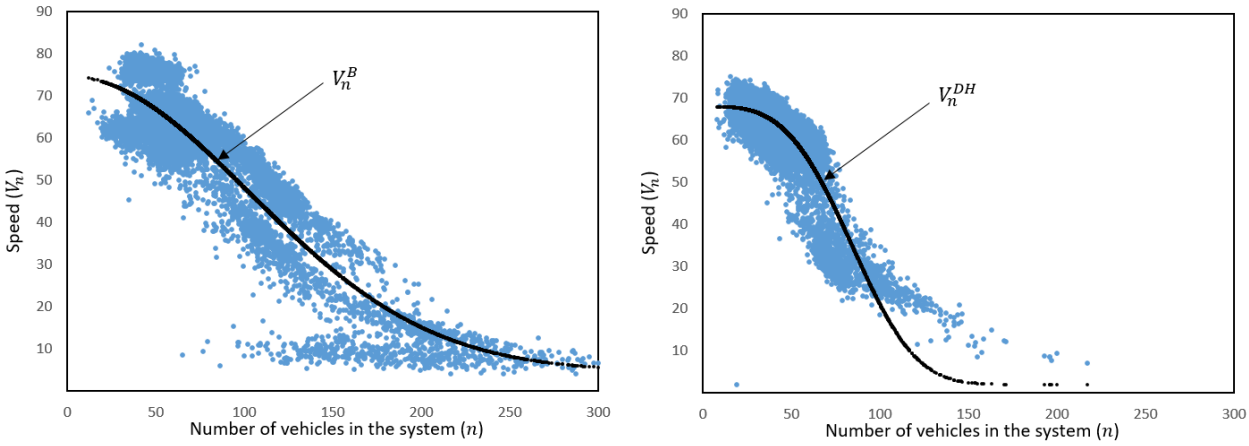
1.4 Model Calibration

In this section we calibrate our queueing model to data. The arrival rate to a highway can differ depending on the highway location, but the speed of vehicles primarily depends on the number of vehicles currently driving on the highway and the headway between them, so we focus on estimating the steady state speed of vehicles in our models. Without loss of generality, we assume L is equal to one mile. Then, we estimate V_n for $n = 1, 2, \cdots, c$, where $c = NJL = NJ$ is the capacity of the highway segment with N lanes and a jam density of J. The value of J is typically between 185 and 250 (Holtzman & Goodman (2012) and Wang et al. (2010)), in our numerical analysis, we use $J = 185$. (Our results are robust to different values of J.) Thus, the capacities of the benchmark case and I policy with $N = 3$ are equal to 555, the capacity of the HV queue under the D policy with $N = 2$ is 370, and the capacity of the AV queue under the D policy with $N = 1$ is 185.

In the rest of this section, we first use the data collected from a highway in Arizona to estimate V_n in the benchmark case. Then we discuss the speed estimation for each of the HV queue and the AV queue under the D policy. For the estimation of V_n in the HV queue, we also use the data from Arizona. For the AV queue we propose a procedure to estimate V_n using several parameters that reflect an AV's driving performance; since a designated lane for AVs has yet to be implemented in reality, V_n cannot be directly estimated from data. Lastly, V_n is estimated under the I policy using a procedure similar to the AV queue under the D policy.

1.4.1 The Benchmark Case

We estimate the state-dependent speed V_n^B based on the data from the Arizona Department of Transportation. Our data include about $10,000$ instances of 5-minute average volume and vehicle



(a) (b)

Figure 1.1: State-dependent speed data and fitted curves: (a) a highway with three lanes ($R^2 = 85\%$), and (b) a highway with two lanes ($R^2 = 76\%$).

speeds collected from segments of Interstate 10 (I-10) with three lanes ($N = 3$) in January 2017. Figure 1.1(a) shows a scatter plot of this data as well as its fitted curve:

$$V_n^B = 70e^{-\frac{n^2}{21{,}049}} + 4.7 \text{ (miles/hour) for } n = 1, 2, \cdots, 555. \tag{1.6}$$

Appendix A.4.1 provides more information about how we estimate this curve.

1.4.2 The D Policy

As mentioned in §1.3.2.3, under this policy the highway is divided into two queues: (1) the HV queue with two lanes, and (2) the AV queue with one lane.

1.4.2.1 The HV Queue

This queue is similar to the benchmark case, except it has one fewer lane. Using about 16,000 data points collected from state route 101 in Arizona with two lanes[5] in January 2017, we estimate the state-dependent speed of the HV queue V_n^{DH} (see Figure 1.1(b)), as follows:

$$V_n^{DH} = 66e^{-\frac{n^{3.4}}{5{,}215{,}902}} + 2 \text{ (miles/hour) for } n = 1, 2, \cdots, 370. \tag{1.7}$$

[5]All the segments of I-10 that we use in our benchmark case have three lanes. Thus, we use data from the 2-lane segments of state route 101 which has the same speed limit as the 3-lane segments of I-10.

By comparing Figures 1.1(a) and 1.1(b), we observe that $V_n^{DH} < V_n^B$ for any given n. Intuitively, when the same number of vehicles travel on two segments, with different number of lanes, the vehicles are able to drive faster on the segment with more lanes.

1.4.2.2 The AV Queue

We use a MAP to estimate the state-dependent speed of the AV queue, V_n^{DA}. Equation (1.2) relates V_n^{DA} to the mean headway h_n^{DA} of vehicles derived from the MAP. In this queue N is equal to one and L is one mile, so $V_n^{DA} = \frac{1}{nh_n^{DA}}$, where h_n^{DA} is derived from equation (1.1); it is a function of mean platoon size, mean intraplatoon headway, and mean interplatoon headway. Therefore, to estimate V_n^{DA}, we need to estimate these three parameters.

First, we specify the mean platoon size. All the previous papers that consider platooning of AVs assume a fixed platoon size: for example, Amoozadeh et al. (2015) and Liu et al. (2018) consider platoons of size 10, and Zhao & Sun (2013) consider platoons of size 6. We take into account randomness in platoon sizes by using a geometric distribution with the same mean value as in Amoozadeh et al. (2015) and Liu et al. (2018), i.e., $1/\delta^{DA} = 10$ vehicles. Appendix A.4.2 contains more details about the estimation of a platoon size distribution.

Second, following Amoozadeh et al. (2015), Vander Werf et al. (2002), and Zhao & Sun (2013), we set the mean intraplatoon headway of AVs equal to 0.55 seconds, i.e., $1/\xi_n^{DA} = 0.55$ seconds for all $n = 1, 2, \cdots, 185$, because the intraplatoon headway of AVs does not depend on congestion or speed[6] (Siciliano & Khatib 2016).

Lastly, we estimate the mean interplatoon headway of AVs. According to Guzzella & Kiencke (1995) and Bergenhem et al. (2012), the interplatoon headway of AVs is set equal to the safe stopping time to avoid chain-reaction crashes. In other words, the interplatoon headway is set such that if one platoon of AVs crashes, the following platoon has enough time to stop before hitting the crashed platoon. We estimate this value by using data from National Highway Traffic Safety

[6]This property also holds for the *intraplatoon* headway of HVs. According to Virginia DMV (2016) and Tientrakool et al. (2011), the intraplatoon headway for HVs reflects the reaction time of human drivers. This reaction time depends neither on speed nor on the congestion caused by a high number of vehicles in a highway segment. The *interplatoon* headway of both AVs and HVs, however, depends on n. This headway is what vehicles maintain to avoid cascade crashes between platoons, so it depends on how fast they are moving and how congested the highway is.

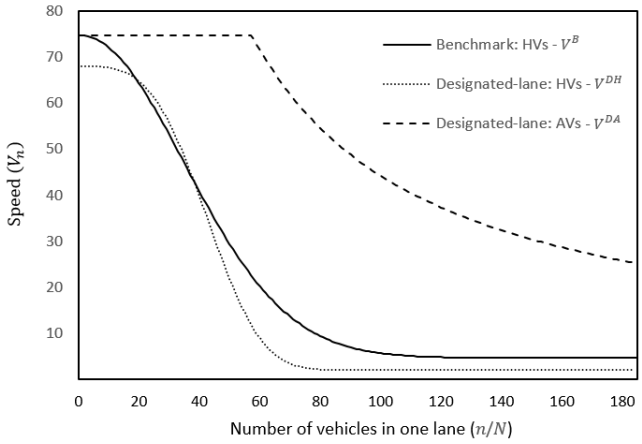


Figure 1.2: Comparison of the state-dependent speeds: the benchmark case vs. the D policy.
Note. The capacities of the AV queue, the HV queue, and the benchmark case are different (185, 370, and 555, respectively), and the speed in each of these queues is shown as a function of the number of vehicles per lane.

Administration (NHTSA 2015) as follows (see Appendix A.4.3 for more details):

$$1/\psi_n^{DA} = 3,600(0.001 - \frac{0.006}{V_n^{DA}}) \text{ (seconds) for } n = 1, 2, \cdots, 185, \tag{1.8}$$

which is less than the mean intraplatoon headway of AVs (i.e. $1/\xi_n^{DA} = 0.55$ seconds).

Having characterized the platooning process, we are able to calculate the mean headway between vehicles. For a MAP with geometric(δ^{DA}) platoon size, exp(ξ_n^{DA}) intraplatoon headway, and exp(ψ_n^{DA}) interplatoon headway, by equation (1.1) we get $h_n^{DA} = \frac{\delta^{DA}}{\psi_n^{DA}} + \frac{(1-\delta^{DA})}{\xi_n^{DA}}$ (see Appendix A.2 for details); substituting the values of δ^{DA}, ξ_n^{DA}, and ψ_n^{DA} into $V_n^{DA} = \frac{1}{nh_n^{DA}}$, we obtain, after simplifications, $V_n^{DA} = \min(74.7, \frac{3,600+2.16n}{0.855n})$, where 74.7 miles per hour is the free flow speed under the D policy.[7] Figure 1.2 illustrates V_n^{DA} and compares it with V_n^B and V_n^{DH}. We observe that the speed of AVs is always higher than that of HVs. For example, when $n/N = 185$, HVs move slowly at about 4 miles per hour in the benchmark case, but AVs have a smooth flow moving at 25 miles per hour. Based on our analysis, we assume that the jam speed in an N-lane highway, V_{Nc}^B, is higher than that in an $(N-1)$-lane highway, $V_{(N-1)c}^{DH}$ in the rest of this paper.[8]

[7]The maximum speed of AVs must be capped, because otherwise the speed keeps increasing as n decreases. For ease of comparison, we set the free flow speed of the highway under the D and I policies equal to that of the benchmark case (which is obtained by setting $n = 1$ in (1.6)).

[8]Figure 1.2 also compares V_n^B and V_n^{DH}. The speed pattern of a 3-lane highway (V_n^B) differs slightly from that of a 2-lane highway (V_n^{DH}). When the number of vehicles per lane is very low (i.e., $n/N \leq 15$), vehicles drive at the free-flow speed, which is higher for a highway with more lanes. At moderately low values of n/N (i.e., $15 < n/N < 55$), the highways are not congested and vehicles are able to still drive fast, so there is not much difference between 3-lane and 2-lane highways. As the highways become more congested (i.e., $n/N \geq 55$), speed decreases on both highways,

1.4.3 The I Policy

Similar to the AV queue under the D policy, we estimate the state-dependent speed V_n^I by employing the mean headway of vehicles derived from a MAP. For any value of $p \in [0,1]$, we characterize the MAP by first specifying the mean platoon size, and then jointly describing the mean intraplatoon headway and the mean interplatoon headway under this policy.

We first estimate the mean platoon size as a function of p. As before, we assume that the platoon size follows a geometric distribution with mean $1/\delta^I$. When all vehicles are AVs (i.e., $p = 1$), as in the AV queue under the D policy, we use $1/\delta^I = 10$ vehicles. On the other hand, when all vehicles are HVs (i.e., $p = 0$), we use $1/\delta^I = 1.5$ vehicles, which is obtained from the data used for the benchmark case. For any $p \in (0,1)$, we assume $1/\delta^I = 3/(2-1.7p)$, so that the mean size of platoons increases with p.[9]

Next we proceed to describe the mean intraplatoon and interplatoon headways. A pair of consecutive vehicles can be of four different types: AV-AV, HV-AV, HV-HV, and AV-HV. For each of these types, we estimate the mean intraplatoon headway and the state-dependent mean interplatoon headway.

- AV-AV: As in the AV queue in the D policy, we set the mean intraplatoon headway $1/\xi_n^{IAA} = 0.55$ seconds, and estimate the mean interplatoon headway $1/\psi_n^{IAA}$ from equations (1.1), (1.2) and (1.8) as follows: $1/\psi_n^{IAA} = \frac{3{,}600N - 6n(1-\delta^{IAA})/\xi_n^{IAA}}{1{,}000N + 6n\delta^{IAA}}$ seconds. Substituting $N = 3$ lanes, $1/\delta^{IAA} = 10$ vehicles, and $1/\xi_n^{IAA} = 0.55$ seconds, $1/\psi_n^{IAA} = \frac{10{,}800 - 2.97n}{3{,}000 + 0.6n}$ seconds for $n = 1, 2, \cdots, 555$.

- HV-AV: Following Zhao & Sun (2013), we set the mean intraplatoon headway $1/\xi_n^{IHA} = 1.4$ seconds: Due to lack of communication between AVs and HVs, an AV maintains a longer intraplatoon headway from an HV than another AV. Similar to the AV-AV pair, we can then estimate the interplatoon headway of this pair as $\psi_n^{IHA} = \frac{10{,}800 - 7.56n}{3{,}000 + 0.6n}$ seconds for $n = 1, 2, \cdots, 555$.

but more so on the 2-lane highway. According to Transportation Research Board (2000), enhanced maneuverability in a 3-lane highway compared to a 2-lane highway tends to increase the average speed of vehicles. Thus, vehicles drive faster on average at a congested highway with more lanes. For further discussion, see Appendix A.5.2.

[9] Note that δ is the probability that a vehicle forms a new platoon instead of joining the very last one. This probability is equal to $1/1.5 = 2/3$ for an HV, and $1/10$ for an AV. Thus, under the I policy, when the proportion of AVs is p, the probability of forming a new platoon is $\delta^I = \frac{2}{3}(1-p) + \frac{1}{10}p = \frac{2-1.7p}{3}$, and $1/\delta^I = \frac{3}{2-1.7p}$.

- HV-HV: Tientrakool et al. (2011) state that on average two HVs maintain 1.1 seconds of intraplatoon headway in practice, so we set $1/\xi_n^{IHH} = 1.1$ seconds for $n = 1, 2, \cdots, 555$. Note that this value is lower than that of an HV-AV pair. The mean interplatoon headway of HVs is derived by rearranging equation (1.3) as $1/\psi_n^{IHH} = \frac{\xi_n^{IHH} NL - (1-\delta^{IHH}) n V_n^B}{n V_n^B \delta^{IHH} \xi_n^{IHH}}$ seconds.[10] Substituting $N = 3$ lanes, $L = 1$ mile, $1/\delta^{IHH} = 1.5$ vehicles, $1/\xi_n^{IHH} = 1.1$ seconds, and $V_n^B = 70e^{-\frac{n^2}{21,049}} + 4.7$ (miles/hour), we obtain

$$1/\psi_n^{IHH} = \frac{10,800 - 0.55n(46.67e^{-\frac{n^2}{21,049}} + 3.13)}{n(46.67e^{-\frac{n^2}{21,049}} + 3.13)} \text{ (seconds)} .$$

- AV-HV: Following the literature (e.g., Zhao & Sun (2013)), we assume that an HV maintains its intraplatoon and interplatoon headways independently of the type of vehicle it is following, so that $1/\xi_n^{IAH} = 1/\xi_n^{IHH} = 1.1$ seconds, and $1/\psi_n^{IAH} = 1/\psi_n^{IHH}$.

As mentioned in §1.3.2.3, the intraplatoon headway of vehicles in the I policy follows a hyper-exponential distribution. Substituting the values of the mean intraplatoon headways for each of the vehicle pairs into equation (1.4), we obtain the mean intraplatoon headway as follows:

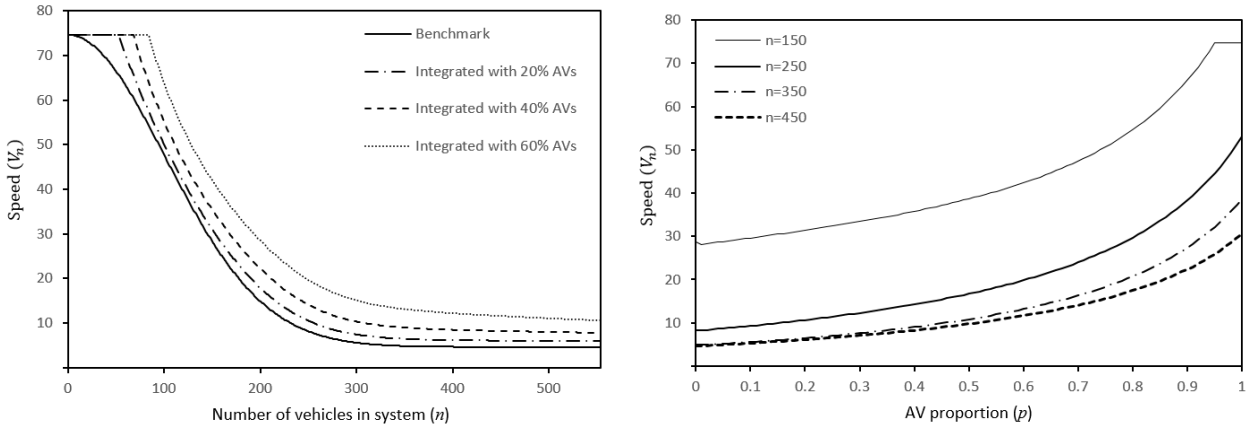
$$1/\xi_n^I = 0.55p^2 + 1.4p(1-p) + 1.1(1-p) \text{ (seconds) for } n = 1, 2, \cdots, 555, \text{ and } p \in [0, 1]. \quad (1.9)$$

Similarly, by substituting the values of mean interplatoon headways for each of the vehicle pairs into equation (1.5), after some simplification, we have the following for $n = 1, 2, \cdots, 555$, and $p \in [0, 1]$:

$$1/\psi_n^I = p\frac{(10,800 - 7.56n) + 4.59np}{3,000 + 0.6n} + (1-p)\frac{10,800 - 0.55n(46.67e^{-\frac{n^2}{21,049}} + 3.13)}{n(46.67e^{-\frac{n^2}{21,049}} + 3.13)} \text{ (seconds)} . \quad (1.10)$$

Finally, substituting δ^I, ξ_n^I, and ψ_n^I into $V_n^I(p) = N/nh_n^I = \frac{N\xi_n^I\psi_n^I}{n(\delta^I\xi_n^I + (1-\delta^I)\psi_n^I)}$ (miles/hour), we have

[10] Note that this depends on how human drivers determine their speed based on the characteristics of the highway such as the number of lanes (N) and the length of the highway segment (L), whereas that of AVs is determined by the safe stopping time.



(a) (b)

Figure 1.3: State-dependent speed under the I policy as a function of: (a) the number of vehicles in system (n), and (b) the proportion of AVs (p).

the following for $n = 1, 2, \cdots, 555$, and $p \in [0, 1]$:

$$V_n^I(p) = \min\{74.7, \frac{10,800N/[n(1+1.7p)]}{\frac{2-1.7p}{1+1.7p}[\frac{(10,800-7.56n)p+4.59np^2}{3,000+0.6n} + \frac{10,800(1-p)}{n(46.67e^{-\frac{n^2}{21,049}}+3.13)} - 0.55(1-p)] + [0.55p^2 + 1.4p(1-p) + 1.1(1-p)]}\}. \tag{1.11}$$

Figure 1.3(a) illustrates V_n^B and $V_n^I(p)$ at $p = 20\%, 40\%$, and 60%, and shows that, for a fixed p, $V_n^I(p)$ has a similar shape as V_n^{DA}. As illustrated in Figure 1.3(b), as the proportion of AVs (p) increases, $V_n^I(p)$ tends to increase. We can explain this result by examining how the weighted mean intraplatoon and interplatoon headways change with p; see Appendix A.5.3 for details.

1.5 Analysis

In §1.5.1 and §1.5.2 we analyze our queueing model under the D policy and the I policy, respectively. In §1.5.3 we compare the performances of these policies. In §1.5.4 we discuss the robustness of our results. All proofs are presented in Appendix A.3.

For the comparison between the two policies, we use two quality of service (QoS) measures: throughput (θ) and mean travel time (W). To compute these measures, we use the steady state

distribution of an $M/G_n/c/c$ queueing model, π_n, which is derived by Cheah & Smith (1994):

$$\pi_n = \big(\frac{(\lambda L)^n}{n! V_n \cdots V_1}\big)\pi_0, \text{ where } \pi_0 = (1 + \sum_{n=1}^{c} \frac{(\lambda L)^n}{n! V_n \cdots V_1})^{-1}. \tag{1.12}$$

For a queueing model with finite capacity, throughput is the rate at which vehicles exit the queue, also equal to the effective arrival rate. In our model, a proportion π_c of vehicles find c other vehicles on the highway upon their arrival, and turn away. Thus, by PASTA (Poisson Arrivals See Time Averages), the effective arrival rate is $\theta = \lambda(1 - \pi_c)$. Highway throughput is an important measure for urban designers. Mean travel time of a single vehicle (also known as mean response time) is obtained by Little's law as follows:

$$W = (\sum_{n=1}^{c} n\pi_n)/\theta. \tag{1.13}$$

This measure mostly concerns individual users. As in the previous sections, the superscript $i \in \{B, DH, DA, D, I\}$ in π_n^i, θ^i, or W^i represents different policies or queues. For example, W^B represents the mean travel time of the benchmark case.

In our analysis, a highway is considered to be *highly (heavily) loaded*, if $\lambda \geq \lambda_{jam}^i$, $i \in \{B, DA, DH, I\}$; see Appendix A.5.4 for more details about λ_{jam}. In this case, the arrival rate to the highway is so high that the highway tends to be full. We call V_{Nc}^B, $V_{(N-1)c}^{DH}$, V_c^{DA}, and $V_{Nc}^I(p)$ *jam speeds*, which are the speed of vehicles when a highway with N lanes each having c capacity is full, in the benchmark case, the HV queue under the D policy, the AV queue under the D policy, and the I policy with AVs constituting a proportion p of arrivals, respectively. In addition, we call the throughput of a jammed highway *jam throughput*, which is equal to the product of the capacity of the system and the jam speeds. For example, in the benchmark case, the rate at which a single vehicle leaves the highway is $\mu_{Nc}^B = V_{Nc}^B$, hence the total rate at which vehicles exit the highway, the jam throughput, is NcV_{Nc}^B vehicles per hour.

Our analysis focuses primarily on high values of λ, because, in general, a policy-maker is focused on improving the performance of a highway when it is heavily loaded. We complement our analytical results with a numerical study that also examines lighter loads.

1.5.1 The D Policy

Under the D policy the highway segment is *split* into two separate queueing systems, as opposed to the benchmark case in which all lanes of the highway are *pooled* in one queueing system. In general, a pooled server is more efficient than multiple servers of the same total capacity. However, in our setting, although the service capacity is divided into two queues, the service rate in the AV queue is higher than that in the HV queue (see §1.4.2). Thus, there is a trade-off between pooling the servers and increasing the service rate by designating one lane to AVs. As a result, the pooled server in the benchmark case can be inferior to the split servers under the D policy. The next proposition, which holds for any values of the model parameters described in §1.3, presents the condition under which each of these factors outweighs the other in terms of W and θ.[11]

Proposition 1.1. *a) There exist* $\underline{\lambda}^{(D,W)}$ *and* $\bar{\lambda}^{(D,W)}$ *such that for* $\lambda \leq \underline{\lambda}^{(D,W)}$, $W^D(p) \geq W^B$ *for* $p \in [0,1]$*; for* $\lambda \geq \bar{\lambda}^{(D,W)}$, $W^D(p) \leq W^B$ *if and only if* $p \geq p^{(D,W)} = \dfrac{V_c^{DA}V_{Nc}^B - V_c^{DA}V_{(N-1)c}^{DH}}{V_c^{DA}V_{Nc}^B - V_{Nc}^B V_{(N-1)c}^{DH}}$. *b)* $\theta^D(p) \geq \theta^B$ *if and only if* $p\pi_c^{DA} + (1-p)\pi_{(N-1)c}^{DH} \leq \pi_{Nc}^B$. *This can be further simplified for* $\lambda \geq \lambda^{(D,\theta)} \equiv \min_p \max\{\lambda_{jam}^{DA}/p, \lambda_{jam}^{DH}/(1-p)\}$, *when* θ^{DH} *and* θ^{DA} *are increasing in* λ*:* $\theta^D(p) \geq \theta^B$ *if and only if* $\underline{p}^{(D,\theta)} \leq p \leq \bar{p}^{(D,\theta)}$, *where* $\underline{p}^{(D,\theta)} = \{\frac{NcV_{Nc}^B - (N-1)cV_{(N-1)c}^{DH}}{L\lambda[1-\pi_c^{DA}(\underline{p}^{(D,\theta)})]}\}^+$ *and* $1-\bar{p}^{(D,\theta)} = \{\frac{NcV_{Nc}^B - cV_c^{DA}}{L\lambda[1-\pi_{(N-1)c}^{DH}(1-\bar{p}^{(D,\theta)})]}\}^+$.[12]

Proposition 1.1(a) indicates that the D policy decreases W compared to the benchmark case only when both the arrival rate λ and the AV proportion p are high. When the highway is lightly loaded with $\lambda \leq \underline{\lambda}^{(D,W)}$, designating a lane to AVs slows down vehicles in $(N-1)$ lanes of the highway segment, while increasing the speed of vehicles in the AV lane only moderately. Even for a heavily loaded highway with $\lambda \geq \bar{\lambda}^{(D,W)}$, dedicating a lane to AVs does not necessarily reduce the mean travel time unless AVs constitute at least a proportion $p^{(D,W)}$ of vehicles; the threshold $p^{(D,W)}$ depends on the jam speed of vehicles in each queue as well as the characteristics of the highway such as the number of lanes and the capacity of each lane. This result cautions against

[11] Throughout this chapter, for thresholds of λ and p, we use superscripts (i,j) for $i \in \{D, I, DI\}$ and $j \in \{W, \theta, W\theta\}$, where D, I, and DI represent the D policy, the I policy, and comparison between D and I policies, respectively; and W, θ, and $W\theta$ indicate thresholds for W, θ, and comparison between W and θ, respectively. We also use underscore and overscore to indicate smaller and larger thresholds, respectively. Table A2 summarizes the notation used in this section.

[12] For the AV queue, θ^{DA} is increasing in λ if nV_n^{DA} is increasing in n (see Lemma A.2 in Appendix A.3). This assumption guarantees that V_n^{DA} decreases in n no faster than linearly. Figure 16(a) in Appendix A.5.4 shows that this property holds for our calibrated model. Similarly, in Lemma A.2, we derive the condition for θ^{DH} to be increasing in λ. This condition also holds for our calibrated model.

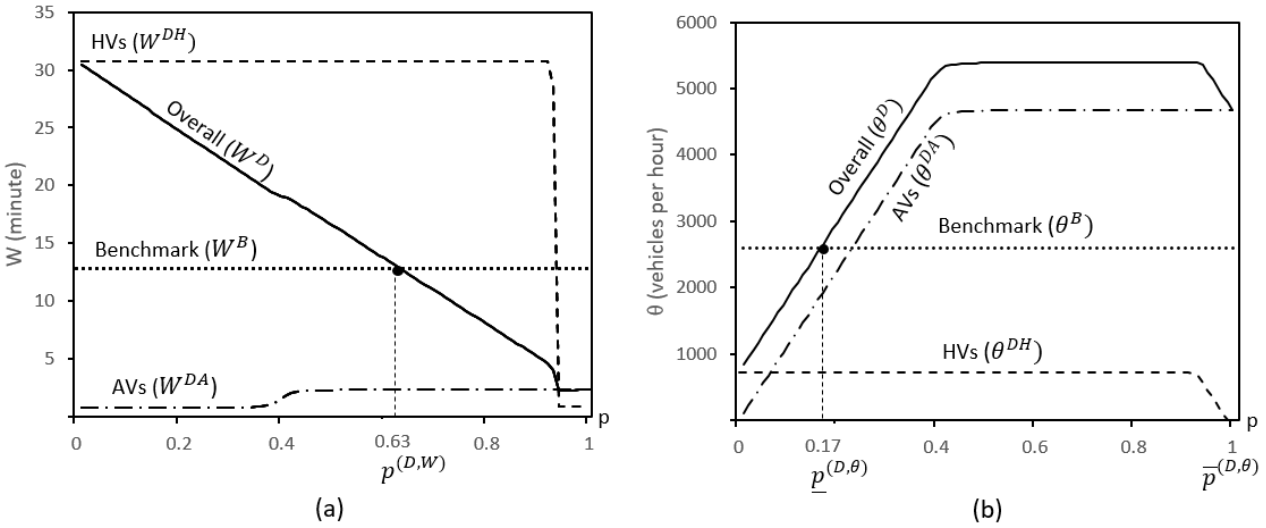


Figure 1.4: QoS measures for the D policy when $\lambda = 11,342$ vehicles per hour: (a) mean travel time, and (b) throughput.
Note. "overall" indicates the overall performance of AVs and HVs under the D policy.

adopting this policy too early, despite recent industry proposals (e.g., Bierstedt et al. 2014 and Aguilar 2018).

We illustrate Proposition 1.1(a) by considering two different values of λ in the calibrated model in §1.4.2: the mean arrival rate of $2,217$ vehicles per hour and the maximum arrival rate of $11,342$ vehicles per hour to I-10 with three lanes in 2016.[13] The former is considered to be a light load, and the latter is a heavy load, because $\underline{\lambda}^{(D,W)} = \overline{\lambda}^{(D,W)} = 2,510$ vehicles per hour.[14] When the highway is lightly loaded with $\lambda = 2,217$, we observe that the D policy can increase W (see Appendix A.5.1 for a discussion of this case).

For $\lambda = 11,342$ vehicles per hour, if AVs constitute at least $p^{(D,W)} = 0.63$ of vehicles, dedicating one out of three lanes to AVs leads to a lower overall mean travel time, W^D, than that of the benchmark case, $W^B = 13$ minutes (see Figure 1.4(a)). The overall mean travel time under the D policy, $W^D = pW^{DA} + (1-p)W^{DH}$, is a weighted average of mean travel time of the AV queue, W^{DA}, and that of the HV queue, W^{DH}. To understand the effect of p on W^D, we discuss its effect on W^{DH} and W^{DA}. First, when $p \leq 0.93$, W^{DH} is substantially higher than W^B, because the HV queue has one fewer lane than the benchmark case. When $p \geq 0.93$, W^{DH} drops significantly,

[13] According to Varaiya (2005), a highway has an ideal throughput of $2,000$ vehicles per hour per lane, so a highway with an arrival rate higher than $2,000 \times N$ vehicles per hour is considered to be highly loaded.

[14] Although in all our numerical expriments $\underline{\lambda}^{(D,W)} = \overline{\lambda}^{(D,W)}$, it is theoretically possible that $\underline{\lambda}^{(D,W)} < \overline{\lambda}^{(D,W)}$, as shown in the proof of Proposition 1.1(a). When $\underline{\lambda}^{(D,W)} < \lambda < \overline{\lambda}^{(D,W)}$, unlike the highly loaded case, $W^D(p)$ depends on V_n^{DH} and V_n^{DA} for all values of n, and $W^D(p)$ can be lower or higher than W^B.

because the arrival rate to the HV queue becomes very low. (Appendix A.5.4 discusses the phase-change-type behavior of W^{DH} at $p = 0.93$.) Second, as p increases, W^{DA} increases from 0.8 minutes to 2.3 minutes; both the magnitude and change of W^{DA} are negligible compared to W^{DH}. Finally, increasing p raises the weight of W^{DA}, and reduces the weight of W^{DH}. Therefore, as p increases, W^D decreases. Moreover, since W^{DH} is significantly higher than W^B, more than 63% AVs are needed to bring W^D below W^B. To have a tangible reduction in W^D, even more AVs are needed; for example, to get 50% reduction in W^B, 80% of vehicles need to be autonomous.

Proposition 1.1(b) provides conditions under which adding AVs to the highway under the D policy improves the throughput over the benchmark case. For any value of λ, the proposition first presents the general condition (i.e., $p\pi_c^{DA} + (1-p)\pi_{(N-1)c}^{DH} \leq \pi_{Nc}^{B}$) that indicates having AVs on the highway under the D policy improves the throughput, if and only if the highway blocks fewer vehicles under this policy than in the benchmark case. For a highly loaded highway with $\lambda \geq \lambda^{(D,\theta)}$, we can further simplify the condition in terms of the proportion of AVs (p). First, when $p < \underline{p}^{(D,\theta)}$, the throughput under the D policy, $\theta^D(p)$, is less than that of the benchmark, θ^B, because both the benchmark case and the HV queue under the D policy are heavily loaded, but the AV lane may still flow relatively freely. In this case, the difference between the jam throughput of the benchmark case and that of the HV queue under the D policy (i.e., $NcV_{Nc}^{B} - (N-1)cV_{(N-1)c}^{DH}$) is larger than the throughput of the AV queue (i.e., $p\lambda(1-\pi_c^{DA})$), so the AV queue's throughput cannot offset this difference. Second, when p is moderate, i.e., $\underline{p}^{(D,\theta)} \leq p \leq \bar{p}^{(D,\theta)}$, $\theta^D(p) > \theta^B$, because under the D policy the highway load is balanced between the HV queue and the AV queue. Finally, when $p > \bar{p}^{(D,\theta)}$, the HV queue is not highly loaded, and the throughput of this queue is lower than the difference between the jam throughput of the benchmark case and that of the AV queue.

To illustrate, we consider the same two values of λ, one below and one above $\lambda^{(D,\theta)} = 2,510$ vehicles per hour. At $\lambda = 2,217$ vehicles per hour, the benchmark case has enough capacity for all the vehicles that enter the highway, and no vehicle is blocked. Thus, assigning a lane to AVs may even reduce θ when p is low (see Appendix A.5.1). However, when the highway is heavily loaded, the D policy is able to improve θ. As Figure 1.4(b) shows, at $\lambda = 11,342$ vehicles per hour, the benchmark case has reached its maximum jam throughput ($555V_{555}^{B}/3 = 2590/3 = 863$ vehicles per hour per lane), but the AV queue has not; the maximum throughput of the AV lane is about five times the throughput of a lane in the benchmark case (i.e., $185V_{185}^{DA} = 4,675$ vs. 863

vehicles per hour per lane). When $p \geq \underline{p}^{(D,\theta)} = 0.17$, the increase in the throughput of the lane dedicated to AVs compensates for the decrease in the throughput of the HV lanes. In addition, when $0.43 \leq p \leq 0.93$ the overall throughput of the highway segment becomes about twice that of the benchmark case (i.e., $5,396$ vs. $2,590$ vehicles per hour). This is same as the ratio of the jam throughput under the D policy to that of the benchmark case (i.e., $\frac{V_c^{DA}+(N-1)V_{(N-1)c}^{DH}}{NV_{Nc}^{B}} = 2.09$). When $p \geq 0.93$, the HV queue is no longer highly-loaded, and throughput of this queue decreases as its arrival rate, $(1-p)\lambda$, decreases. Eventually, as p approaches 1, θ^D converges to the jam throughput of the AV lane (i.e., cV_c^{DA}). Since the jam throughput of the designated lane alone is higher than that of the benchmark case (see Figure 1.4(b)), when $p \geq \underline{p}^{(D,\theta)}$, $\theta^D(p)$ does not go below θ^B, hence $\bar{p}^{(D,\theta)} = 1$.

Parts (a) and (b) of Proposition 1.1 together show that the effect of employing the D policy on throughput is not always the same as that on mean travel time. Corollary 1.1 presents the condition that determines which metric improves first, by comparing the thresholds $p^{(D,W)}$, $\underline{p}^{(D,\theta)}$, and $\bar{p}^{(D,\theta)}$.

Corollary 1.1. *(a)* $p^{(D,W)} \geq \underline{p}^{(D,\theta)}$*, if* $\lambda \geq \max\{\bar{\lambda}^{(D,W\theta)}, \bar{\lambda}^{(D,W)}, \lambda^{(D,\theta)}\}$*, where*

$$\bar{\lambda}^{(D,W\theta)} = \frac{V_{Nc}^{B}(V_c^{DA} - V_{(N-1)c}^{DH})[NcV_{Nc}^{B} - (N-1)cV_{(N-1)c}^{DH}]}{LV_c^{DA}(V_{Nc}^{B} - V_{(N-1)c}^{DH})[1 - \pi_c^{DA}(\underline{p}^{(D,\theta)})]}.$$

(b) $p^{(D,W)} \leq \bar{p}^{(D,\theta)}$*, if* $\lambda \geq \max\{\underline{\lambda}^{(D,W\theta)}, \bar{\lambda}^{(D,W)}, \lambda^{(D,\theta)}\}$*, where*

$$\underline{\lambda}^{(D,W\theta)} = \frac{V_{Nc}^{B}(V_c^{DA} - V_{(N-1)c}^{DH})[NcV_{Nc}^{B} - cV_c^{DA}]}{L\{\pi_{(N-1)c}^{DH}(1-\bar{p}^{(D,\theta)})V_{Nc}^{B}[V_c^{DA} - V_{(N-1)c}^{DH}] + V_{(N-1)c}^{DH}[V_c^{DA} - V_{Nc}^{B}]\}}.$$

Corollary 1.1(a) indicates that if the arrival rate to the highway is high, then the D policy improves θ over the benchmark case before improving W; i.e., $0.63 = p^{(D,W)} \geq \underline{p}^{(D,\theta)} = 0.17$ in Figure 1.4. The intuition behind this result is that W^D is the weighted average of W^{DH} and W^{DA}, while θ^D is the sum of θ^{DH} and θ^{DA}. Consequently, when $(1-p)$ (i.e., the weight of W^{DH} in W) is high, the low performance of the HV queue has a more significant impact on W than θ. In addition, Corollary 1.1(b) shows that there exists an interval for p (e.g., $0.63 \leq p \leq 1$ in Figure 1.4), in which the D policy simultaneously improves θ and W over the benchmark case.

In a nutshell, the performance of the D policy depends significantly on arrival rate and proportion of AVs. Although this policy has the potential to reduce mean travel time as well as throughput, this requires a substantial proportion of vehicles to be AVs, even for a congested highway. This finding contributes to the AV literature which has neglected the role of arrival rate. For example, in the simulation study performed by Liu et al. (2018), at a fixed value of λ, one lane is dedicated to AVs when $p = 0.4$, resulting in about 24% improvement in θ. However, Liu et al. (2018) is silent on whether the benchmark case without AVs has reached its maximum throughput at this value of λ, and how throughput θ would change with different values of λ.

1.5.2 The I Policy

This section analyzes the effect of AVs on highway congestion under the I policy, and compares its performance with that of the benchmark case. We first present Proposition 1.2, which compares this policy with the benchmark case for highly loaded highways.

Proposition 1.2. *a) For any given $p \in [0,1]$, there exists $\lambda^{(I,W)} \geq 0$ such that for $\lambda \geq \lambda^{(I,W)}$, the I policy has a lower mean travel time W than the benchmark case, if and only if $V^I_{Nc}(p) \geq V^B_{Nc}$, or equivalently*

$$\delta^I\left(\frac{2-p}{\psi^{IHH}_{Nc}} - \frac{1-p}{\psi^{IHA}_{Nc}} - \frac{1-p}{\psi^{IAH}_{Nc}} - \frac{p}{\psi^{IAA}_{Nc}}\right) + (1-\delta^I)\left(\frac{2-p}{\xi^{IHH}_{Nc}} - \frac{1-p}{\xi^{IHA}_{Nc}} - \frac{1-p}{\xi^{IAH}_{Nc}} - \frac{p}{\xi^{IAA}_{Nc}}\right) \geq \left(\frac{\delta^{IAA}}{\psi^{IHH}_{Nc}} + \frac{1-\delta^{IAA}}{\xi^{IHH}_{Nc}}\right) - \left(\frac{\delta^{IHH}}{\psi^{IHH}_{Nc}} + \frac{1-\delta^{IHH}}{\xi^{IHH}_{Nc}}\right). \tag{1.14}$$

b) For any given $p \in [0,1]$, there exists $\lambda^{(I,\theta)} \geq 0$ such that for $\lambda \geq \lambda^{(I,\theta)}$, the I policy has a higher throughput θ than the benchmark case, if and only if (1.14) holds.

Proposition 1.2 states that when the highway is highly loaded, for any given p the I policy outperforms the benchmark case in terms of both W and θ, if and only if the jam speed of vehicles is higher than that of the benchmark case (i.e., $V^I_{Nc}(p) \geq V^B_{Nc}$).

We illustrate this result using the calibrated model for $\lambda = 2,217$ and $11,342$ vehicles per hour as in §1.5.1. We compute $\lambda^{(I,W)} = \lambda^{(I,\theta)} = 2,594$ vehicles per hour, so $2,217$ and $11,342$ vehicles per hour are considered as light load and high load, respectively. For light traffic, our numerical analysis in Appendix A.5.1 shows that AVs do improve the performance of the highway under the I policy, but the amount of improvement is not substantial because the benchmark model already

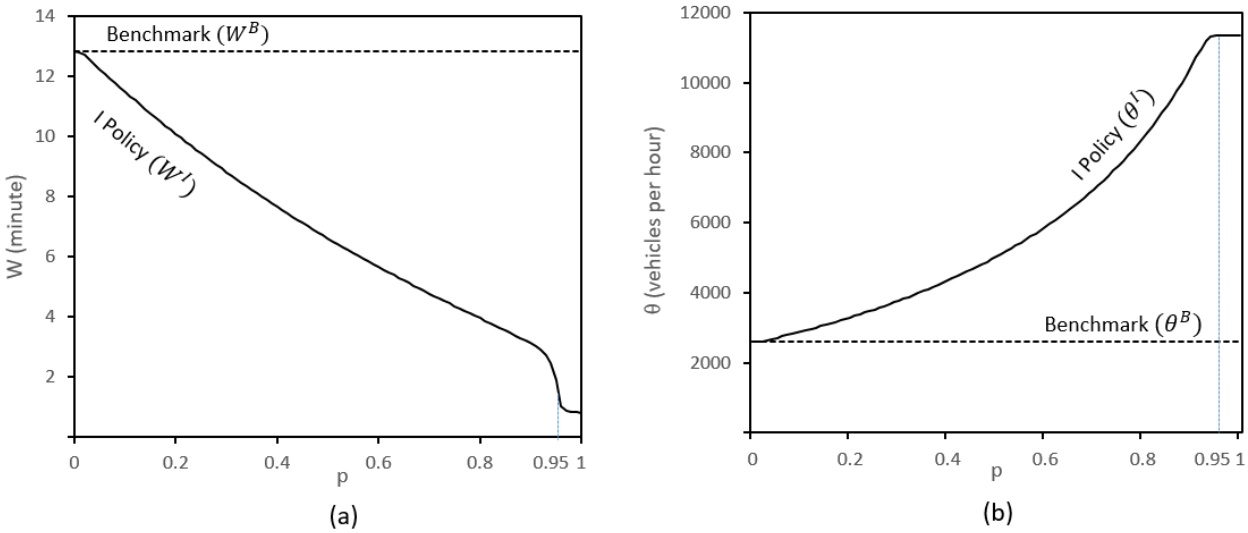


Figure 1.5: QoS measures for the I policy when $\lambda = 11,342$ vehicles per hour: (a) mean travel time, and (b) throughput.

performs quite well. At $\lambda = 11,342$ vehicles per hour, Figure 1.5(a) illustrates that $W^I(p)$ is lower than W^B for all values of p. This happens because, as shown in Figure 1.3(a), the jam speed of vehicles under the I policy, $V^I_{Nc}(p)$, is increasing in p for $p \in [0, 1]$, where $p = 0$ corresponds to the benchmark case. When $p < 0.95$, $\lambda = 11,342$ vehicles per hour is a high load (i.e., for these values of p, the jam service rate, $NcV^I_{Nc}(p)$, is lower than λ), so vehicles drive at the jam speed, $V^I_{Nc}(p)$. In this case, since $V^I_{Nc}(p)$ is increasing in p, $W = 1/V^I_{Nc}(p)$ decreases in p. When $p \geq 0.95$, AVs are so prevalent that vehicles drive at the free-flow speed of the highway, and $W^I(p)$ is minimal. In Appendix A.5.4, we offer further discussion about the sharp decrease in W at $p = 0.95$. Figure 1.5(b) shows that θ under the I policy is strictly increasing for $p \leq 0.95$. For $p \geq 0.95$, θ is equal to $\lambda = 11,342$ vehicles per hour, and it cannot grow any further, because vehicles are traveling at the free-flow speed.

Since our model captures various characteristics of a mixed traffic flow, our results offer deeper insights than previous studies. Liu et al. (2018) and Bierstedt et al. (2014) state that in order for the I policy to improve the performance of the benchmark case (in terms of θ) by about 30%, the AV proportion should be substantial – 60% and 75%, respectively. However, we observe that the performance of this policy crucially depends on λ: Whereas the I policy does not have a significant impact (about 5%) on W or θ at $\lambda = 2,217$ vehicles per hour for any p, 50% AVs halve the mean travel time and double the throughput at $\lambda = 11,342$ vehicles per hour (see Figure 1.5). This

discrepancy may stem from the fact that these two studies focus on the role of AVs in reducing the mean *intraplatoon* headway. However, as described in Appendix A.5.3, the capability of AVs to reduce the weighted mean *interplatoon* headway is the primary driver for the speed increase under the I policy.

Ghiasi et al. (2017) consider the I policy for a one-lane highway. They show that θ increases in p, if and only if $1/\xi^{IAA} \leq \frac{1/\psi_{\bar{n}}^{IAH}+1/\psi_{\bar{n}}^{IHA}}{2} \leq 1/\psi_{\bar{n}}^{IHH}$, where $\bar{n}$ is a fixed number. In contrast, for our calibrated model in §1.4.3, although there exist values of n such that the condition in Ghiasi et al. (2017) does not hold (e.g., at $\bar{n} = 100$, $\frac{3.28+2.81}{2} \not\leq 2.81$ seconds), θ is still increasing in p. This happens because the number of vehicles on a highway is not a fixed value, and speed of vehicles (as well as the mean headway) changes with n even if p is fixed. Thus, by considering only one instance of n, Ghiasi et al. (2017) do not fully capture the effect of congestion on speed.

1.5.3 Comparison of the D Policy and the I Policy

Building on the analyses in §1.5.1 and §1.5.2, we compare the performance of the D policy to that of the I policy. Under the premise of Proposition 1.1, Proposition 1.3 presents conditions under which the I policy outperforms the D policy in terms of W and θ, and vice versa.

Proposition 1.3. *Suppose $\lambda \geq \lambda_{jam}^{I}(1)$, and $V_{Nc}^{I}(p)$ is concave everywhere or convex everywhere, and it is increasing in p.*[15] *Then,*

$p \to 0$	$p \to 1$	$p \not\to 0$ or 1
$W^I(p) \to W^B \leq W^D(p)$	$W^I(p) \to W^D(p) \leq W^B$	$W^I(p) \leq W^B \leq W^D(p)$, if $p \leq p^{(D,W)}$ $W^I(p) \leq W^D(p) \leq W^B$, if $p^{(D,W)} \leq p \leq p^{(DI,W)}$ $W^D(p) \leq W^I(p) \leq W^B$, if $p \geq p^{(DI,W)}$
$\theta^I(p) \to \theta^B \geq \theta^D(p)$	$\theta^I(p) \geq \theta^B \geq \theta^D(p)$, if $N \geq \frac{V_c^{DA}}{V_{Nc}^{B}}$ $\theta^I(p) \geq \theta^D(p) \geq \theta^B$, if $N \leq \frac{V_c^{DA}}{V_{Nc}^{B}}$	For $\lambda \geq \max(\lambda_{jam}^{I}(1), \lambda^{(D,\theta)})$: $\theta^D(p) \leq \theta^B \leq \theta^I(p)$, if $p \leq \underline{p}^{(D,\theta)}$ or $p \geq \bar{p}^{(D,\theta)}$ $\theta^B \leq \theta^D(p) \leq \theta^I(p)$, if $\underline{p}^{(D,\theta)} \leq p \leq \underline{p}^{(DI,\theta)}$ or $\bar{p}^{(DI,\theta)} \leq p \leq \bar{p}^{(D,\theta)}$ $\theta^B \leq \theta^I(p) \leq \theta^D(p)$, if $\underline{p}^{(DI,\theta)} \leq p \leq \bar{p}^{(DI,\theta)}$

where $p^{(DI,W)}$ is the smallest p such that $-\frac{V_c^{DA}V_{(N-1)c}^{DH}}{V_c^{DA}-V_{(N-1)c}^{DH}} \leq V_{Nc}^{I}(p)\left(p - \frac{V_c^{DA}}{V_c^{DA}-V_{(N-1)c}^{DH}}\right)$, $\underline{p}^{(DI,\theta)}$ is the smallest p such that $NcV_{Nc}^{I}(p) - p\lambda(1-\pi_c^{DA}) = (N-1)cV_{(N-1)c}^{DH}$, and $\bar{p}^{(DI,\theta)}$ either satisfies

[15] These conditions on V can be expressed in terms of parameters (see the proof of Proposition 1.3). These assumptions are not restrictive; they are required to hold only at $n = Nc$. As Figure 1.3(b) shows, these conditions are satisfied in our calibrated model for all the heavily loaded queues we consider. Note that when $V_{Nc}^{I}(p)$ is increasing in p and $\lambda \geq \lambda_{jam}^{I}(1)$ (as assumed in the statement of Proposition 1.3), all queues are highly loaded. This happens because $V_{Nc}^{I}(1) = V_c^{DA}$ is the maximum jam speed among all queues (i.e., the benchmark case, the D policy, and the I policy) for all values of p. Therefore, $\lambda_{jam}^{I}(1)$ is at least as high as both $\lambda^{(I,W)}$ and $\lambda^{(I,\theta)}$.

$NcV^I_{Nc}(\bar{p}^{(DI,\theta)}) - (1-\bar{p}^{(DI,\theta)})\lambda(1-\pi^{DH}_{(N-1)c}) = cV^{DA}_c$, *or it is the largest* p *such that* $NcV^I_{Nc}(p) - p\lambda(1-\pi^{DA}_c) = (N-1)cV^{DH}_{(N-1)c}$.

Proposition 1.3 characterizes ranges of p where the D policy performs better than the I policy, and vice versa. When p converges to 0 or 1, almost all vehicles are HVs or AVs, respectively. In this case, the I policy outperforms the D policy as well as the benchmark case. When $p = 0$, the I policy, which is equivalent to the benchmark case, is superior to the D policy. This happens because under the D policy, the arrival rate to the AV queue approaches zero, so the highway does not utilize its full capacity. When p approaches 1, the I policy performs better than the benchmark case and the D policy, since it replaces all the HVs with fast AVs. In this case, the throughput of the benchmark case is higher than that under the D policy, only if the highway has a sufficient number of lanes (i.e., $N \geq \frac{V^{DA}_c}{V^B_{Nc}}$), such that they collectively produce a higher throughput than one fast lane of AVs.

When $0 < p < 1$, Proposition 1.3 specifies intervals for p where the D policy performs better than the I policy in terms of W and θ. We first discuss W, and then θ. As discussed in §1.5.1 and §1.5.2, unlike the D policy that decreases W over the benchmark case only when $p \geq p^{(D,W)}$, the I policy has a lower mean travel time, W^I, than that of the benchmark case, W^B, for any value of p, under the premise in Proposition 1.3. As a result, when $p \leq p^{(D,W)}$, the I policy performs better than the D policy. When $p^{(D,W)} \leq p \leq p^{(DI,W)}$, although the D policy has a lower mean travel time, W^D, than the benchmark case, the I policy still results in the lowest W, because the I policy enables HVs to travel faster by mixing them with fast moving AVs on all lanes. When $p \geq p^{(DI,W)}$, W^D is lower than W^I. In this case, the arrival rate to the HV queue is so low that this queue flows freely. Thus, W^D is determined primarily by the jam speed of vehicles on the heavily loaded AV lane. Under the I policy, the highway is also heavily loaded, but the jam speed of the mixture of vehicles is lower than the jam speed of the AV queue under the D policy.

Proposition 1.3 also compares θ between the two policies. For highly loaded highways, throughput is increasing in p under the I policy, and this policy outperforms the benchmark case for any p (see §1.5.2). When $p \leq \underline{p}^{(D,\theta)}$, the I policy should also be chosen over the D policy, because the D policy leads to a lower throughput, θ^D, than that of the benchmark case, θ^B. In addition, when $\underline{p}^{(D,\theta)} \leq p \leq \underline{p}^{(DI,\theta)}$ or $\bar{p}^{(DI,\theta)} \leq p \leq \bar{p}^{(D,\theta)}$, even though θ^D is higher than θ^B, throughput under

the I policy, θ^I, is still the highest. Only when $\underline{p}^{(DI,\theta)} \leq p \leq \bar{p}^{(DI,\theta)}$ does the D policy outperform the I policy in terms of throughput. In this case, the arrival rate to the AV queue of the D policy is so high that this queue works at its jam throughput. The high throughput of this fast AV lane makes up for the low throughput of the HV lanes, and increases the overall throughput under the D policy. In contrast, under the I policy, the state-dependent speed is balanced between fast AVs and slow HVs, so θ^I is not as high as θ^D. When $p \geq \bar{p}^{(D,\theta)}$, θ^I is higher than θ^D: AVs, that constitute the majority of vehicles, can run on all lanes under the I policy, whereas they can run only on one designated lane under the D policy.

We illustrate Proposition 1.3 using the calibrated model in §1.4 for highly loaded highways with $\lambda = 11,342$ vehicles per hour. As Figure 1.6(a) shows, W^I is lower than W^D for all values of p, except when p is between $p^{(DI,W)} = 0.93$ and 0.94.[16] Figure 1.6(b) compares throughput under the I policy with that under the D policy. As stated in Proposition 1.3, when $p \in [0.25, 0.55]$ (where $\underline{p}^{(DI,\theta)} = 0.25$ and $\bar{p}^{(DI,\theta)} = 0.55$), θ^D is higher than θ^I. When the highway is lightly loaded with $\lambda = 2,217$ vehicles per hour, in terms of both W and θ, the I policy performs at least as well as the benchmark case, while the D policy can be inferior to the benchmark case. In this case, integrating AVs and HVs improves the performance of the highway more than assigning one lane to AVs for any value of $p \in [0, 1]$ (see Appendix A.5.1).

Proposition 1.3 shows, interestingly, that on some interval for p, either policy outperforms the other in terms of W or θ. Corollary 1.2 specifies this interval.

Corollary 1.2. $\underline{p}^{(DI,\theta)} \leq \min\{p^{(DI,W)}, \bar{p}^{(DI,\theta)}\}$.

Corollary 1.2 implies that there exists an interval for p, $[\underline{p}^{(DI,\theta)}, \min\{p^{(DI,W)}, \bar{p}^{(DI,\theta)}\}]$ (e.g., $[0.25, 0.55]$ in Figure 1.6(b)), such that, in terms of throughput, the D policy performs better than the I policy, but in terms of mean travel time, it is worse. The driver of this trade-off is the fact that, under the D policy, the highway is divided into a fast queueing system for AVs and a slow one for HVs. On the other hand, the I policy balances the quality of service received by HVs and AVs, by mixing fast AVs with slow HVs. When $p \in [\underline{p}^{(DI,\theta)}, \min\{p^{(DI,W)}, \bar{p}^{(DI,\theta)}\}]$, the throughput derived from the balanced speed of vehicles ($V^I_{Nc}(p)$) under the I policy is lower than the high

[16] Note that when $p > 0.94$, since $V^I_{Nc}(p)$ is increasing in p, $\lambda = 11,342$ vehicles per hour becomes lower than the jam load for this highway, i.e., the condition $\lambda \geq \lambda^I_{jam}(1)$ in Proposition 1.3 is violated. In addition, the mean travel time of the I policy in this case is lower than that when the highway is jammed. Therefore, $W^I(p)$ is not necessarily higher than $W^D(p)$ anymore.

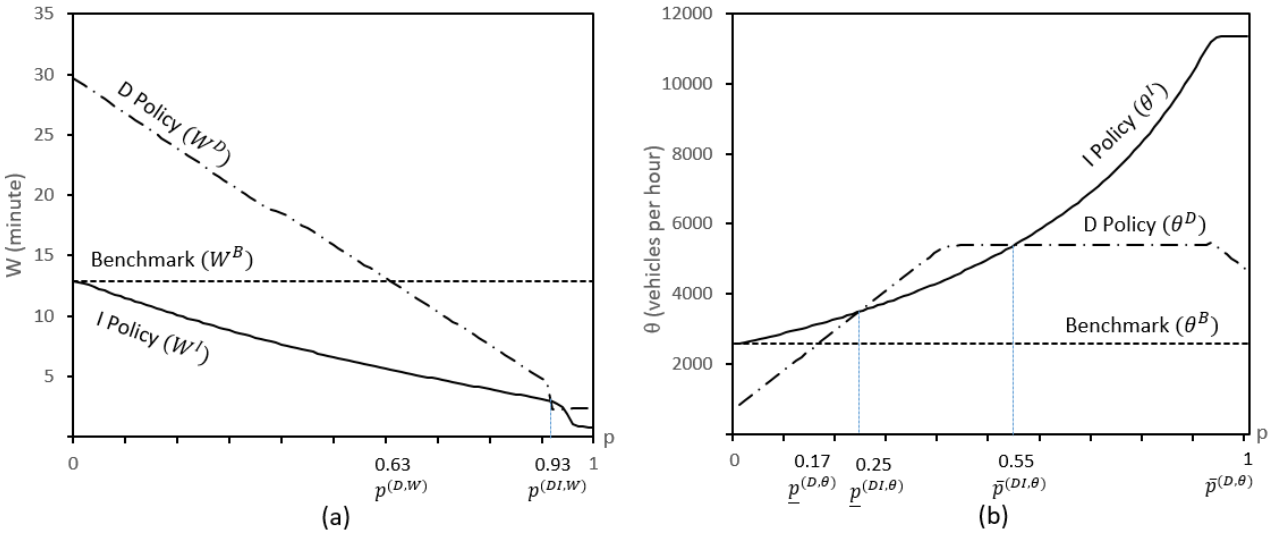


Figure 1.6: A comparison between the D policy and the I policy when $\lambda = 11,342$ vehicles per hour: (a) mean travel time, and (b) throughput.

throughput the D policy achieves by utilizing the fast AV lane.

Under the D policy, it seems intuitive to designate more than one lane to AVs when the AV proportion p is high. Numerically analyzing the D policy with two lanes designated to AVs, we observe that although this policy increases throughput, it still performs poorly in terms of mean travel time. See Appendix A.5.5 for further discussion.

1.5.4 Robustness of the Results

We demonstrate the robustness of our results as follows. In Appendix A.6.1, we show analytically that our main results hold for any platoon size distribution, as well as for a broad class of distributions for intraplatoon and interplatoon headways. In Appendix A.6.2, we report the results of sensitivity analyses on several model parameters. We observe that, although the intervals $[p^{(DI,W)}, 1]$ and $[\underline{p}^{(DI,\theta)}, \bar{p}^{(DI,\theta)}]$ change as parameters vary, the main qualitative insights continue to hold: If the performance metric considered by policy makers is W, we recommend the I policy; otherwise, we recommend the D policy in a moderate region of p, and the I policy in other regions.

1.6 Simulation Study

We create a discrete-event agent-based simulation for a multi-lane highway. The goal of this simulation is to verify the validity of our queueing model and analysis. We present a summary of our simulation model and results in this section, while providing more details in Appendix A.7.

We develop a discrete-event cellular automata simulation (DECAS). This approach combines two methods, Discrete-event simulation (DES) and cellular automata (CA), used in prior literature. DES is commonly used to simulate queueing models (e.g., Law et al. (2000) and Ross (2006)). CA models are capable of explicitly representing individual vehicle interactions and relating these interactions to macroscopic traffic flow metrics, such as mean travel time and throughput (e.g., Benjaafar et al. 1997). Thus, DECAS is appropriate for simulating traffic flow in our setting.

In our DECAS model, a highway is modeled as a grid. Each cell of the grid can be occupied by at most one vehicle. In a typical CA model, the state of the system (i.e., speed and location of vehicles that are present on the grid) evolves according to a predefined set of rules at every time step (usually one second). Instead of such a discrete-time simulation, we employ a discrete-event simulation by updating the state of the system when one of the following "events" happens: (i) arrival of a new vehicle to the highway segment ("arrival" for short) and (ii) departure of an existing vehicle from the highway segment ("departure" for short). This approach significantly improves the speed of large-scale simulation in our setting.

As compared to the calibrated analytical model presented in §4, our simulation model incorporates the following general features:

• Length of the highway segment $L > 1$ (whereas we use the normalized length $L = 1$ in the numerical analysis presented in §1.5).

• Lane changing: If the speed of a vehicle is higher than the vehicle immediately in front, the vehicle is allowed to leave its current platoon (of which the size can be one or larger) and to create a new platoon in an adjacent lane or merge into the existing platoon in an adjacent lane, if the following conditions hold. First, there is enough space (e.g., at least one empty cell) between this vehicle and the vehicle immediately in front in the adjacent lane. Second, the gap between this vehicle and the vehicle immediately behind in the adjacent lane is so high that, if the vehicle behind travels at the maximum speed of the highway, the advancement of this vehicle is smaller than the gap.

• Platoon formation process: As a vehicle arrives to the highway, it decides whether or not to join the existing platoon immediately in front. A vehicle can also leave its current platoon and create a new platoon (or merge into another platoon), as it changes its lane.

• A transient behavior of vehicles: The simulation model allows speed-up, which is defined as the ability of a vehicle to increase its speed if there is a large gap between this vehicle and the vehicle immediately in front.

Our simulation implicitly includes two more features of traffic flows: the negative effect of lane-changing on speed, and mergers of platoons. First, right after a vehicle changes its lane, if its speed is lower than that of the vehicle immediately behind, the vehicle immediately behind either reduces its speed to match the speed of the vehicle in front or changes its lane. Second, when the speed of a vehicle is higher than that of the vehicle immediately in front, and the vehicle is not able to change its lane, this vehicle is forced to reduce its speed to match the speed of vehicle immediately in front; then the follower vehicle merges into the platoon immediately in front.

As in our calibrated model in §4, we run the simulation for a highway segment with three lanes, and a jam density of 185 vehicles per mile per lane. We consider a segment of four miles: we use the first mile as warm-up state to place vehicles on the segment and form platoons, and use the results from the remaining three miles. The arrival rate to the segment is $11,342$ vehicles per hour.

To investigate the performance of the D and I policies, for each value of $p \in \{10, 20, \cdots, 100\}$ we run the simulation 30 times for a time horizon of four hours, and report the average of these 30 values in Figure 1.7. For a highly loaded highway, the running time of each instance of our simulation is, on average, about half an hour; for a lightly loaded highway, it is about one minute. The confidence intervals of these values are very small: The largest length of a 95% confidence interval, which belongs to the benchmark case, is equal to 0.2 minutes. As Figure 1.7(a) shows, similar to our numerical results in §5, for any given value of p, the I policy outperforms the benchmark model in terms of W. When p is at least 64%, the D policy is also capable of reducing W over that of the benchmark model. Similarly, Figure 1.7(b) depicts that, whereas the I policy has a higher θ than the benchmark model for all values of p, the D policy increases θ over the benchmark model when $p \geq 0.16$, and it outperforms the I policy when $p \in [0.24, 0.46]$. Although W^D and W^I are very similar to what we observe in §5, θ is a bit lower in the simulation than that in the queueing model, especially when p is small. This is most likely due to the use of a more stringent blocking

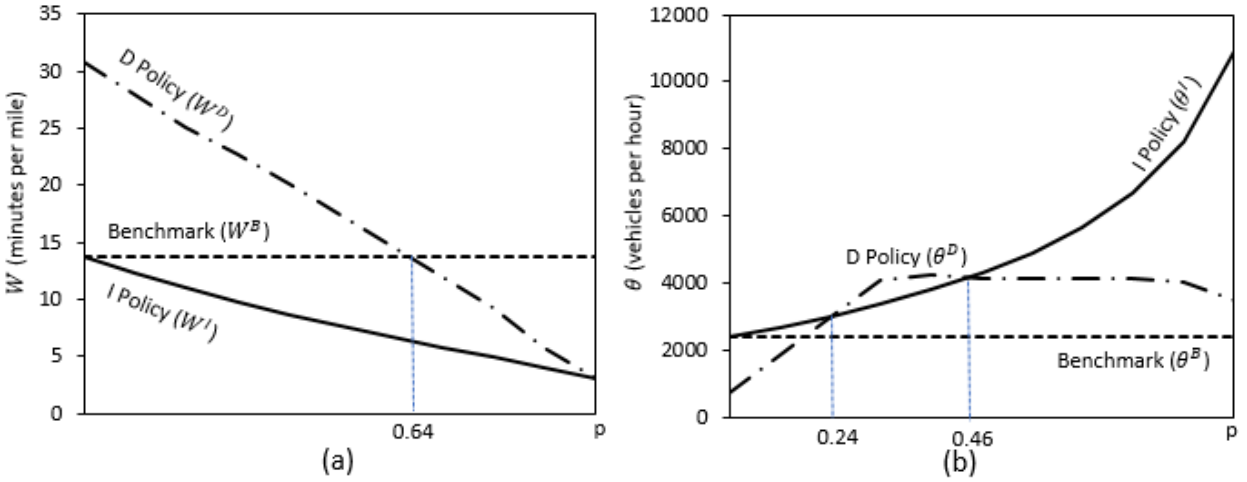


Figure 1.7: A comparison between the D policy and the I policy when $\lambda = 11,342$ vehicles per hour: (a) mean travel time of the simulation, (b) throughput of the simulation.

rule in the simulation: Whereas in the queueing model a vehicle is blocked when the entire highway segment is full, in the simulation a vehicle is blocked when the first mile of the highway is full. See Appendix A.7 for a more in-depth discussion of the results.

1.7 Policy Recommendation and Conclusion

In this paper we investigate the effects of autonomous vehicles on highway traffic flow under two policies: the D and the I policies. We model traffic flow as a queueing system, and calibrate this model to data. Then, we analyze each of these policies, as well as the benchmark case, and provide recommendations about when each of these policies should be considered. We further validate our findings with a simulation study that takes into account several general features.

We use two metrics to measure the performance of a highway: the mean travel time of a single vehicle, and the throughput of the highway. The former concerns users of the highway, while the latter is important for urban designers. Interestingly, these two metrics are not always aligned: A high utilization of the highway does not guarantee a short travel time, and vice versa. Thus, it is crucial for policy makers to take both of these metrics into account.

The performance of different policies depends crucially on arrival rate to the highway, which has been overlooked in the AV literature to date, as well as the proportion of AVs on the road. Since AVs are primarily intended to alleviate congested highways, we focus on high arrival rates. For

each metric, we recommend policies in three different regions of the AV proportion: low, moderate, and high. Our analysis indicates:

1) When the AV proportion is low, the I policy is recommended. This is intuitive because the number of AVs on the highway is so low that it is not worth designating a lane to them. Thus, the Departments of Transportations in several states, including Colorado, Wisconsin and Washington, should recognize that their plan to designate a lane to AVs (Aguilar 2018) could lead to significantly more congestion than the I policy (although it may have other benefits, such as helping human drivers become acquainted with the new era of AVs, and incentivizing adoption).

2) When the AV proportion is moderate, a policy maker could consider adopting either policy, depending on which metric he or she cares about more. If a lower travel time for vehicles is of more importance, the I policy is the solution; but if improving the overall utilization (throughput) of the highway is of high priority, the D policy should be considered. Under the D policy, the fast AV lane increases throughput rapidly beyond that of the I policy, but the slow HV lanes lead to a longer mean travel time than that under the I policy.

3) When the AV proportion is high, the I policy is again recommended based on both metrics. This is because AVs, which constitute the majority of vehicles, are allowed to use only one lane of the highway under the D policy, but they can drive on any lane under the I policy. Whereas Bierstedt et al. (2014) predict that, under the I policy, even high values of the AV proportion will not have a significant impact on throughput, (e.g., 60% AVs increase throughput by only 30%), we observe that 60% AVs increase throughput by about 130%. Moreover, in line with KPMG's prediction that AVs could increase the capacity of highways by 500%, our calibrated model shows that, when all vehicles are autonomous, the I policy increases the throughput of a congested highway by 437%.

Our paper is the first to model and analyze a multi-lane highway using a queueing system coupled with a MAP. There are several interesting avenues to expand this research. First, since AVs are not commercially available yet, we conduct our numerical analysis based on experimental data used in prior literature. Although we perform sensitivity analyses, more accurate data that will become available may yield more precise results. Second, incorporation of detailed highway characteristics such as entry or exit ramps, a large-scale network structure of highways and roads, and microscopic-level behaviors of vehicles will further enrich our model. In particular, the validity of modeling a highway as a loss queue is worth further investigation in such a general setting.

Third, one may explore endogenizing the adoption rate of AVs based on quality of traffic flow on highways and regular roads: If having AVs helps reduce congestion, more people may be interested in trading their conventional vehicle for an autonomous one. As a result, the adoption rate of AVs may be determined depending on various factors that include the amount of improvement in congestion on regular roads and highways, the price of AVs (which is unknown yet), possibly additional trips due to the autonomous feature (which gives free time to drivers), and so on. Finally, it would be interesting to study the effect of AVs on highway crash rates. On the one hand, AVs are more capable of preventing crashes than HVs, due to their low reaction time. On the other hand, platoons of AVs are longer and denser than platoons of HVs, so an accident between two AVs could propagate through the entire platoon and affect a longer string of vehicles. Thus, it is not immediately clear if AVs will improve highway safety.[17]

[17]Our queueing model captures some safety features implicitly; for example, we consider a larger headway for an HV-AV pair (i.e., when an AV follows an HV) than for an AV-AV pair (i.e., when an AV follows an AV). We also conduct a sensitivity analysis on the value of the mean intraplatoon headway of an AV-HV pair in Appendix A.6.2. We leave explicit modeling of other safety features for future research.

Chapter 2

Can Autonomous Vehicles Solve the Commuter Parking Problem?

2.1 Introduction

Autonomous Vehicles (AVs) are expected to play a determinant role in shaping the future of mobility. AVs are already operating on the roadways of several major cities in the United States (U.S.) and other countries (Bayern 2020). As such, now is a rare window of opportunity for social planners to put in place policies that pave the way for the inevitable mass arrival of AVs. Among several urban transportation issues that need to be addressed, this paper concerns *parking* and *traffic congestion* during morning rush hour. Currently, to accommodate high parking demands, cities typically surrender large spaces to build parking structures; for example, an astounding 14 percent of Los Angeles county land is dedicated to parking (Jaffe 2015).[1] Nevertheless, morning commuters usually struggle to find an affordable parking spot in a convenient location. Another significant problem of morning commute is traffic congestion. According to Ingraham (2019), a report from the U.S. Census Bureau (2019) indicates that, while the average daily commute time to work in the U.S. is 26.1 minutes, a notable 6% of morning commuters (about 9 million Americans) travel at least an hour to work, committing over 10 full days per year to the time just getting to work. AVs have potential to address these issues, as they can drop commuters off at their workplaces

[1] We gratefully acknowledge that this statistic was borrowed from the keynote talk ("OR and the Transportation Tech Revolution") of Garrett van Ryzin during the 2018 INFORMS Annual Meeting.

downtown and park in a suburban area with cheaper parking.[2] Such self-parking ability of AVs not only allows commuters to avoid high parking fees downtown, but also reduces a city's need to build or maintain large parking structures in its downtown area. AVs may also help reduce congestion. This is because these vehicles can be made aware of the location of available parking spots using vehicle connectivity technologies, so they do not need to search for parking downtown. However, the additional traffic from downtown to suburban parking areas can add extra congestion, so the effect of AVs on the morning commute is unclear.

In this paper, we study the effect of AVs on the morning commute and provide guidance to social planners (e.g., mobility and infrastructure departments of mayoralties, city councils, town councils and town boards) on how to adapt parking fees, congestion tolls, and infrastructure to the special needs and characteristics of AVs. Specifically, we address the following research questions: (1) In comparison to human-driven vehicles (HVs), can AVs reduce or even eliminate commuters' need for parking downtown? (2) How should a social planner determine the optimal parking fees and congestion tolls, and design infrastructure (e.g., drop-off capacity and downtown parking capacity) to minimize the total system cost of the morning commute (which includes the travel cost of all commuters and the commuters' penalties for not arriving on time at work)?

To answer these questions, we examine the problem faced by morning commuters who use AVs to travel from home to work in a downtown area. The goal of each commuter is to minimize her total transportation and parking costs during the morning commute. (For convenience, we use the pronouns "she/her/hers" to refer to a commuter.) Three major elements affect an individual commuter's costs: travel time, arrival time at work, and parking fee. The individual commuters' decisions on their departure times from home collectively affect the level of congestion on roadways; hence, these decisions affect both travel time and arrival time of an individual commuter. To avoid a late arrival, a commuter may decide to depart home early, but the commuter may not want to depart too early because there exists an inconvenience cost for arriving too early at work (Hendricks 2015). As such, commuters encounter a trade-off between experiencing congestion and arriving too early to work. Each commuter also decides on her parking location between a central parking area (located downtown) and an external parking area (located outside downtown). The external

[2]Large-scale commercial fully autonomous cars (which have the ability to drop off commuters) are expected to be available by 2025 (Faggella 2020). As such, many cities, including Los Angeles and Cincinnati, have planned for converting downtown parking spaces to curbside drop-off zones (Shaver 2019).

parking area is cheaper than the central parking area, but farther away from work, so AVs need to exit the congested downtown to travel to the external parking area. This presents commuters with a second trade-off.

These trade-offs faced by commuters are captured in a continuous-time game-theoretic traffic model. Our model is motivated by infrastructure and traffic patterns in the City of Pittsburgh. In this model, commuters decide on their departure times from home and their parking locations between the central and external areas. These commuters take an inbound highway (e.g., I-376 in Pittsburgh) to travel to work downtown. Upon arrival at work, the commuters are dropped off as soon as there is an available curbside drop-off spot, and their AVs drive off to the parking areas chosen by the commuters. Our model takes into account the congestion that the commuters experience on the inbound highway to downtown, as well as the extra congestion that their AVs create as they leave downtown to go to the external parking area. We calibrate our model to traffic and parking data from Pittsburgh.

By analyzing this model, we characterize the departure time and parking location patterns of commuters under user equilibrium (UE). Under UE, no commuter can unilaterally change her departure time and/or parking location to reduce her cost. We show that there exist two different user equilibria, depending on the distance of the external parking area from downtown and the parking fee of the central area (relative to that of the external area). For cities, such as Pittsburgh, where the parking fee of the central area outweighs the cost of traveling to the external parking area, commuters choose the external area until the congestion downtown, caused by AVs traveling to the external parking area, becomes so high that the remaining commuters prefer to park in the expensive central parking area. In this case, commuters tend to leave early to avoid the snowballing downtown congestion. The results of this case are starkly different from the status quo where all commuters drive HVs and the option of parking in the external area does not exist. This results suggest that, after the mass adoption of AVs, large areas may not be devoted to parking spaces in Downtown Pittsburgh (unless extra parking fees and congestion tolls are imposed as we discuss below). However, for cities with relatively low downtown parking fees, our analysis shows that commuters still prefer to park in the central parking area.

We next analyze the problem of a social planner who aims to minimize the total system cost of the morning commute by dictating the departure time and parking location of all commuters. The

analysis of a social optimum (SO) allows us to identify a gap between the total system cost under UE and that of SO. Our analysis shows that there exist two different SOs. For cities such as Pittsburgh, where the travel time from work to the central parking area is lower than that to the external parking area, the social planner routes AVs to the central area to minimize the total system cost. In other cities, the social planner routes AVs to the external area until the downtown congestion becomes so high that the travel time to this area (which includes the downtown congestion time and free-flow travel time from downtown to the external area) grows beyond the travel time to the central area, so the travel cost to the central area becomes lower than that to the external area; at this point, the social planner routes AVs to the central area until the downtown congestion becomes so low that the external area becomes the better option again. This cycle of alternately routing AVs to the external and central areas continues until all AVs have assigned parking spots or the central area is full (in which case the remaining AVs must go to the external area). Under either SO, the decisions made by the social planner are different from those made by commuters under UE, and thus the total system cost under UE is higher than the SO cost.

To close the performance gap between the SO and the UE, we examine both short-term and long-term measures a social planner can implement. As for short-term measures, a social planner may adjust parking fees and impose congestion tolls. These measures can reduce the total system cost by inducing commuters to choose the departure times and parking locations desired by the social planner. Our numerical analysis of the Pittsburgh data indicates that these measures can reduce the total system cost of the morning commute by 51%. In the long run, a social planner can lower the total system cost further by adapting the infrastructure to the special needs and characteristics of AVs. For example, since AVs have the ability to drop off their commuter at work and park outside downtown, it may be beneficial to increase the number of curbside drop-off spaces while reducing the number of parking spots downtown. For Pittsburgh, we find that converting all downtown parking spots to curbside drop-off spots in the long run can lead to an additional 70% reduction in the total system cost of the morning commute. Even by converting only 10% of the parking spots to curbside drop-off spots (which increases the drop-off capacity to the inbound highway capacity) in Downtown Pittsburgh, we can achieve an almost 21% reduction. These results suggest that major cities can benefit significantly from the mass adoption of AVs by adjusting their short-term and long-term transportation and infrastructure policies.

The rest of this paper is organized as follows. In §2.2, we review the related literature. Our morning commute travel model is presented in §2.3. In §2.4 and §2.5, we analyze the UE and the SO, respectively. In §2.6, we propose both short-term and long-term plans to reduce the total system cost of the morning commute. We conclude in §2.7.

2.2 Related Literature

This work is related to two streams of research: smart city operations and transportation science. Under the umbrella of the smart city operations literature, there are various studies on ride-sharing (e.g., Benjaafar et al. (2018), He et al. (2018), Qi et al. (2018), and Bai et al. (2019)) and electric vehicles (e.g., Mak et al. (2013), Lim et al. (2014), and Pelletier et al. (2016)). The role of AVs in the future of transportation is a nascent part of the emerging literature on smart city operations. Qi et al. (2020) study the potential of shared autonomous electric vehicles (SAEVs) for improving the self-sufficiency and resilience of solar-powered urban microgrids. They show that even a moderate-sized SAEV fleet can improve microgrid self-sufficiency, and microgrid resilience can be significantly enhanced by SAEVs. As discussed by Mak (2020) and Hasija et al. (2020), AV technology is a promising research direction in the smart city operations domain. We expand this evolving stream of research by examining the effect of AVs on congestion and parking during the morning commute.

The early research in transportation science literature that studies the morning commute problem focuses on HVs. Morning commuters choose their departure times from home based on multiple factors such as congestion, schedule delays, parking fees and availabilities. Vickrey (1969) considers a finite group of commuters who decide on their departure time from home to their work places downtown. He shows that there exists an equilibrium departure time pattern when all commuters attempt to minimize their own travel costs. Arnott et al. (1991) extend Vickrey (1969) by examining commuters' decisions on both departure times from home and parking locations. They consider a combination of congestion tolls and parking fees to minimize the total system cost, and show that the optimal departure rate from home must be equal to the capacity of the inbound highway. Xu et al. (2019) study different flat congestion toll schemes that are easier to implement to reduce the total system cost. The model of Arnott et al. (1991) is further extended by accounting for other features such as joint morning and evening commute (e.g., Zhang et al. (2008)), multiple parking

clusters downtown (e.g., Qian et al. (2011)), multiple residential areas (e.g., He et al. (2015)), and positive search time to find an empty parking spot (e.g., Qian & Rajagopal (2014) and Qian & Rajagopal (2015)). Our model also builds on the fundamental structure of Arnott et al. (1991) while enriching it by capturing the specific characteristics of AVs, such as their ability to drop off their commuters and park outside downtown without carrying passengers.

Recently, there are a few studies that incorporate AVs in the morning commute problem. Nourinejad & Amirgholy (2018) consider the morning commute problem in which AVs can drive to remote, but cheap, parking spots downtown. They show that AV owners depart home later and park farther away from work than HV owners. Zhang et al. (2019) extend the model of Nourinejad & Amirgholy (2018) to the joint morning and evening commute. Similar to our paper, Liu (2018) considers the morning commute problem for commuters who use AVs. In his model, commuter drop-offs happen without any delay, and all parking spots are distributed evenly along a line from work outwards in the downtown area. Liu (2018) shows that under equilibrium commuters choose the closest available parking spot to work. In addition, he shows that the number of commuters who leave home per hour must be equal to the capacity of the inbound highway in a social optimum, confirming the result of Arnott et al. (1991).

We contribute to this literature by examining the ability of AVs that can drop off commuters at work and then self-park. To properly model this ability of AVs, different from the prior literature reviewed above, we consider an external parking area (located outside downtown) for AVs, the downtown congestion caused by AVs, and the capacity constraint at the drop-off location. Modeling the external parking option is essential because this is one of the most important benefits of AVs that helps reduce downtown parking demands. We show that the external parking area can be the primary parking option chosen by commuters under equilibrium. This implies that ignoring this additional parking option may result in an overestimation of commuter travel cost. However, the extra trip of AVs to the external parking area can create more congestion downtown, so the aggregate effect of the external parking option is not clear without a careful analysis. In addition, our model captures the fact that the curbside space dedicated to commuter drop-offs is limited, potentially causing delays during commuter drop-offs. Such delays are illustrated in a simulation study by Overtoom et al. (2020). Different from the prior literature, in our social optimum solution, the number of commuters who leave home per hour does not exceed the number of curbside drop-off

spots (which is lower than the capacity of the inbound highway) to eliminate congestion at drop-off. Finally, we offer insight into both short-term and long-term measures a social planner can implement in anticipation of mass adoption of AVs. In short, our paper offers a unique perspective into the future of smart cities by investigating the impact of AVs on parking and congestion downtown.

2.3 Model

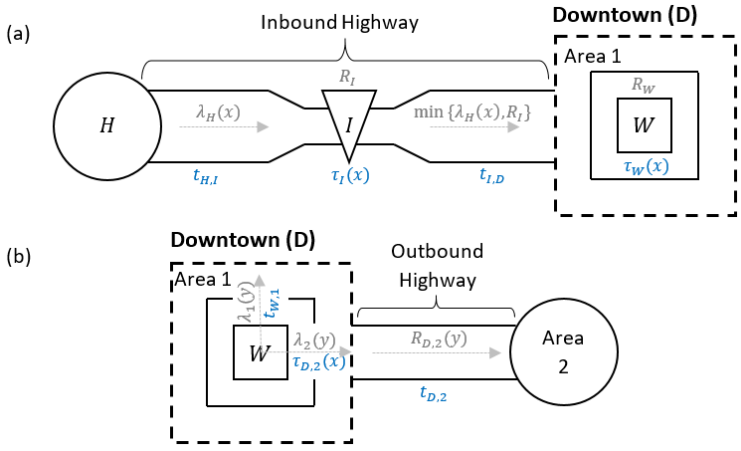
We study the problem of N morning commuters who use AVs. All commuters go from home (H) to work (W) that is located downtown (D). Appendix A provides a table for the summary of our notation.

Figure 2.1(a) depicts the travel route of AVs from H to W while they carry passengers. As shown in this figure, the route from H to D is assumed to be a highway which has an inbound bottleneck (I) with the capacity of R_I vehicles per hour (where $0 < R_I < \infty$). This means that if more than R_I commuters arrive at this bottleneck per hour, these commuters experience a delay on their way to W. As AVs arrive at D, they drop their commuters off at W. Due to limited curbside space at W, at most R_W (> 0) drop-offs per hour are possible. If more than R_W AVs per hour arrive at W to drop off their commuters, these commuters experience a drop-off delay.

After AVs drop off their commuters, they move from W to parking areas, as illustrated in Figure 2.1(b). There are two parking areas available: central (Area 1) and external (Area 2). Area 1 consists of all downtown parking spots surrounding W, and Area 2 is a parking area located outside D. The price of parking at Area j ($\in \{1, 2\}$) is p_j, and the capacity of Area j, K_j, is defined as the number of parking spaces in this area. On the way out from W to Area 2, AVs move in different directions, affecting the travel rate $R_{D,2}$ of AVs in D and potentially causing delays. We assume that AVs do not experience any congestion on the outbound highway from D to Area 2. This assumption is justified because the capacity of the outbound highway is comparable to that of the inbound highway; but, due to congestion in I and D, the rate of AVs coming out of D to go to Area 2 is lower than the rate of AVs entering D (which is at most equal to the capacity of I, R_I).

Every morning a commuter makes two decisions: departure time x ($\in [0, x_{max}]$) from H (where the latest departure time x_{max} is determined endogenously), and parking location j ($\in \{1, 2\}$). The

Figure 2.1: An illustration of the morning commute: (a) the passengered part of the trip, and (b) the passengerless part of the trip.



departure rate of commuters from H at x is denoted by $\lambda_H(x)$. Following the prior literature (e.g., Arnott et al. (1991), Qian et al. (2011), Qian & Rajagopal (2015), and Liu (2018)), the official start time at work is denoted by T for all commuters. According to the U.S. Bureau of Labor Statistics (2015), most workers in the U.S. are on the job between 8 a.m. and 5 p.m, so it is justified to assume that the majority of companies start their workday at a specific time (e.g., 8 am).

As illustrated in Figure 2.1, travel time of AVs that leave H at time x is divided into: (1) free-flow travel time, (2) delay in the inbound bottleneck, (3) delay caused by drop-off congestion at W, and (4) delay caused by congestion within D. We elaborate on each of these travel times.

(1) Free-flow travel time: This is the travel time of vehicles when they are able to move freely on segments of highways and roads without a capacity constraint. Specifically, the free-flow travel time occurs on the following four segments: (i) between home H and inbound bottleneck I, (ii) between inbound bottleneck I and the boundary of downtown D, (iii) between work W and Area 1, and (iv) between the boundary of D and Area 2. Free-flow travel time in each of these segments, which has a fixed duration, is denoted respectively by $t_{H,I}, t_{I,D}$, $t_{W,1}$, and $t_{D,2}$. Since all commuters experience $t_{H,I}$ and $t_{I,D}$, without loss of generality, we normalize them to zero.

(2) Inbound queueing delay: If the rate of AVs that arrive to the inbound bottleneck exceeds the capacity of this bottleneck, R_I, these AVs experience a delay, which is called the inbound queueing delay, $\tau_I(x)$.

(3) Drop-off congestion time at W: Before dropping off their commuters at W in D, AVs may

experience delays due to the limited curbside space at W. We call this delay the drop-off congestion time and denote it by $\tau_W(x)$.

(4) Congestion time in D: Immediately after dropping off commuters, AVs that head to Area 2 may experience delays in D. The downtown area D is a network of roads, where AVs move in different directions as they exit D on their way to Area 2. These multi-directional traffic flows may create congestion, and add to the travel time of AVs that go to Area 2. This is in accord with the macroscopic fundamental diagrams of traffic flows in a region, which indicate that, as traffic density in a region increases, the traffic flow in this region decreases and congestion forms (see Geroliminis et al. (2007) for further details). The added travel time due to this congestion is called the congestion time in D, denoted by $\tau_{D,2}(x)$.

Taken together, the total travel time of AVs that go to Area 1 is $\tau_I(x) + \tau_W(x) + t_{W,1}$, and that of AVs that go to Area 2 is $\tau_I(x) + \tau_W(x) + \tau_{D,2}(x) + t_{D,2}$. Let y denote the drop-off time at W for a commuter who leaves H at time x; then,

$$y = x + \tau_I(x) + \tau_W(x), \tag{2.1}$$

where $0 \leq y \leq y_{max} = x_{max} + \tau_I(x_{max}) + \tau_W(x_{max})$. For $j \in \{1, 2\}$, we let $\lambda_j(y)$ represent the rate of commuters in D who are dropped off at W at time y and choose Area j. Then the departure rate of AVs from W to the parking areas is at most equal to either the departure rate $\lambda_H(x)$ from H when there is no congestion ($\tau_I(x) = \tau_W(x) = 0$ and $y = x$) or the drop-off rate R_W; i.e., $\lambda_1(y) + \lambda_2(y) = \min\{\lambda_H(x), R_W\}$. AVs that head to the parking areas from W form two separate streams: one for the AVs headed to area 1 and the other one for the AVs headed to Area 2. We denote the travel rate of AVs in D that head to Area j ($\in \{1, 2\}$) as $R_{D,j}(y)$, and its sum as $R_D(y) = R_{D,1}(y) + R_{D,2}(y)$ for $y \in [0, y_{max}]$. Due to congestion in D, $R_D(y)$ may be lower than $\lambda_1(y) + \lambda_2(y)$.

We make the following assumptions throughout the paper:

A1. Since Area 1 is located in D, it has a higher parking fee and a lower capacity than Area 2, i.e., $p_1 > p_2$ and $K_2 > K_1$. Since there is an abundance of parking spaces in Area 2 (i.e., $K_2 \gg N$), the parking fee p_2 of this area is normalized to zero. In addition, we assume that the parking fee at Area 1, $p_1(x)$, is an increasing function of x ($\in [0, x_{max}]$), because

the commuters choose the cheapest parking spots available, so the parking fee goes up as a commuter's departure time from H increases, as cheaper spaces are taken.

A2. Once a commuter decides on her parking location, a parking spot is assigned to her AV. This means that AVs that go to Area 1 are directly routed to their assigned parking spots without any delay. In other words, the travel rate in D for AVs that are headed to Area 1, $R_{D,1}(y)$, is equal to the departure rate of AVs from W to Area 1, i.e., $\lambda_1(y)$.

A3. The travel rate in D for AVs that drop off their commuters at W at time y, $R_D(y)$, is a linear decreasing function of the number of AVs present in this area at time x, i.e., $R_D(y) = M - \theta \int_0^y [\min\{\lambda_H(z), R_I, R_W\} - R_D(z)]dz$, where M and θ are positive constants. This assumption is in line with the empirical evidence, presented in Daganzo (2007) and Geroliminis et al. (2007), which shows that as the traffic density in an area increases, the flow decreases.

A4. The capacities of different segments on the way from H to parking areas decrease, i.e., $R_I > R_W > R_D(y)$ for $y \in [0, y_{max}]$. If any of these inequalities does not hold, the congestion time associated with that inequality disappears: when the inbound bottleneck capacity R_I is lower than the drop-off rate R_W at W, there is no drop-off congestion. Similarly, if the drop-off rate R_W is lower than the exit rate $R_D(y)$, there is no congestion in D. Our analysis can be easily extended to those simpler cases with no congestion.

A5. We model each segment on the way from H to the parking areas (inbound bottleneck, drop-off area, and downtown) as a queue with deterministic time-varying arrival and service rates. For example, the inbound bottleneck is a queue with arrival rate $\lambda_H(x)$ for $x \in [0, x_{max}]$ and service rate R_I vehicles per hour. As discussed in Kim & Whitt (2013), calculating the exact wait time (i.e., congestion time in a segment) for each individual AV that arrives to that segment is complex. Thus, we estimate the individual wait time of an AV that leaves H at time x as the number of AVs that are present in the queue divided by the average throughput of the queue. We divide by the average throughput because AVs that are present in the queue, but have left H before time x, might leave the queue at different rates depending on their arrival time to the queue. Specifically, the congestion time $\tau_I(x)$ that a commuter traversing the I bottleneck at time x experiences is equal to $\frac{\int_0^x [\lambda_H(u) - R_I]^+ du}{(\int_0^x \min\{\lambda_H(u), R_I\} du)/x}$. The numerator of

this expression is the total number of AVs present in the bottleneck at time x, and the denominator is the average departure rate from I between time zero and time x. Similarly, we calculate the drop-off congestion time as $\tau_W(x) = \frac{\int_0^{x+\tau_I(x)} [\min\{R_I, \lambda_H(u)\} - R_W]^+ du}{\int_0^{x+\tau_I(x)} \min\{\lambda_H(u), R_W\} du / [x+\tau_I(x)]}$, and congestion in D as $\tau_{D,2}(x) = \frac{\int_0^{x+\tau_I(x)+\tau_W(x)} [\lambda_2(u) - R_{D,2}(u)]^+ du}{\int_0^{x+\tau_I(x)+\tau_W(x)} \min\{\lambda_2(u), R_{D,2}(u)\} du / [x+\tau_I(x)+\tau_W(x)]}$.[3]

Now that we have modeled the travel time of commuters, we next consider the costs associated with their commutes. The cost that a commuter incurs consists of three elements: travel time cost, work schedule penalty, and parking cost. First, the travel time cost is the sum of the cost associated with the 'passengered' travel time and the cost associated with the 'passengerless' travel time. The part of the morning commute during which AVs carry passengers includes inbound queueing delay τ_I and drop-off congestion time τ_W, and the part of this trip during which AVs do not carry passengers includes free-flow travel time $t_{W,1}$ for AVs that choose Area 1, and free-flow travel time $t_{D,2}$ and congestion time in D, $\tau_{D,2}$, for AVs that choose Area 2. Let α and α' (where $\alpha > \alpha' > 0$) represent the monetary values of one unit of passengered travel time and one unit of passengerless travel time, respectively. Both of these travel costs include the vehicle usage costs (e.g., gas/electricity, depreciation, mileage, etc.), and the passengered travel cost also accounts for the value of commuters' time. The travel time cost is then equal to $\alpha[\tau_I(x) + \tau_W(x)] + \alpha' t_{W,1}$ and $\alpha[\tau_I(x) + \tau_W(x)] + \alpha'[\tau_{D,2}(x) + t_{D,2}]$ for commuters who choose Area 1 and Area 2, respectively.

The second cost item is a work schedule penalty, which is a penalty that a commuter incurs when she arrives before or after time T. Following the prior literature (e.g., Arnott et al. (1991), Qian et al. (2011), and Liu (2018)), we define a work schedule penalty as the difference between the actual arrival time, $y = x + \tau_I(x) + \tau_W(x)$, and the official start time at work, T. Let β and γ represent the monetary cost of early and late start of work, respectively. Then, the work schedule penalty for commuters who arrive at W early is equal to $\beta(T - y)$, and that for commuters who arrive late is equal to $\gamma(y - T)$. Finally, a commuter who leaves H at time x and chooses Area 1 pays parking fee $p_1(x)$, while a commuter who leaves H at time x and chooses Area 2 does not incur a parking cost (see Assumption A1).

[3] For the drop-off congestion time $\tau_W(x)$, the numerator is the total number of AVs that are present at the drop-off zone at time $x + \tau_I(x)$ (which is the time when commuters who leave H at time x arrive at the drop-off zone in W), and the denominator of $\tau_W(x)$ is the average drop-off rate at time $x + \tau_I(x)$. For the congestion time $\tau_{D,2}(x)$ in D, the numerator is the total number of AVs that are present in D and head to Area 2 at time $x + \tau_I(x) + \tau_W(x)$ (which is the time when AVs that leave H at time x and head to Area 2 drop off their commuters at W), and the denominator of $\tau_{D,2}(x)$ is the average travel rate in D at time $x + \tau_I(x) + \tau_W(x)$.

Before defining the total cost of commuters, we state our final assumptions:

A6. The penalty of starting work early is lower than the monetary value of passengered travel time, i.e., $0 < \beta < \alpha$. This is the standard assumption in the literature, and it is empirically supported (e.g., Small (1982)). In addition, similar to Liu (2018), the marginal increase in the parking fee is lower than the marginal decrease in the work schedule penalty, i.e., $p_1'(x) \leq \beta \ (< \alpha)$ for $x \in [0, x_{max}]$. This means that a commuter prefers to arrive to work closer to time T, because if she delays her arrival time at W by one unit of time, her marginal saving in the work schedule penalty (β) is higher than her marginal increase in the cost of parking downtown ($p_1'(x)$).

A7. No commuter intentionally decides to arrive late at work. In other words, the monetary value of arriving late at work, γ, is so high that all commuters are dropped off at W before the work start time T. In Appendix B.5, we show that our main results hold when this assumption is relaxed.

Putting the three cost elements together, we can express the total cost of a commuter who departs H at time x and chooses Area 1 or Area 2, respectively, as follows:

$$C_1(x) = \alpha[\tau_I(x) + \tau_W(x)] + \beta(T - y) + \alpha' t_{W,1} + p_1(x), \tag{2.2}$$

$$C_2(x) = \alpha[\tau_I(x) + \tau_W(x)] + \beta(T - y) + \alpha'[\tau_{D,2}(x) + t_{D,2}]. \tag{2.3}$$

Lastly, we define the total system cost (also known as the social cost) as follows.

$$\int_0^{x_{max}} \lambda_H(x)\{\alpha[\tau_I(x) + \tau_W(x)] + \beta(T - y)\}dx + \int_0^{y_{max}} \lambda_1(y)\alpha' t_{1,W} + \lambda_2(y)\alpha'[t_{2,D} + \tau_{2,D}(x)]dy. \tag{2.4}$$

The cost in (2.4) consists of two terms: The first term is the sum of the passengered congestion cost and the work schedule penalty for all commuters. The second term is the sum of the passengerless congestion cost for all AVs that choose Area 1 ($\int_0^{y_{max}} \lambda_1(y)\alpha' t_{1,W} dy$) and the passengerless congestion cost for all AVs that choose Area 2 ($\int_0^{y_{max}} \lambda_2(y)\alpha'[t_{2,D} + \tau_{2,D}(x)]dy$). The total system cost does not include the parking fee $p_1(x)$, because the parking fees paid by the commuters who go to Area 1 cancel out the parking fees collected by the social planner. In other words, the parking revenues collected by the social planner is considered as a part of social welfare, so it is not counted towards the total system cost.

In our subsequent analyses, we illustrate our analytical results using the parameter values

estimated from the Pittsburgh Metropolitan Area. We summarize our calibrated parameter values in Table 1 while presenting details in Appendix B.2.

Table 2.1: Summary of the calibrated model parameters

Parameter	Value	Parameter	Value	Parameter	Value
R_I	4,600 vehicles per hour	$t_{W,1}$	2 minutes	K_1	10,000 AVs
N	20,000 commuters	$t_{D,2}$	15 minutes	R_W	3,600 drop-offs per hour
$p_1(0)$	\$1.80	α	\$4.50 per hour	θ	1
$p_1(x_{max})$	\$16	α'	\$2.25 per hour	a	0.2
$p_1(x)$	$14.20x/x_{max} + 1.80$ dollars	β	\$3.90 per hour	M	59.8

2.4 User Equilibrium Analysis

In this section, we present a user equilibrium (UE) in which individual commuters decide on their departure times and parking locations to minimize their costs. We define a UE as an equilibrium that satisfies the following two conditions:

Condition 1: $C_1(x) = C_2(x)$ for any $x \in [0, x_{max}]$ such that $\lambda_1(y) \neq 0$ and $\lambda_2(y) \neq 0$.

Condition 2: For each $j \in \{1, 2\}$, $\frac{\partial C_j(x)}{\partial x} = 0$ for any $x \in [0, x_{max}]$ such that $\lambda_j(y) \neq 0$.

Conditions 1 and 2 guarantee that all commuters incur the same total cost, regardless of their choices of parking location and departure time, respectively. This means that, under a UE, no commuter can unilaterally change her departure time and/or parking location to reduce her cost.

We next characterize the UE solutions. Proposition 2.1 indicates that there are two possible forms of a UE: UE1 and UE2. We use the following additional notation to describe these equilibria: x_j and $x_{j,max}$ for $j \in \{1, 2\}$ respectively denote the earliest and latest departure times of the commuters who choose Area j, and y_j is equal to $x_j + \tau_I(x_j) + \tau_W(x_j)$. Proofs are provided in Appendix B.3.

Proposition 2.1. *(a) [UE1] Suppose $\alpha' t_{D,2} \leq \alpha' t_{W,1} + p_1(0)$. There exists a UE which is presented in Table 2.2, where $A = \frac{\alpha(\int_0^y R_W - a\theta e^{\theta z} dz)^2}{(\alpha-\beta)(\int_0^y R_W - a\theta e^{\theta z} dz)^2 - \alpha'(\int_0^y a\theta e^{\theta z} dz)^2 + \alpha' a\theta e^{y} R_W y^2} R_W$, $B = \frac{\alpha - p_1'(x)}{\alpha - \beta} R_W$, $y_1 = x_1 + \alpha'[t_{D,2} + \tau_{D,2}(x_1)]$, $C = ae^{\theta y} + \frac{yae^{\theta y} + \int_0^y a\theta e^{\theta z} dz}{t_{W,1} - t_{D,2} + p_1(x)/\alpha'} - \frac{[yp_1'(x)/\alpha'] \int_0^y a\theta e^{\theta z} dz}{[t_{W,1} - t_{D,2} + p_1(x)/\alpha']^2}$, and x_1 satisfies $\alpha' t_{W,1} + p_1(x_1) = \alpha'[t_{D,2} + \tau_{D,2}(x_1)]$.*
(b) [UE2] Suppose $\alpha' t_{D,2} \geq \alpha' t_{W,1} + p_1(0)$. There exists a UE which is presented in Table 2.3, where $D = ae^{\theta(y-y_2)} + \frac{(y-y_2)ae^{\theta(y-y_2)} + \int_0^{y-y_2} a\theta e^{\theta z} dz}{t_{W,1} - t_{D,2} + p_1(x)/\alpha'} - \frac{p_1'(x)[(y-y_2)/\alpha'] \int_0^{y-y_2} a\theta e^{\theta z} dz}{[t_{W,1} - t_{D,2} + p_1(x)/\alpha']^2}$.

Proposition 2.1 shows that there exist two different user equilibria: UE1 and UE2. In addition, as presented in Table 2.2 and Table 2.3, depending on the capacity of Area 1, K_1, UE1 can take

Table 2.2: A characterization of UE1.

Condition	$\lambda_H(x)$	$\lambda_1(y)$	$\lambda_2(y)$
(i) $K_1 \geq N - R_W y_1 - a(e^{\theta(N/R_W - y_1)} - 1)[1 + \frac{N/R_W}{t_{W,1} - t_{D,2} + p_1(x_{max})/\alpha'}]$	A for $0 \leq x \leq x_1$ B for $x_1 \leq x \leq x_{max}$	0 for $0 \leq y \leq y_1$ $R_W - C$ for $y_1 \leq y \leq y_{max}$	R_W for $0 \leq y \leq y_1$ C for $y_1 \leq y \leq y_{max}$
(ii) $K_1 < N - R_W y_1 - a(e^{\theta(N/R_W - y_1)} - 1)[1 + \frac{N/R_W}{t_{W,1} - t_{D,2} + p_1(x_{max})/\alpha'}]$	A for $0 \leq x < x_1$ or $x_2 \leq x \leq x_{max}$ B for $x_1 \leq x < x_2$	0 for $0 \leq y \leq y_1$ or $y_{1,max} \leq y \leq y_{max}$ $R_W - C$ for $y_1 \leq y \leq y_{1,max}$	R_W for $0 \leq y \leq y_1$ or $y_{1,max} \leq y \leq y_{max}$ C for $y_1 \leq y \leq y_{1,max}$

Table 2.3: A characterization of UE2.

Condition	$\lambda_H(x)$	$\lambda_1(y)$	$\lambda_2(y)$
(i) $K_1 \geq N - [\frac{N/R_W}{t_{W,1} - t_{D,2} + p_1(x_{max})/\alpha'} + 1]a(e^{\theta(\frac{N}{R_W} - y_2)} - 1)$	B for $0 \leq x \leq x_{max}$	R_W for $0 \leq y \leq y_2$ $R_W - D$ for $y_2 \leq y \leq y_{max}$	0 for $0 \leq y \leq y_2$ D for $y_2 \leq y \leq y_{max}$
(ii) $K_1 \leq \frac{R_W}{\alpha - \beta}[\alpha p_1^{-1}(\alpha'(t_{D,2} - t_{W,1})) - \alpha'(t_{D,2} - t_{W,1}) + p_1(0)]$	B for $0 \leq x < x_{1,max}$ 0 for $x_{1,max} \leq x < x_2$ A for $x_2 \leq x \leq x_{max}$	R_W for $0 \leq y \leq y_{1,max}$ 0 for $y_{1,max} < y \leq y_{max}$	0 for $0 \leq y < y_2$ R_W for $y_2 \leq y \leq y_{max}$
(iii) $[\alpha p_1^{-1}(\alpha'(t_{D,2} - t_{W,1})) - \alpha'(t_{D,2} - t_{W,1}) + p_1(0)](\frac{R_W}{\alpha - \beta}) < K_1 < N - a(e^{\theta(\frac{N}{R_W} - y_2)} - 1)[\frac{N/R_W}{t_{W,1} - t_{D,2} + p_1(x_{max})/\alpha'} + 1]$	B for $0 \leq x < x_2$ A for $x_2 \leq x \leq x_{max}$	R_W for $0 \leq y \leq y_2$ $R_W - D$ for $y_2 \leq y \leq y_{1,max}$ 0 for $y_{1,max} < y \leq y_{max}$	0 for $0 \leq y \leq y_2$ D for $y_2 \leq y \leq y_{1,max}$ R_W for $y_{1,max} < y \leq y_{max}$

two different forms (i) and (ii), and UE2 can take three different forms (i) to (iii). In §2.4.1 and §2.4.2 we provide an in-depth discussion of UE1 and UE2, respectively.

2.4.1 UE1

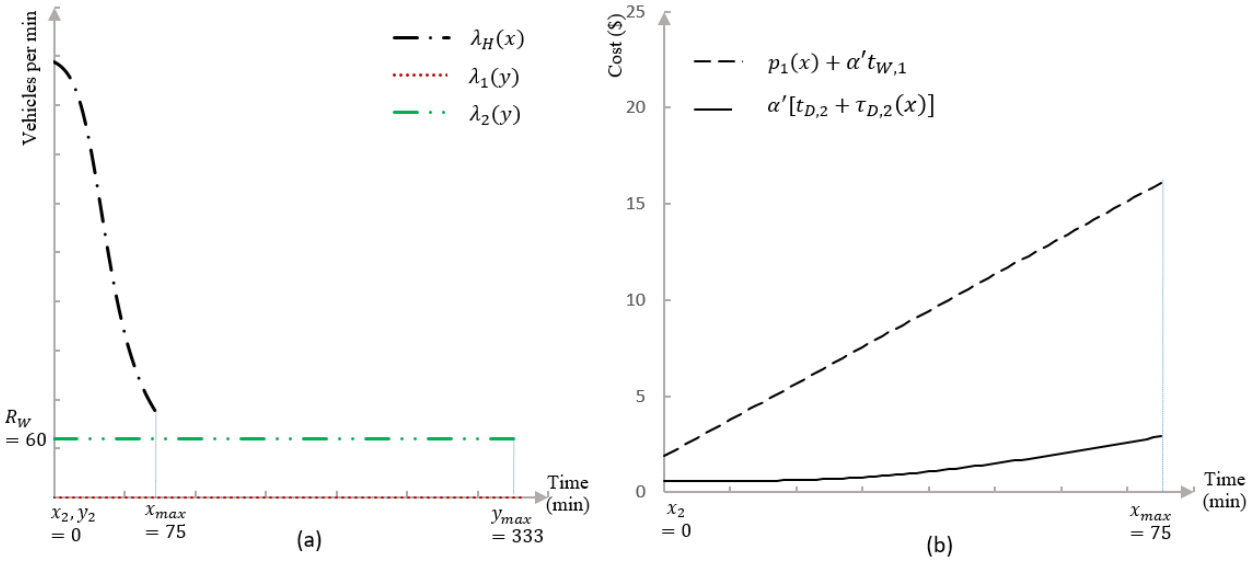
UE1 is observed under the condition that $\alpha' t_{D,2} \leq \alpha' t_{W,1} + p_1(0)$. This condition compares the cost associated with parking in Area 1 against that in Area 2. In determining which parking area is cheaper, commuters face a trade-off between the proximity of the parking location to W and the parking price. On the one hand, by parking in Area 1 that is closer to W, a commuter's AV does not go through the outbound highway to Area 2, nor does it experience any outbound congestion in D, saving the outbound free-flow travel time, $t_{D,2}$, and the congestion time in D, $\tau_{D,2}(x)$. On the other hand, the commuter incurs the cost of traveling to Area 1 and a high parking fee $p_1(x)$. Note that the cost associated with parking in one area entails not only its parking fee, but also the cost of traveling to that area. Specifically, $p_1(x) + \alpha' t_{W,1}$ is the cost associated with parking in Area 1 and $\alpha'[t_{D,2} + \tau_{D,2}(x)]$ is that in Area 2. UE1 is attained when it is cheaper for commuters to park in Area 2 at time zero (i.e., $\alpha'[t_{D,2} + \tau_{D,2}(0)] = \alpha' t_{D,2} \leq \alpha' t_{W,1} + p_1(0)$). As more commuters leave H over time, congestion time $\tau_{D,2}(x)$ increases, and when $\tau_{D,2}(x)$ becomes significantly high (i.e., D becomes very congested), some commuters start to park in Area 1; i.e., $\lambda_1(y) = 0$ for $y \leq y_1$ and $\lambda_1(y) > 0$ for $y > y_1$. In this respect, under UE1, Area 2 is the primary parking area chosen by commuters and Area 1 is the auxiliary parking area.

UE1 can take different forms depending on the capacity of Area 1, K_1. Under UE1, all commuters who leave H early (i.e., $x < x_1$) choose Area 2, but commuters who leave H late (i.e., $x \geq x_1$) may split between Area 2 and Area 1. In fact, if Area 1 has sufficient capacity (i.e., case (i) in Table 2.2 where $K_1 \geq N - a(e^{\theta N/R_W} - 1)[1 + \frac{N/R_W}{t_{W,1} - t_{D,2} + p_1(x_{max})/\alpha'}]$), some late commuters may choose this area. However, if Area 1 does not have enough capacity to accommodate all the commuters who choose this area (i.e., case (ii) in Table 2.2), this area becomes full at time $y_{1,max}$, and the remaining commuters have no choice but going to Area 2. When we use the calibrated model parameters presented in Table 2.1, we observe case (i) of UE1. In this case, the cost of traveling to Area 1 is higher than that to Area 2, so even late commuters do not choose Area 1. One may find this result rather surprising, but the observation of UE1 in our calibrated model is robust to a wide range of parameter values (see Appendix B.4.1).

We next discuss the commuters' decisions and offer insight into the departure pattern of commuters. Figures 2.2 and 2.3 depict commuters' departure rates and costs under UE1 for our calibrated model, respectively. Figure 2.2(a) illustrates that all commuters choose Area 2 over Area 1, i.e., $\lambda_1(y) = 0$ and $\lambda_2(y) = R_W$ for $0 \leq y \leq y_{max} = 333$ minutes. This is because, as depicted in Figure 2.2(b), at any time $x \in [0, x_{max} = 75]$, the cost associated with parking in Area 1, $p_1(x) + \alpha' t_{W,1}$, is higher than that in Area 2, $\alpha'[\tau_{D,2}(x) + t_{D,2}]$. This indicates that there is no need to have a designated parking area in Downtown Pittsburgh. This could occur in many other cities with a similar structure to Pittsburgh, where the downtown passengerless congestion cost is not as high as the cost of downtown parking. In these cities, unless a new policy is developed to regulate commuters' decisions (e.g., the policies we analyze in §2.6), we may observe a significant downtown land-use change (e.g., repurposing downtown parking spots to commercial and residential areas).

Figure 2.2(a) also illustrates that commuters prefer to leave H early (i.e., $\lambda_H(x)$ decreases over time). As depicted in Figure 2.3(a), all $N = 20,000$ commuters leave H relatively early, i.e., $x \in [0, 75]$. However, due to the limited inbound highway and drop-off capacities, the cumulative flow of commuters who arrive at W is much lower than the cumulative flow of commuters who leave H. In addition, since the travel rate in D decreases, the cumulative flow of AVs that arrive at Area 2 becomes even lower, creating congestion downtown. This figure also depicts the congestion time (inbound, drop-off, and downtown) that commuters who leave at any time $x \in [0, 75]$ experience. In particular, the inbound congestion time $\tau_I(x)$ is the horizontal distance from any point on

Figure 2.2: An illustration of case (i) of UE1 in the calibrated model: (a) departure rates, and (b) costs associated with parking in Areas 1 and 2.

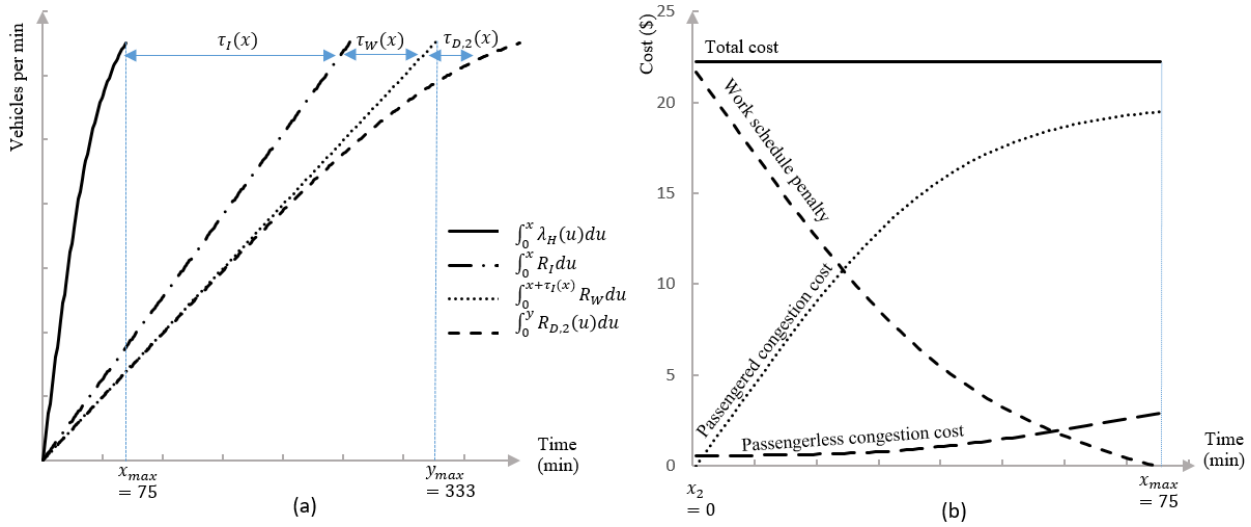


Notes. In (a), $\lambda_H(x)$ denotes the departure rate from H at time x, and $\lambda_1(y)$ and $\lambda_2(y)$ denote the departure rates from W at time y to Areas 1 and 2, respectively. In (b), $p_1(x) + \alpha' t_{W,1}$ represents the cost associated with parking in Area 1, and $\alpha'[t_{D,2} + \tau_{D,2}(x)]$ is that in Area 2.

the curve that represents the cumulative departure rate from H to the curve that represents the cumulative departure rate from the inbound highway. The drop-off congestion time $\tau_W(x)$ and the downtown congestion time $\tau_{D,2}(x)$ are illustrated similarly. The total congestion time together with the work schedule penalty determines the commuters' departure time decisions. The departure time decision is not trivial for commuters. In fact, as depicted in Figure 2.3(b), when a commuter decides on her departure time from H, she faces a trade-off between the work schedule penalty and the congestion costs. If the commuter wants to arrive close to the official work start time T (i.e., high values of x in this figure) to reduce her work schedule penalty, she might experience a higher passengered congestion cost (which includes the inbound and drop-off congestion costs) and passengerless congestion cost (which includes the cost of congestion in D). In fact, the choice of commuters to park in a cheaper parking area, i.e., Area 2, not only causes extra congestion in D, but also leads to an even higher congestion on the inbound bottleneck. To avoid these congestions, most commuters leave early (i.e., low values of x in Figure 2.3(b)) at the risk of paying some work schedule penalty.

The departure pattern from H in our model is different from the commuters' departure pattern

Figure 2.3: An illustration of case (i) of UE1 in the calibrated model: (a) cumulative flows, and (b) different cost components commuters incur.



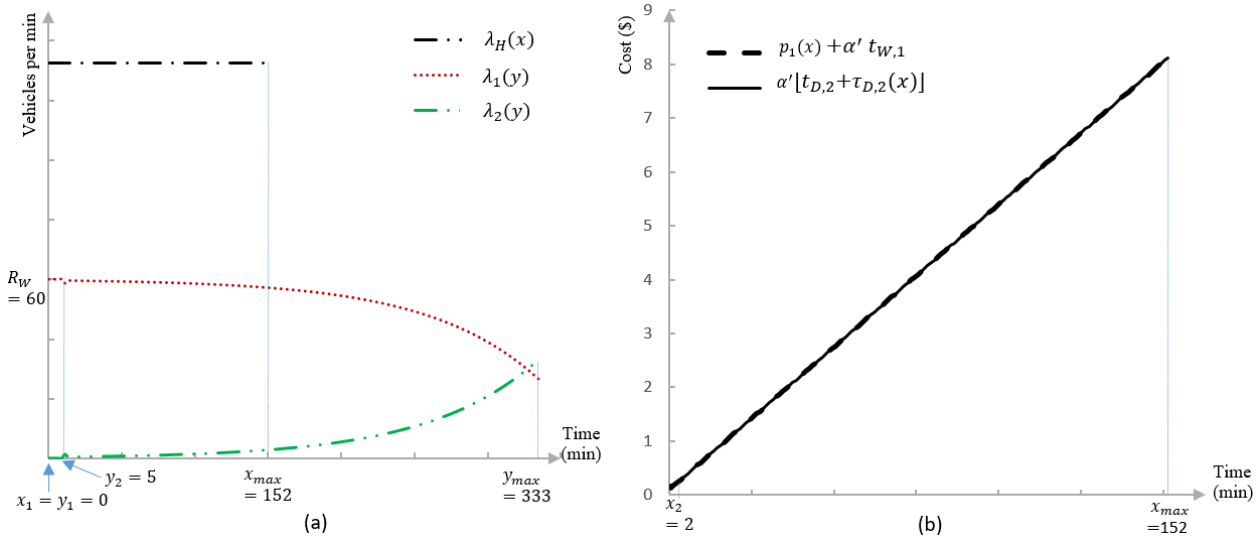
Notes. In (a), $\int_0^x \lambda_H(u)du$, $\int_0^x R_I du$, $\int_0^{x+\tau_I(x)} R_W du$, and $\int_0^y R_{D,2}(u)du$ represent cumulative departure rate from H, cumulative departure rate from I, cumulative departure rate from W, and cumulative departure rate from D to Area 2, respectively. In (b), the total cost $C_2(x)$ has three components: the work schedule penalty, which is equal to $\beta[T - x - \tau_I(x) - \tau_W(x)]$, the passengered congestion cost, which is equal to $\alpha[\tau_I(x) + \tau_W(x)]$, and the passengerless congestion cost, which is equal to $\alpha'[t_{D,2} + \tau_{D,2}(x)]$.

observed in previous studies. This is primarily because those studies overlooked congestion in D caused by commuters who are traveling to parking areas. For example, Arnott et al. (1991) consider the setting where commuters drive HVs, park their vehicles in Area 1, and walk to W. In their model, the only trade-off that a commuter faces is between the amount of time the commuter spends in congestion in I and her work schedule delay (which is affected by congestion time in I). Hence, commuters randomly choose their departure time from H, creating a constant flow out of the residential area H, i.e., $\lambda_H(x)$ is constant. Similarly, Liu (2018) considers the setting where an AV instantly drops off its commuter upon arrival to W and drives away to park in one of the parking spots, which are evenly distributed along a line from W within D. Liu (2018) shows that commuters fill up the parking spots close to W in the same order that they leave H. The AV commuters in that model face a similar trade-off to the HV commuters in Arnott et al. (1991) (because the cost of driving to an empty parking spot increases at a constant rate and can be consolidated with the work schedule penalty), so the departure rate from H, $\lambda_H(x)$, remains constant. However, as discussed before, our model offers deeper insights into the parking decision

behavior of AVs, as we consider the effect of the rapidly growing congestion in D caused by AVs that travel outside downtown to park.

Finally, we examine the effect of AVs on the morning commute by comparing our results with a morning commute model for HVs. In this model, commuters drive HVs, park in Area 1, and walk to W. These commuters do not experience any drop-off congestion at W (because HVs do not have the ability to drop off their commuters and self-drive to Area 1 or 2), nor do they have the option to park outside downtown (because the walking time to W is prohibitively high). However, they might cause and experience some congestion as they drive around downtown to find empty parking spots; also known as the parking cruising congestion time. To have a fair comparison between the HV case and the AV case, we also allow AV commuters in our model to drive to Area 1, park their AVs and walk to W. We next summarize our main findings for the calibrated model while providing details in Appendix B.4.5. First, AVs reduce the travel cost of a commuter from \$22.61 in the HV case to \$22.23, but their resultant total system cost increases from $\$283,243$ in the HV case to $\$444,600$. This is because the AV commuters experience drop-off congestion and downtown congestion, so they incur high congestion costs, which increase the total system cost substantially; in contrast, HV commuters park downtown and incur higher parking fees that are not included in the total system cost. Second, due to limited drop-off curbside space, AVs extend the duration of the morning commute (i.e., the time period between when the first group of commuters leave H until the time that last group of commuters arrive at W) from 298 minutes in the HV case to 333 ($= y_{max}$) minutes. In addition, AVs change the travel pattern from H to W. Whereas in the HV case commuters leave H over a span of 254 minutes, in our model commuters' departure window is 75 ($=x_{max}$) minutes (see Figure 2.2(a)). This means that AV commuters cause a sudden jam in the inbound bottleneck early on, but HV commuters gradually increase the inbound congestion. As a result, the total vehicle hours traveled (VHT) of AVs ($54,875$ hours) is much longer than that of HVs ($5,571$ hours). AVs also cause a higher vehicle miles traveled (VMT) than HVs, because all AVs travel to Area 2 which is farther than Area 1 from W. Therefore, for cities where the parking fee of the central area outweighs the cost of travel to the external parking area, the adoption of AVs is beneficial for individual commuters, but it does not necessarily lower the total system cost. Therefore, it is essential for city planners to make policies on parking demand and infrastructure that influence commuters' decisions and reduce the total system cost, as we discuss later in §2.6.

Figure 2.4: An illustration of case (i) of UE2: (a) departure rates, and (b) costs associated with parking in Areas 1 and 2.



Notes. In (a), $\lambda_H(x)$ denotes the departure rate from H at time x, and $\lambda_1(y)$ and $\lambda_2(y)$ denote the departure rates from W at time y to Areas 1 and 2, respectively. In (b), $p_1(x) + \alpha' t_{W,1}$ represents the cost associated with parking in Area 1, and $\alpha'[t_{D,2} + \tau_{D,2}(x)]$ is that in Area 2. In this numerical example, we use the same parameter values as in Table 2.1, except for the Area 1 parking fee $p_1(x)$, travel time to Area 2, $t_{D,2}$, and the capacity of Area 1, K_1 (see Appendix B.4.2 for more details).

2.4.2 UE2

UE2 is observed under the condition that $\alpha' t_{D,2} > \alpha' t_{W,1} + p_1(0)$. This condition indicates that commuters choose to park in Area 1 at time zero, since at time zero the cost of parking in Area 2 (i.e., $\alpha'[t_{D,2}+\tau_{D,2}(0)] = \alpha' t_{D,2}$) is higher than that in Area 1 (i.e., $\alpha' t_{W,1}+p_1(0)$). As the departure time x from H approaches the ideal start time at work, T, the parking fee $p_1(x)$ increases, and when $p_1(x)$ becomes significantly high, some commuters park in Area 2; i.e., $\lambda_1(y) = R_W$ and $\lambda_2(y) = 0$ for $y \leq y_2$, and $\lambda_1(y), \lambda_2(y) > 0$ for $y > y_2$. Thus, Area 1 is the primary parking area chosen by commuters and Area 2 is the auxiliary parking area. Cities with small downtown areas that have relatively cheap parking fees and high passengerless travel costs may observe UE2.

Under UE2, Area 1 and Area 2 are chosen by commuters in the reverse order of that under UE1. In other words, all commuters who leave H early (i.e., $x < x_2$) choose Area 1, but commuters who leave H late (i.e., $x \geq x_2$) split between Area 1 and Area 2. UE2 can take different forms depending on the capacity of Area 1, K_1. In particular, if Area 1 has enough capacity to accommodate

all commuters who choose this area (i.e., case (i) in Table 2.3), this area is chosen during the entire duration of the morning commute, while Area 2 may also be chosen simultaneously by some commuters. Figure 2.4(a) illustrates the auxiliary role of Area 2 under case (i) of UE2 using a numerical example. In this example, as Figure 2.4(b) shows, for a short period of time $x \in [0, 2]$, Area 1's low parking fee $p_1(x)$ leads to a slightly lower cost of choosing Area 1, i.e., $\alpha' t_{W,1} + p_1(x)$, than that of Area 2, i.e., $\alpha' t_{D,2}$, so all commuters who leave H before $x_2 = 2$ minutes choose Area 1. Note that as long as Area 2 is not chosen by any commuters, there is no downtown congestion, i.e., $\tau_{D,2}(x) = 0$ for $x \leq 2$. At $x_2 = 2$ minutes, due to the increased parking fee for Area 1, the cost of going to Area 2 becomes equal to the cost of going to Area 1, and some commuters start to go to Area 2. This results in $\lambda_2(y)$ becoming positive at $y_2 = 5$ minutes, which is the time when commuters who leave H at $x_2 = 2$ minutes are dropped off at W, i.e., $y_2 = x_2 + \tau_I(x_2) + \tau_W(x_2)$. When $x > x_2 = 2$ minutes, the cost of going to Area 2 remains equal to the cost of going to Area 1, so both Area 1 and Area 2 continue to be chosen by commuters. Figure 2.4(a) also illustrates that $\lambda_H(x) = 132.4$ commuters per minute is constant, which is different from what we observed in Figure 2.2. Unlike UE1, in this case having Area 1 as the primary parking area chosen by commuters prevents the congestion level in D from growing exponentially, so commuters can leave H evenly during a longer period of time than what we observed in Figure 2.2. This departure behavior of commuters prevents the creation of a substantial level of congestion on the inbound highway early in the morning.

We also compare this numerical example of UE2 with an analogous numerical example for the case when all commuters use HVs. As in §2.4.1, in the HV case, commuters park in Area 1 and walk from this area to W. In this case the departure window from H is equal to 316 minutes, which is longer than the departure window of 152 minutes in the AV case. In fact, unlike UE1 in our calibrated model, AVs not only reduce the duration of the morning commute, but also reduce the total system cost from $\$518,351$ in the HV numerical example to $\$171,054$. This happens because AVs primarily choose Area 1, so having the option to park in Area 2, which is an additional option available only to AVs, lowers the total system cost. Therefore, for cities where downtown parking fees are low and/or Area 2 is located relatively far from downtown, the adoption of AVs is beneficial and reduces the cost of the morning commute.

There exist two other cases of UE2 based on Area 1's parking capacity K_1. Similar to case (i),

in these two cases, early commuters choose Area 1, but due to the limited capacity of Area 1, when this area becomes full, all remaining commuters have to go to Area 2. In particular, in case (ii) shown in Table 2.3, Area 1's capacity is so low that commuters who wish to find a spot in Area 1 must leave H earlier than they would have if Area 1's capacity was as high as that in case (i). Once Area 1 reaches its full capacity at $x_{1,max}$, no commuter leaves H until time x_2. This is because the remaining commuters have to go to Area 2, which has abundant capacity, so there is no need to rush. In case (iii) shown in Table 2.3, Area 1 has a moderate capacity. In this case, commuters do not need to leave as early as they do in case (ii), but when Area 1 becomes full, all the remaining commuters must go to Area 2. Further discussion of these cases is provided in Appendix B.4.2.

2.5 Social Optimum

We analyze the case in which a social planner dictates the departure rates from H, i.e., $\lambda_H(x)$ for $x \in [0, x_{max}]$, and the departure rates of AVs from W to Areas 1 and 2, i.e., $\lambda_j(y)$ for $y \in [0, y_{max}]$ and $j \in \{1, 2\}$, that minimize the total system cost in (2.4). This is different from the UE, under which the goal of a commuter is to minimize her own travel cost regardless of how her decision affects other commuters. As mentioned in §2.4, commuters' decisions under UE may cause congestion, hence increasing the total system cost. This is not a desirable social outcome, so we find the *socially optimum* (SO) that minimizes the total system cost. We can state this problem as follows:

$$\min_{\lambda_H(x),\lambda_1(y),\lambda_2(y)} \int_0^{x_{max}} \lambda_H(x)\{\alpha[\tau_I(x)+\tau_W(x)]+\beta(T-y)\}dx + \int_0^{y_{max}} \lambda_1(y)\alpha' t_{1,W} + \lambda_2(y)\alpha'[t_{2,D}+\tau_{2,D}(x)]dy$$

$$\begin{aligned} \text{subject to: } & y = x + \tau_I(x) + \tau_W(x) \\ & \lambda_1(y) + \lambda_2(y) = \min\{\lambda_H(x), R_W\} \\ & 0 \le \lambda_H(x) \le N \\ & 0 \le \lambda_1(y), \lambda_2(y). \end{aligned} \tag{2.5}$$

In the following proposition, we describe the values of $\lambda_H(x)$, $\lambda_1(y)$ and $\lambda_2(y)$ that minimize the total system cost in (2.5) and satisfy the four constraints.[4] Proposition 2.2 demonstrates that

[4]The first constraint describes the relationship between the departure time x from H and the departure time y from W, as presented earlier in (2.1). The second constraint states that the number of AVs that travel from W to the parking areas at time y, i.e., $\lambda_1(y) + \lambda_2(y)$, is equal to the number of AVs that drop off their commuters at time y, $\min\{\lambda_H(x), R_W\}$. The third constraint guarantees that the departure rate $\lambda_H(x)$ from H is positive and it does not exceed the total number N of commuters. The last constraint guarantees non-negative $\lambda_1(y)$ and $\lambda_2(y)$.

there are two forms of SOs: SO1 and SO2. Let x_D denote the earliest departure time from H for commuters who go to Area 2 and experience downtown congestion.

Proposition 2.2. *(a) [SO1] Suppose $t_{D,2} \geq t_{W,1}$. There exists an SO which is presented in Table 2.4, where $y_2 = x_2 = K_1/R_W$, x_D satisfies $\beta a x_D(R_W - a)/R_W + \beta a^2 K_1/R_W^2 - \alpha'(R_W - a)\tau_{D,2}(T - x_D) = 0$, $y_{max} = x_{max} = T = \frac{N}{R_W}$ for cases (i), and $y_{max} = x_{max} = T = \frac{N + a(x_D - x_2)}{R_W}$ for case (ii).*

Table 2.4: A characterization of SO1.

Condition	$\lambda_H(x)$	$\lambda_1(y)$	$\lambda_2(y)$
(i) $K_1 \geq N$	R_W for $0 \leq x \leq x_{max}$	R_W for $0 \leq y \leq y_{max}$	0 for $0 \leq y \leq y_{max}$
(ii) $K_1 < N$	R_W for $0 \leq x < x_2$ and $x_D \leq x \leq x_{max}$ $R_W - a$ for $x_2 \leq x \leq x_D$	R_W for $0 \leq y < y_2$ 0 for $y_2 \leq y \leq y_{max}$	0 for $0 \leq y < y_2$ $R_W - a$ for $y_2 \leq y < y_D$ R_W for $y_D \leq y \leq y_{max}$

(b) [SO2] Suppose $t_{D,2} < t_{W,1}$. There exists an SO which is presented in Table 2.5, where $x_D = y_D = \frac{\alpha' R_W}{a\beta}\tau_{D,2}(\min\{y_1, T - x_D - y_2\lfloor\frac{T - x_D}{y_2}\rfloor\})$, $y_1 = \tau_D^{-1}(t_{W,1} - t_{D,2})$, $y_2 = y_1 + t_{W,1} - t_{D,2}$, $y_{1,max}$ satisfies $K_1/R_W = y_{1,max} - y_1(\frac{N - (R_W - a)x_D}{R_W y_2} + 1)$, $y_{max} = x_{max} = T = \frac{N + a x_D}{R_W}$, and $n = 0, 1, \cdots, \lfloor\frac{N - (R_W - a)x_D}{R_W y_2}\rfloor$.[5]

Table 2.5: A characterization of SO2.

Condition	$\lambda_H(x)$	$\lambda_1(y)$	$\lambda_2(y)$
(i) $K_1 \geq (\frac{N - (R_W - a)x_D}{R_W y_2} + 1)y_1 R_W$	$R_W - a$ for $0 \leq x < x_D$ R_W for $x_D \leq x \leq x_{max}$	0 for $ny_2 \leq y \leq \min\{y_D + ny_2 + y_1, y_{max}\}$ R_W otherwise	R_W for $ny_2 \leq y \leq \min\{y_D + ny_2 + y_1, y_{max}\}$ 0 otherwise
(ii) $K_1 < (\frac{N - (R_W - a)x_D}{R_W y_2} + 1)y_1 R_W$	$R_W - a$ for $0 \leq x < x_D$ R_W for $x_D \leq x \leq x_{max}$	0 for $ny_2 \leq y < y_D + ny_2 + y_1$ and $y_D + y_{1,max} \leq y \leq y_{max}$ R_W otherwise	R_W for $ny_2 \leq y < y_D + ny_2 + y_1$ and $y_D + y_{1,max} \leq y \leq y_{max}$ 0 otherwise

Proposition 2.2 states that, depending on the locations of Areas 1 and 2, the social planner's decision on commuters' departure times and parking locations follows either SO1 or SO2. In addition, depending on the capacity of Area 1, K_1, there exist two forms (i) and (ii) that each of SO1 and SO2 takes. In §2.5.1 and §2.5.2 we discuss when SO1 and SO2 are attained, respectively.

2.5.1 SO1

SO1 is observed when the travel time from D to Area 2 is longer than that from W to Area 1, i.e, $t_{D,2} \geq t_{W,1}$. In this case, regardless of the congestion cost in D, the cost of routing a commuter to Area 1 ($\alpha' t_{W,1}$) is lower than that to Area 2 ($\alpha'[t_{D,2} + \tau_{D,2}(x)]$), so the social planner routes AVs to Area 1 as long as this area is not full, and to Area 2 afterwards. SO1 takes two different

[5]The parameter n counts the number of switches from routing AVs to Area 2 to routing them to Area 1.

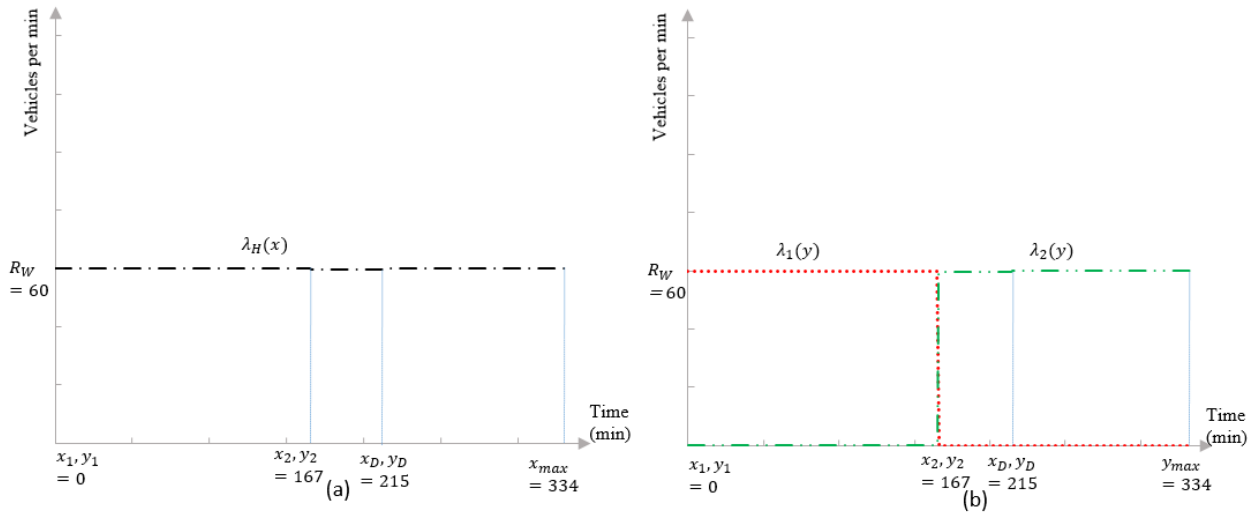
forms depending on the capacity K_1 of Area 1. When Area 1 can accommodate all commuters (i.e., case (i) where $K_1 \geq N$), the social planner routes all commuters to this area, i.e., $\lambda_1(y) = R_W$ and $\lambda_2(y) = 0$. In this case, the social planner sets the rate at which commuters leave H, $\lambda_H(x)$, equal to the drop-off rate at W, R_W, to eliminate the passengered congestion (which consists of the inbound and drop-off congestions). The drop-off rate at W determines the departure rate from H because, when all commuters are routed to Area 1, the drop-off space becomes the most downstream bottleneck that AVs go through. This means that the focus of the social planner should shift from creating parking spaces downtown to creating drop-off curbside spaces. When Area 1 does not have enough capacity (i.e., case (ii) where $K_1 < N$), as this area becomes full at time y_2, the remaining commuters must go to Area 2. In our calibrated model with the parameter values presented in Table 2.1, we observe case (ii) of SO1, as it satisfies the two conditions for this case: $t_{D,2} = 15 > 2 = t_{W,1}$ and $K_1 = 10,000 < 20,000 = N$. In Appendix B.4.3, we show that the observation of case (ii) of SO1 is robust over a wide range of model parameters.

Figure 2.5 illustrates the SO departure rates for our calibrated model. When $0 \leq x < x_2 = y_2 = 167$ minutes, all commuters are routed to Area 1, and the departure rate from H is equal to the drop-off capacity, i.e., $\lambda_H(x) = \lambda_1(y) = 60$ AVs per minute for $x = y \in [0, 167)$.[6] At $y_2 = 167$ minutes, all $10,000$ parking spots in Area 1 are filled, and the social planner must route the remaining $10,000$ commuters to Area 2. These AVs might create congestion in D, so the social planner may set the departure rate from home, $\lambda_H(x)$, to either the travel rate of AVs in D, $R_{D,2}(y)$, or the drop-off capacity R_W. The former option has a higher aggregate work schedule penalty than the latter (because, by Assumption A4, $R_{D,2}(y)$ is less than or equal to R_W), but it has no passengerless congestion cost (because the rate of AVs that depart H becomes equal to the the number of AVs that can travel in D, $R_{D,2}(y) = R_W - a = 60 - 0.2 = 59.8$ AVs per minute).[7] To balance the congestion cost in D and the work schedule penalty, the social planner first sets the departure rate from H equal to the travel rate in D, i.e., $\lambda_H(x) = \lambda_2(y) = R_{D,2}(y) = 59.8$ AVs per minute for $x = y \in [x_2 = 167, x_D = 215)$ minutes, and then increases it to the drop-off capacity as the ideal start time at work, T, approaches; i.e., $\lambda_H(x) = \lambda_2(y) = R_W = 60$ AVs per minute

[6] Given that the passengered congestion time, $\tau_I(x) + \tau_W(x)$, is eliminated under SO, the drop-off time at W, $y = x + \tau_I(x) + \tau_W(x)$, is the same as the departure time x from H. Note that this is due to our (innocuous) assumption of normalizing the free-flow travel time on the inbound highway to zero.

[7] Note that when there is no congestion in D, the travel rate $R_{D,2}(y)$ does not decrease from its maximum value of $R_W - a$.

Figure 2.5: An illustration of SO1 in the calibrated model: (a) departure rate from H, and (b) departure rates from W to Area 1 and Area 2.



Notes. In (a), $\lambda_H(x)$ denotes the departure rate from H. In (b), $\lambda_1(y)$ and $\lambda_2(y)$ denote the departure rates from W to Area 1 and Area 2, respectively.

for $x = y \geq x_D = 215$ minutes. As such, only AVs that leave home late incur a passengerless congestion cost.

The SO pattern observed in Figure 2.5 is different from the UE pattern depicted in Figure 2.2(a). In fact, the daily total system cost under SO ($\$218,728$) is less than half of that under UE ($\$444,600$). This discrepancy stems from high passengered congestion time and passengerless travel time under UE. Under SO, the departure rate from H is at most equal to the drop-off rate, i.e., $\lambda_H(x) \leq R_W = 60$ commuters per minute, and commuters do not experience any passengered congestion. In contrast, under UE, the departure rate $\lambda_H(x)$ from H is always higher than the drop-off rate $R_W = 60$ commuters per minute, so all commuters incur a positive passengered congestion cost (see Figure 2.3(b)). In addition, the passengerless travel time is longer under UE than under SO, because under UE the high parking fee at Area 1 deters the commuters from choosing this area even though it is closer than Area 2. The total system cost can be reduced even more, if Area 2 becomes closer to D. In Appendix B.4.4 we discuss an alternative location for Area 2 in Pittsburgh.

Finally, we compare SO1 with SO in the model discussed in §2.4.1 where all commuters drive HVs. We observe that AVs have three major impacts on the morning commute: (1) reducing the

total system cost, (2) changing the order at which parking areas are occupied, and (3) extending the morning commute time window. First, AVs reduce the total system cost from $\$252,905$ in the HV case to $\$218,728$ in SO1 (see Appendix B.4.5 for detail), because they do not experience downtown parking cruising congestion. Second, AVs change the order at which parking areas fill up from inwards to outwards: HVs fill parking spots from the farthest (which are also the cheapest) parking area towards W, but AVs fill Area 1 (which is the closest parking area to W) before Area 2. This discrepancy directly stems from the ability of AVs to drop off commuters at W before self-parking. In the HV case, the social planner prefers commuters to walk a longer distance from where they park to W to reduce the work schedule penalty. But AVs can drop off their commuters before they park, so their parking location has no impact on the work schedule penalty. Therefore, it is better for AVs to choose the closest parking area to reduce the commute cost. Lastly, the duration of the morning commute is longer in SO1 than that in the HV case. As mentioned before, the social planner aims to remove congestion by setting the departure rate from H equal to the capacity of the most downstream bottleneck, which is the inbound bottleneck for HVs. This departure rate is higher than the drop-off capacity in our model (since $R_W \leq R_I$ by Assumption A4), resulting in an expansion of the morning commute duration. This shows that the ability of AVs to drop off commuters at W does not necessarily lead to earlier arrival of commuters at W, as long as the curbside space in D is limited.

2.5.2 SO2

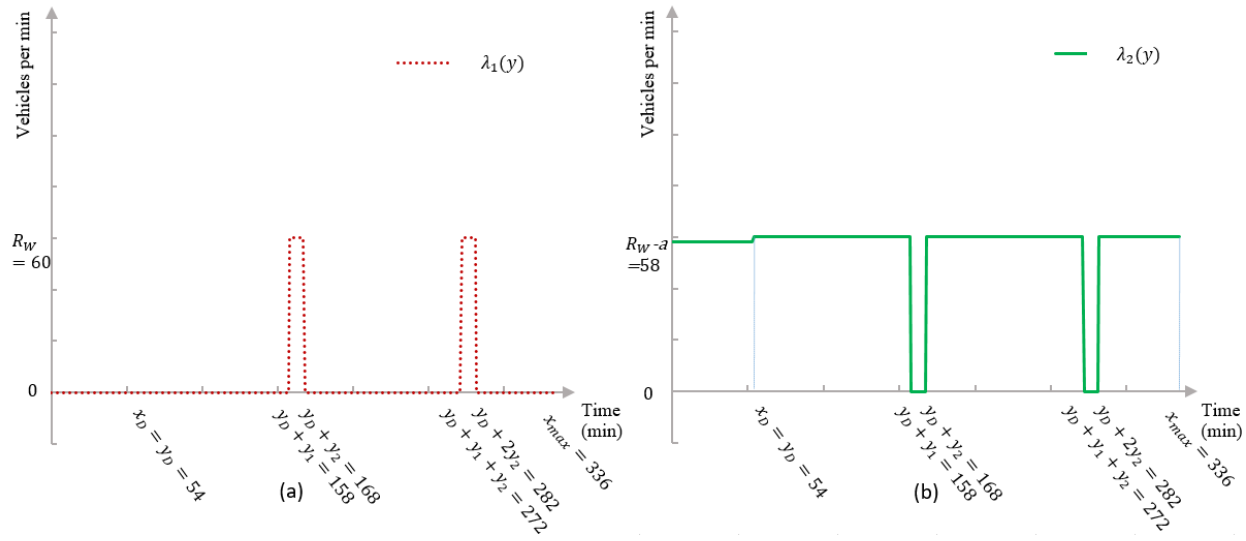
SO2 is observed in cities with large downtown areas where the travel time from D to Area 2 is shorter than that from W to Area 1, i.e, $t_{D,2} < t_{W,1}$. SO2 takes two forms depending on the capacity of Area 1, K_1. When Area 1 has enough capacity for all AVs that are routed to this area (i.e., case (i) where $K_1 \geq (\frac{N-(R_W-a)x_D}{R_W y_2}+1)y_1 R_W$), the social planner alternately routes AVs to Area 2 and Area 1. We illustrate this case using the same numerical example used in §2.4.2, except that $t_{W,1} = 15$ minutes. In this example, $t_{D,2} = 5 < 15 = t_{W,1}$ and $K_1 = 10,000 > N - (\lfloor \frac{N}{y_2 R_W} \rfloor + 1)y_1 R_W = 1,280$. Figure 2.6(a) and Figure 2.6(b) depict the departure rates from W to Area 1 and Area 2, respectively. The social planner wants to reduce the downtown congestion cost without increasing the duration of the morning commute substantially. As such, when $x \leq x_D = 54$ minutes, the social planner eliminates the downtown congestion by setting the departure rate from H equal to the travel rate

in D, i.e., $\lambda_H(x) = R_{D,2}(y) = R_W - a = 60 - 2 = 58$ AVs per minute for $x \in [0, 54]$, and routes all AVs to Area 2 (because the travel time from W to Area 2, i.e., $t_{D,2} + \tau_{D,2}(x) = t_{D,2}$, is lower than that from W to area 1, i.e., $t_{W,1}$). During this period of time, there is no downtown congestion, but the work schedule penalty is high due to the low departure rate of AVs from H. Hence, at time $x_D = 54$ minutes, to reduce the commuters' work schedule penalty, the social planner increases the departure rate from H to the drop-off capacity $R_W = 60$ commuters per minute (which creates downtown congestion because $R_W \geq R_{D,2}(y)$ by Assumption A4), and routes AVs to Area 2 until the downtown congestion becomes so high that the cost of routing commuters to Area 1 becomes lower than that to Area 2, i.e., $\alpha' t_{W,1} \leq \alpha'[t_{D,2} + \tau_{D,2}(x)]$; at that point, the social planner routes AVs to Area 1. When the downtown congestion is reduced, the social planner routes AVs to Area 2 again. This cycle of alternately routing AVs to Area 2 and Area 1 continues until all $20,000$ AVs are assigned to the parking locations. This is shown in Figure 2.6, where $\lambda_1(y) = 0$ and $\lambda_2(y) = R_W = 60$ AVs per minute for $y \in [54, 158]$, $y \in [168, 272]$ and $y \in [282, 336]$, and $\lambda_1(y) = R_W = 60$ AVs per minute and $\lambda_2(y) = 0$ for $y \in [158, 168]$ and $y \in [272, 282]$. So the social planner uses Area 1 as a mitigator for the downtown congestion by routinely taking a break from routing AVs to Area 2 and allowing the downtown congestion to dissolve completely. This is a different approach than what commuters do under UE: when the downtown congestion becomes high, although some commuters choose Area 1, there can still be commuters who choose Area 2. These commuters, who prefer experiencing downtown congestion to paying a high parking fee in Area 1, cause the downtown congestion to continue to grow (although at a slower rate, because some commuters go to Area 1).

As in SO1, if all commuter use HVs, then the departure rate from H increases from at most $R_W = 60$ to $R_I = 76.77$ commuters per minute. This high departure rate form H shortens the window of the morning commute for HVs. However, in our case, since the downtown congestion level stays low and there is no parking cruising congestion, AVs decrease the total system cost from $\$311,749$ in the HV case to $\$224,608$ in our model.

When Area 1 does not have enough capacity (i.e., case (ii) in Table 2.5), the social planner routes the remaining AVs to Area 2 as Area 1 becomes full at time $y_{1,max}$, i.e., $\lambda_1(y) = 0$ and $\lambda_2(y) = R_W$ for $y \geq y_{1,max}$. In this case, the earliest time when congestion forms downtown, x_D, is higher than that in case (i), because downtown congestion grows inevitably after $y_{1,max}$, so the

Figure 2.6: An illustration of SO2: (a) departure rate from W to Area 1, and (b) departure rate from W to Area 2.



social planner sends more AVs to Area 2 (i.e., increases x_D) when there is no downtown congestion.

2.6 Reducing the Total System Cost of the Morning Commute

As discussed in §2.5, commuters' decisions under UE are different from SO. In this section, we examine solutions that the social planner may adopt in order to close the gap between the total system cost under UE and that under SO: a short-term solution of regulating parking fees and congestion tolls in §2.6.1 and a long-term solution of adjusting parking and curbside drop-off capacities in §2.6.2.

2.6.1 Pricing and Tolling Schemes

To reduce the total system cost, the social planner can use two levers that are commonly used in practice (Federal Highway Administration 2020): parking fees and congestion tolls. We consider a dynamic parking pricing scheme that enables the social planner to regulate Area 1's parking fee $p_1(x)$ based on departure time x.[8] A tolling scheme $\pi_2(x)$, which can be interpreted as a road usage

[8]Recall that commuters decide on their parking location when they leave their homes. However, by (2.1), the departure time from H uniquely determines drop-off time, so Area 1's parking fee also depends on the drop-off time y.

charge, is imposed on commuters who choose Area 2 based on their departure time x to balance congestion in D. Proposition 2.3 describes these pricing and tolling schemes.

Proposition 2.3. *(a) Suppose $t_{2,D} \geq t_{1,W}$. Under equilibrium, $\lambda_H(x)$, $\lambda_1(y)$ and $\lambda_2(y)$ follow those in Table 2.4, when $p_1(x)$ and $\pi_2(x)$ for $0 \leq x \leq x_{max}$ are as presented in Table 2.6, where $\epsilon_1(x) \cdot \epsilon_2(x) = 0$, $\epsilon_1(x) = \alpha'(t_{D,2} - t_{W,1})$ for $x_2 \leq x$, and $\epsilon_2(x) = \alpha'(t_{D,2} - t_{W,1})$ for $x \leq x_2$.*

Table 2.6: A characterization of parking pricing and congestion tolling schemes for SO1

Case	$p_1(x)$	$\pi_2(x)$
(i)	βx for $0 \leq x \leq x_{max}$	$\beta x - \alpha'(t_{D,2} - t_{W,1}) + \epsilon_2(x)$ for $0 \leq x \leq x_{max}$
(ii)	$\beta x + \max\{\alpha'[t_{D,2} - t_{W,1} + \tau_{D,2}(x_{max})] - \beta T, \alpha'(t_{D,2} - t_{W,1}) - \beta x_2, 0\} + \epsilon_1(x)$ for $0 \leq x \leq x_{max}$	$\beta x + \max\{-\beta x_2, \alpha'\tau_{D,2}(x_{max}) - \beta T, \alpha'(t_{W,1} - t_{D,2})\} + \epsilon_2(x)$ for $x < x_D$ $\beta x - \alpha'\tau_{D,2}(x) + \max\{-\beta x_2, \alpha'\tau_{D,2}(x_{max}) - \beta T, \alpha'(t_{W,1} - t_{D,2})\}$ for $x \geq x_D$

(b) Suppose $t_{2,D} < t_{1,W}$. Under equilibrium, $\lambda_H(x)$, $\lambda_1(y)$ and $\lambda_2(y)$ follow those in Table 2.5, when $p_1(x)$ and $\pi_2(x)$ for $0 \leq x \leq x_{max}$ are as presented in Table 2.7, where $\epsilon_1(x) = \alpha'(t_{W,1} - t_{D,2})$ and $\epsilon_2(x) = 0$ for $nx_2 \leq x < \min\{x_D + nx_2 + x_1, x_{max}\}$ in case (i) and for $nx_2 \leq x < x_D + nx_2 + x_1$ and $x_D + x_{1,max} \leq x \leq x_{max}$ in case (ii), and $\epsilon_1(x) = 0$ and $\epsilon_2(x) = \alpha'(t_{W,1} - t_{D,2})$ otherwise.

Table 2.7: A characterization of parking pricing and congestion tolling schemes for SO2

Case	$p_1(x)$	$\pi_2(x)$
(i) and (ii)	$\beta x + \max\{-\alpha'(t_{W,1} - t_{D,2}), -\beta(x_D + x_1)\} + \epsilon_1(x)$ for $0 \leq x \leq x_{max}$	$\beta x + \max\{\alpha'(t_{W,1} - t_{D,2}) - \beta(x_D + x_1), 0\} + \epsilon_2(x)$ for $x < x_D$ $\beta x - \alpha'\tau_{D,2}(x) + \max\{\alpha'(t_{W,1} - t_{D,2}) - \beta(x_D + x_1), 0\} + \epsilon_2(x)$ for $x \geq x_D$

Proposition 2.3 indicates that there exist a parking fee scheme $p_1(x)$ and a congestion toll scheme $\pi_2(x)$ such that the SO presented in Proposition 2.2 results in the same travel cost for all commuters regardless of their departure time or their parking location. In other words, when these parking fees and congestion tolls are imposed, UE matches SO. This short-term solution is particularly important to the social planner, as it enables them to influence commuters' decisions and reduce the aggregate congestion and travel costs of the morning commute. Table 2.6 provides the two forms (i) and (ii) of the pricing and tolling schemes associated with the two forms (i) and (ii) of SO1, respectively. Similarly, Table 2.7 corresponds to SO2. We first discuss Table 2.6, which pertains to our calibrated model, and then Table 2.7 using the numerical example of SO2 discussed in §2.5.2.

For our calibrated model, Figure 2.7(a) illustrates the optimal parking fees and congestion tolls, presented in case (ii) of Table 2.6. Under these schemes, the total system cost of the morning commute in Pittsburgh is reduced by 51% (which amounts to $56.5 million in annual savings) as

compared to that under UE. When these parking fees and congestion tolls are imposed during the morning commute window, commuters' decisions on both departure time and parking location under UE follow those of SO1 depicted in Figure 2.5 rather than those in Figure 2.2(a). Specifically;

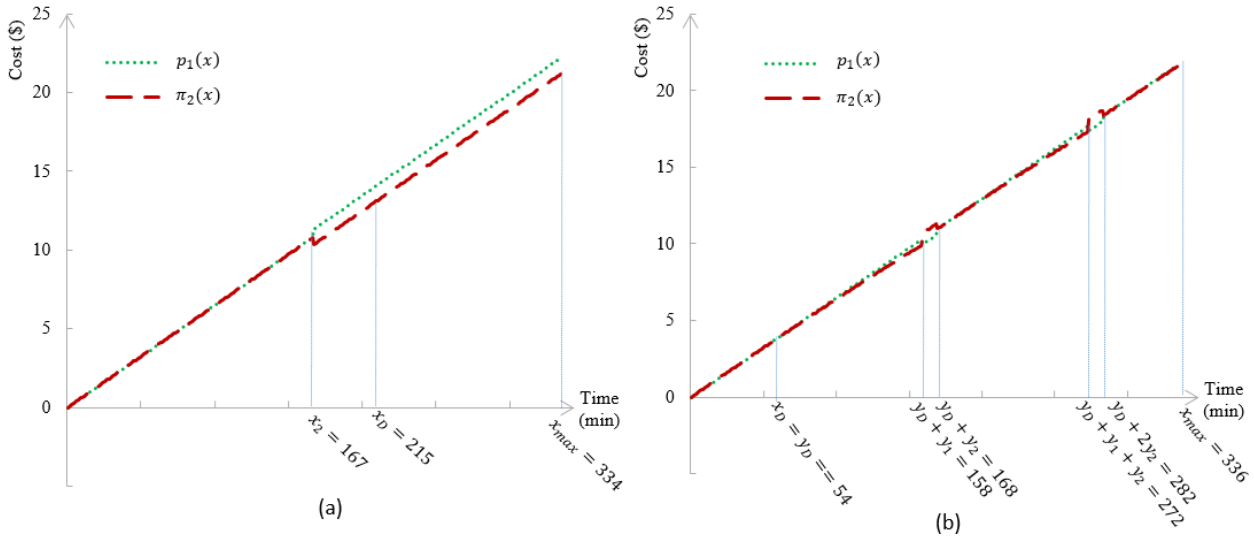
• To guarantee that the commuters' departure rate from H follows the SO pattern, for $x \leq x_2 = 167$ minutes, all commuters incur the same passengered and passengerless travel costs, but not the same work schedule penalty (which is equal to $\beta(T - x)$ for a group of commuters who leave at time x). Hence, the social planner sets Area 1's parking fee $p_1(x)$ equal to βx to ensure equal travel costs for all, i.e., $\beta(T-x)+\beta x = \beta T$. As the work schedule penalty linearly decreases in departure time x, the parking fee linearly increases, so all commuters who follow SO incur the same total travel cost. For $x > x_2 = 167$ minutes, the social planner needs to set the congestion toll $\pi_2(x)$ such that it not only leads to the same work schedule penalty for all commuters, but also assures the same passengerless travel cost for the commuters who go to Area 2 (i.e., $\alpha'(t_{W,1} - t_{D,2})$). In addition, for $x > x_D = 215$ minutes, the social planner reduces the congestion toll $\pi_2(x)$ by $\alpha'\tau_{D,2}(x)$, so it accounts for the extra congestion costs these commuters incur in D. This congestion toll scheme again leads to the same travel cost for all commuters. Thus, no commuter has an incentive to unilaterally deviate from this equilibrium, and the departure rate from H is the same as that in Figure 2.5.

• To guarantee that commuters abide by the social planner's parking location decisions, for commuters who leave H before $x_2 = 167$ minutes, the social planner sets the congestion toll $\pi_2(x)$ equal to the parking fee $p_1(x)$. Given that Area 1 is closer than Area 2, i.e., $t_{D,2} \geq t_{W,1}$, this congestion toll induces commuters to choose Area 1. For commuters who leave after x_2, the social planner increases Area 1's parking fee by a positive factor equal to $\alpha'(t_{D,2} - t_{W,1})$ to discourage commuters from choosing this area.[9]

Proposition 3 states that the social planner needs to adopt both dynamic parking fee and congestion toll schemes *simultaneously* in order to induce commuters to follow SO. One natural question is what happens if the social planner can impose one but not both. To address this question, we consider two benchmarks. We present the summary of our findings while presenting detail in Appendix B.4.6. First, we consider a benchmark when the social planner adjusts the

[9]The pricing and tolling schemes for case (i) of SO1 follow a similar logic. In particular, since Area 1 has enough capacity for all AVs in case (i) (i.e., $K_1 > N$), the parking pricing and congestion tolling schemes for case (i) is similar to those for case (ii) when Area 1 is chosen (i.e., $x \leq x_2$).

Figure 2.7: An illustration of parking fee scheme and congestion toll scheme for: (a) SO1, and (b) SO2.



downtown parking fee, but does not impose a congestion toll. In this case, since Area 1 is closer than Area 2, the social planner sets the parking fee such that commuters choose Area 1 as their preferred choice. This new parking pricing scheme leads to the same parking location decisions as those under SO, but the commuters' departure times do not match those under SO. In this case, the total system cost is reduced by 3% from $\$444,600$ under UE to $\$432,654$. Second, we consider a benchmark when the social planner does not adjust the parking fee, but imposes a congestion toll to persuade the commuters to choose Area 1. In particular, this congestion toll (which is a flat toll of $\$3.50$ for our calibrated model) is set in a way that the cost associated with choosing Area 1 is lower than the cost of choosing Area 2 as long as Area 1 is not full. Similar to the first benchmark, commuters' departure times from H do not follow those under SO, but the total system cost is reduced by 6% from $\$444,600$ under UE to $\$416,076$. These two benchmarks reveal the importance of adopting both dynamic parking fee and congestion toll schemes, which lead to substantially higher savings with a 51% reduction in the total system cost.

Lastly, we discuss the pricing and congestion tolling schemes, illustrated in Figure 2.7(b), that induce commuters to follow case (i) of SO2 for the numerical example discussed in §2.5.2. Similar to Figure 2.7(a), the parking fees and congestion tolls are set such that all commuters incur the same work schedule penalty. During the cycles in which Area 2 is chosen, the congestion toll $\pi_2(x)$ is reduced by $\alpha' \tau_{D,2}(x)$ to account for the extra congestion cost in D. In addition, during the cycles

in which Area 1 (resp., Area 2) is chosen, the amount of $\alpha'(t_{W,1} - t_{D,2})$ is added to the congestion toll $\pi_2(x)$ (resp., parking fee $p_1(x)$) to deter commuters from choosing Area 2 (resp., Area 1). Using these parking pricing and congestion tolling schemes, the social planner can reduce the total system cost by 49% from $\$437,150$ under UE to $\$224,608$.[10]

2.6.2 Improving the Infrastructure

Although parking pricing and congestion tolling are practical tools for reducing the cost of the morning commute, the social planner should also explore a long-term sustainable plan to adapt infrastructure to the special characteristics of the AV technology. In particular, a unique characteristic of AVs is their ability to drop off commuters at work and park outside downtown. This may translate into a need for more drop-off spots and a lower demand for parking downtown. Since increasing the drop-off capacity is possible through converting regular parking spaces in D to drop-off stations, there is a trade-off between the curbside drop-off capacity, R_W, and the number of parking spots downtown, K_1. On one hand, increasing R_W reduces the total work schedule penalty. This is because curbside space is the most downstream bottleneck that determines the departure pattern of commuters from H, so more curbside space allows some commuters to leave their homes later than before. On the other hand, reducing K_1 increases the number of trips to Area 2 as well as the amount of passengerless travel time.

The following corollary derives the optimal drop-off capacity, denoted by R_W^*, and the optimal capacity of Area 1, denoted by K_1^*, that minimize the total system cost of the morning commute under SO.[11]

Corollary 2.1. *There exist* $R_W^* \in [0, \min\{K_1 + R_W, R_I\}]$ *and* $K_1^* = K_1 + R_W - R_W^*$*, where* R_W *and* K_1 *represent the current values of drop-off capacity and Area 1's capacity, respectively.*

Corollary 2.1 shows that the optimal value of curbside space, R_W^*, can range from zero to either the inbound bottleneck capacity R_I or the total available space downtown, which is the sum

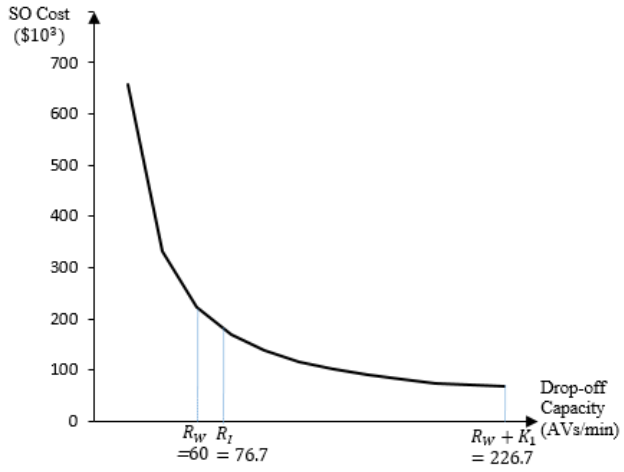
[10] For case (ii) of SO2, the pricing and tolling schemes are similar to those of case (i), except that after Area 1 becomes full at $y_{1,max}$, the congestion toll $\pi_2(x)$ is reduced by $\alpha'\tau_{D,2}(x)$ to account for the growing downtown congestion cost.

[11] We assume there is no cost associated with converting a parking space to a drop-off space. However, such a cost can easily be incorporated in our model, as it is linear to the number of converted parking spots.

of current curbside drop-off space R_W and parking space, K_1. The inbound bottleneck capacity limits the improvement in the drop-off capacity because increasing the drop-off capacity beyond the inbound bottleneck capacity does not reduce cost, as commuters will experience congestion in the inbound bottleneck. To effectively improve the total system cost, the social planner should increase (or decrease) the inbound bottleneck capacity simultaneously with the drop-off capacity, because otherwise, one of them becomes the bottleneck and the other one has some underutilized capacity. The optimal capacity of Area 1, K_1^*, is then equal to the remaining space after allocating the curbside drop-off, i.e., $R_W + K_1 - R_W^*$. This implies that, depending on the current values of R_W and K_1, adding more parking spots downtown can lead to a higher or a lower total system cost. Hence, AVs do not necessarily reduce the number of downtown parking spots. In fact, when downtown parking leads to a lower passengerless travel time than parking in the external area, especially when downtown congestion is high, the social planner prefers AVs to park downtown, although it extends the duration of the morning commute and increases the total work schedule penalty for all commuters.

Figure 2.8 illustrates the total system cost as a function of the drop-off capacity for our calibrated model under SO1. In this case, since the unit work schedule penalty is higher than the unit passengerless travel cost ($\beta = \$3.9 > \$2.25 = \alpha'$ per hour), even though Area 1 is closer than Area 2 ($t_{W,1} < t_{D,2}$), the positive effect of increasing R_W on the work schedule penalty dominates the negative effect of decreasing K_1 on the passengerless travel time. Hence, as this figure shows, the social planner should increase the drop-off capacity as much as possible, because the higher drop-off capacity the lower the total system cost. In other words, the drop-off capacity should increase from the current value of $R_W = 60$ commuters per minute to $R_W^* = 226.7$ commuters per minute (or $13,600 = 3,600 + 10,000 = R_W + K_1$ commuters per hour), and AVs eliminate the downtown parking demand. This can be achieved by converting all downtown parking spots to drop-off stations (i.e., reducing $K_1 = 10,000$ parking spots to $K_1^* = 0$). However, the drop-off capacity is bounded by the inbound bottleneck capacity, so if we want to increase the drop-off capacity beyond the current value of R_I (76.7 commuters per minute), we need to simultaneously increase the inbound bottleneck capacity R_I to the same level, i.e., 226.7 commuters per minute. This reduces the total system cost under SO by 69.3%. Even increasing the drop-off capacity R_W from its current value of 60 commuters per minute to the inbound bottleneck capacity R_I of 76.7

Figure 2.8: An illustration of total system cost as a function of drop-off capacity R_W for SO1.



commuters per minute results in a 20.9% reduction in the total system cost.[12]

2.7 Conclusion

In this paper we investigate the effect of AVs on the morning commute and infrastructure (drop-off and parking spaces) usage. We characterize the departure time and parking location patterns of commuters under UE based on the downtown parking fee, the location of the external parking area, and the availability of downtown parking. Our model also takes into account the capacity of roadways and parking spaces and the monetary value of passengerless travel time. As AV technology advances and becomes more reliable, the operational costs of AVs decrease and the monetary value of passengerless travel time decreases, which can lead to an even more pronounced impact on traffic patterns and parking usages. In the case when commuters prefer parking downtown due to the high travel cost from downtown to Area 2 (i.e., UE1), the AV technological advancement can reduce the cost of passengerless travel and induce commuters to mainly park in the external area (i.e., UE2). Yet, a setback in the AV technology can change commuters' decisions from choosing the external parking area to the downtown parking area, as the distrust in the AV technology may increase the cost of passengerless travel time for commuters due to the potential extra supervision required.

We compare these patterns under UE against those determined by a social planner who aims to

[12]The numerical example of SO2 discussed in §2.5.2 shows a similar pattern of the total system cost to Figure 2.8, and R_W^* is equal to the minimum of $R_W + K_1$ and R_I.

minimize the total system cost. For cities with small downtown areas, regardless of the downtown parking fee, since the social planner aims to minimize the total travel time, she wants the downtown parking area to be filled before any commuter goes to the external area. For cities with bigger downtown areas, the social planner wants to alternately route commuters to the external and downtown parking areas to avoid creating a significant amount of congestion downtown. Since the commuters take the downtown parking fee into account when deciding on their parking location, the socially optimum (SO) differs from what commuters decide under UE, unless new policies are developed to regulate commuters' decisions. For example, for our calibrated model, under UE all commuters choose to park outside downtown, but SO is to choose the downtown parking area (as long as it is not full) to reduce the total system cost. In addition, in comparison to HV commuters, AV commuters tend to leave early to avoid congestion downtown under UE. This exacerbates congestion on the inbound highway, as well as the drop-off and downtown congestions, causing a significantly higher total system cost.

To reduce the high congestion cost created by AVs during the morning commute, we recommend two complementary approaches to social planners. First, we characterize optimal parking fees and congestion tolls that reduce the total system cost as a short-term solution. For cities with small downtown areas, where the downtown parking area is closer to work than the external one, the socially optimal decision is to route AVs to the downtown parking area. But, depending on the downtown parking fee, either this area or the external area can be the primary parking area chosen by commuters. We recommend optimal parking fees and congestion tolls that provide sufficient incentives for commuters to choose the downtown parking area as long as it has capacity under equilibrium. For cities with larger downtown areas, where the downtown parking area is farther away, the social planner as well as commuters want the external area to be the primary parking area. However, since AVs that go to the external area may create substantial downtown congestion, the cost associated with traveling to the external area can surpass that of the downtown area. So we recommend that the social planner simultaneously uses parking fees and congestion tolls as levers to either direct AVs to the downtown parking area when downtown congestion is high or extend the duration of morning commute to eliminate downtown congestion.

We also derive the optimal curbside drop-off capacity and the downtown parking capacity that minimize the total system cost as a long-term solution. The social planner prefers to keep the

departure rate from home under the drop-off capacity to eliminate the passengered congestion time. This shows that the social planner may shift the focus from expanding parking spaces downtown to creating more designated curbside drop-off spaces. In line with this observation, we recommend infrastructure improvements to the social planner. Increasing the drop-off capacity is possible through reducing the number of downtown parking spots. For cities with small downtown areas, this reduces the capacity of the closer parking area, which can lead to increasing the SO cost. However, increasing the drop-off capacity also means that more commuters can depart home per hour, so the morning commute duration shrinks and the aggregate work schedule penalty decreases. We recommend optimal values of drop-off capacity and downtown parking spots that result in the lowest total system cost in the long run. Depending on the city characteristics, the optimal decision can be anything from converting all downtown parking spots to drop-off spots or reducing the drop-off capacity from its current value. So, surprisingly, there is a chance that the social planner needs to expand the number of downtown parking spots. Even in the case when increasing the number of downtown parking spots is recommended, the social planner should take the inbound highway capacity into account. Increasing the drop-off capacity beyond the inbound highway capacity does not further reduce the total system cost, as the inbound highway becomes the new bottleneck for commuters. Thus, we recommend that the social planner increases (or reduces) the inbound bottleneck capacity simultaneously with the drop-off capacity. For our model calibrated to Pittsburgh, we show that converting all downtown parking spots to curbside drop-off spots is the optimal choice, as it facilitates the drop-off process.

Our paper is the first to characterize the impact of AVs on the morning commute problem by modeling an external parking area as well as downtown congestion caused by AVs. There are many plausible future scenarios where AVs may be used during the morning commute, such as the use of shared autonomous transportation services or the option of cruising around instead of parking. We hope that our model and analysis provide a foundation for other investigations on this topic.

Chapter 3

Modeling and Managing Curbside Ride-Hailing Drop-offs

3.1 Introduction

With the proliferation of mobile devices and advancements in accurate positioning technologies, ride-hailing services, provided by *transportation network companies* (TNCs), are an indispensable component in mobility systems. These convenient and low-cost ride-hailing services provide morning commuters with an attractive travel alternative. During morning commute hours, managing parking and traffic congestion is one of the main issues that a city planner deals with. To accommodate high parking demands, a city has to surrender large spaces to build parking structures; for example, an astounding 14 percent of Los Angeles county land is dedicated to parking. Ride-sharing vehicles might be able to alleviate this problem, as they have the ability to drop commuters off at their workplaces in a city center without parking. This allows commuters to avoid high parking fees in a city center, and reduces a need to build or maintain large parking structures in a city center. However, ride-hailing services contribute to traffic congestion by exploiting more curbside space than driving and public transit require (Castiglione et al. 2016, Li et al. 2016, Castiglione et al. 2018, Agarwal et al. 2019): Ride-hailing pickups and drop-offs may disturb or block the traffic flow of conventional vehicles (Goodchild et al. 2019). This negative externality of ride-hailing services could push some commuters to change their mode of transportation from driving their personal

vehicles to ride-hailing, creating even more curbside congestion.

In this chapter, we study the effect of ride-hailing drop-offs on the morning commute and provide guidance to city planners on how to adjust infrastructure. In particular, we investigate the potential effect of ride-hailing on the morning commute patterns that commuters face when traveling to a central business district for work by tackling the following fundamental questions: (1) How do ride-hailing services affect congestion and downtown parking demand? (2) How should municipal governments design infrastructure (curbside drop-off zones) to minimize aggregate travel costs of commuters?

To answer these questions, we develop a continuous-time traffic model in which commuters decide when to leave their residences and what transportation mode they choose from one of the two available options: driving and ride-hailing. The goal of a commuter is to minimize his total transportation costs (including parking fees, ride-hailing fares, and imputed costs of early arrivals) during the morning commute. (For convenience, we use the pronouns "he/him/his" to refer to a commuter.) Three major elements affect an individual commuter's costs: travel time, arrival time at work, and parking fee/ride-hailing fare. The individual commuters' decisions on their departure times from home collectively affect the level of congestion on roadways; hence, these decisions affect both travel time and arrival time of an individual commuter. To avoid congestion, a commuter may decide to depart home early, but the commuter may not want to depart too early because there exists an inconvenience cost for arriving too early at work (Hendricks 2015). As such, commuters encounter a trade-off between experiencing congestion and arriving too early to work. In addition, there is a trade-off in choosing between the two transportation modes (i.e., driving and ride-hailing): the ride-haling passengers experience more congestion at drop-off than personal vehicle passengers do, but ride-haling passengers do not need to walk from parking area to work. These trade-offs faced by commuters are captured in a continuous-time game-theoretic traffic model. Our model is motivated by infrastructure and traffic patterns in the City of Pittsburgh. In this model, commuters decide on their departure times from home and their modes of transportation between driving and ride-hailing. These commuters take an inbound highway (e.g., I-376 in Pittsburgh) to travel to work downtown. Upon arrival downtown, the ride-hailing commuters are dropped off as soon as there is an available curbside drop-off spot, and the rest of the commuters drive to downtown parking areas to park their vehicles. The latter group of commuters walks to work from the parking areas.

Our model takes into account the congestion that commuters experience on the inbound highway to downtown, the extra congestion that ride-hailing commuters experience before drop-off, and the extra congestion that commuters who drive experience on their way to the parking areas. We calibrate our model to traffic and parking data from Pittsburgh.

By analyzing this model, we characterize the departure time and transportation mode decisions of commuters as well as ride-hailing fare pricing decisions of the TNC under equilibrium. Under equilibrium, no commuter can unilaterally change his departure time and/or parking location to reduce his cost nor can the TNC change its fares to increase its profit. We show that there exist two different cases of equilibrium, depending on when the TNC starts serving commuters during the morning commute. In determining when to offer rides, the TNC faces a trade-off between the total number of commuters it serves and the fare it charges. On the one hand, if the TNC offers rides early during the morning commute period, the fares should be low to be competitive with the low downtown parking fees. On the other hand, due to the low fares, more commuters choose ride-hailing. For cities where the downtown parking fees are relatively low, e.g., there is an early bird discount, it is not profitable for the TNC to offer rides, so it sets its fares so high that no commuter chooses ride-hailing. For cities with relatively high early morning downtown parking fees, the TNC sets its fare just below the cost of driving such that it deters commuters from driving. These two cases coincide when the downtown parking fees are moderate, such as in Pittsburgh, so both driving and ride-hailing are chosen by some commuters throughout the entire duration of the morning commute. Under all equilibrium cases, the departure rate of ride-hailing commuters from home is higher than the capacity of the drop-off zone, causing drop-off congestion for these commuters as well as disturbing the flow of commuters who drive through the drop-off zone to go to the parking areas. This makes driving a less desirable option than when there is no ride-hailing services, and may push more commuters who currently drive to switch to ride-hailing, leading to even more congestion than in the case with no ride-hailing vehicles.

We next analyze the problem of a social planner who aims to minimize the total system cost of the morning commute by dictating the departure time and transportation modes of all commuters as well as the ride-hailing fares of the TNC. Our analysis shows that there exist two different social optima (SOs). For cities such as Pittsburgh, where the walking time from the parking areas to work is relatively short, the social planner directs some commuters to drive and some to use ride-hailing.

Otherwise, she wants all commuters to use ride-hailing. (For convenience, we use the pronouns "she/her/hers" to refer to the social planner.) The SO analysis allows us to identify that the social optimum that minimizes the aggregate costs of all commuters may differ from the decisions of the commuters and the TNC under equilibrium. Thus the total system cost under equilibrium is higher than the SO cost.

As a way to reduce the gap between the individual and social optima, we examine both short-term and long-term measures a social planner can implement. As for short-term measures, we analyze a social planner's decisions on parking fees as well as curbside drop-off tolls. These measures can reduce the total system cost by inducing commuters to choose the departure times and transportation modes desired by the social planner. Our numerical analysis of the Pittsburgh data indicates that these measures can reduce the total system cost of the morning commute to the SO cost (a 77% reduction). In the long run, a social planner can further lower the total system cost by adapting the infrastructure so it better suits ride-hailing operations. For example, it may be beneficial to increase the number of curbside drop-off spaces downtown. For Pittsburgh, we find that by increasing the drop-off capacity in Downtown Pittsburgh to the inbound highway capacity, we can achieve an 8% reduction in the total system cost. These results suggest that cities can benefit significantly from ride-haling services by adjusting their short-term and long-term transportation and infrastructure policies.

The rest of this chapter is organized as follows. In §3.2, we review the related literature. Our morning commute travel model is presented in §3.3. In §3.4 and §3.5, we analyze the equilibrium and the SO, respectively. In §3.6, we propose both short-term and long-term plans to reduce the total system cost of the morning commute. We conclude in §3.7.

3.2 Related Literature

This work is closely related to the literature on smart city operations and the literature on transportation science. Under the umbrella of the smart city operations literature, there are various studies that focus on different aspects of ride-hailing, including: surge pricing (e.g., Besbes et al. (2021), Garg & Nazerzadeh (2021), Guda & Subramanian (2019), and Hu et al. (2022)), matching

and dispatching (e.g., He et al. (2020), and Özkan & Ward (2020)), supply and demand management (e.g., Lian & Van Ryzin (2021), and Yu et al. (2020)), and transit planning (e.g., Wei et al. (2021)). For a comprehensive review of ride-hailing studies, we refer readers to Benjaafar & Hu (2020) and Chen et al. (2018). We expand this evolving stream of research by examining the effect of ride-hailing on congestion and parking during the morning commute.

The early research in the transportation science literature that studies the morning commute problem focuses on driving as the sole mode of transportation. Morning commuters choose their departure times from home based on multiple factors such as congestion, schedule delays, parking fees and availability. Vickrey (1969) considers a finite group of commuters who decide on their departure time from home to their work places downtown. He shows that there exists an equilibrium departure time pattern when all commuters attempt to minimize their own travel costs. Arnott et al. (1991) extend Vickrey (1969) by examining commuters' decisions on both departure times from home and parking locations. They consider a combination of congestion tolls and parking fees to minimize the total system cost, and show that the optimal departure rate from home must be equal to the capacity of the inbound highway. Xu et al. (2019) study different flat congestion toll schemes that are easier to implement to reduce the total system cost. The model of Arnott et al. (1991) is further extended by accounting for other features such as joint morning and evening commute (e.g., Zhang et al. (2008)), multiple parking clusters downtown (e.g., Qian et al. (2011)), multiple residential areas (e.g., He et al. (2015)), and positive search time to find an empty parking spot (e.g., Qian & Rajagopal (2014) and Qian & Rajagopal (2015)). Our model also builds on the fundamental structure of Arnott et al. (1991) while enriching it by capturing the specific characteristics of ride-hailing (including commuter drop-off at work and extra congestion) as a mode of transportation.

Ride-hailing is incorporated in the morning commute problem by Su & Wang (2019). Similar to our paper, they consider the problem of morning commute when commuters have two options for commuting from home to work: driving and ride-hailing. In their model, commuters who drive pay a fixed parking fee to park their vehicles at their work place (i.e., the walking time from the parking areas to work is assumed to be negligible), and commuters who use ride-hailing pay a fixed fare and are dropped off at work without experiencing or causing congestion. Su & Wang (2019) show that under equilibrium ride-hailing might be used at the late stage of the morning peak only, which

leads to an extended morning rush hour duration and a high total system cost. They recommend that, to achieve the minimum total system cost, a social planner should restrict the number of downtown parking spots as well as ride-hailing trips, and derive these optimal values, however, no policies to enforce such restrictions is discussed.

We contribute to this literature by examining the effect of ride-hailing on congestion and parking during the morning commute. Modeling ride-hailing is essential because it helps reduce downtown parking demands. We show that having ride-hailing as a transportation mode option results in a lower total system cost. This implies that ignoring this option may result in an overestimation of the cost. However, ride-hailing has some potential negative impacts, so the aggregate effect of ride-hailing as a commuting option is not clear without a careful analysis. To properly model this mode of transportation, different from the prior literature reviewed above, we consider the congestion experienced by ride-hailing commuters during drop-offs, as well as the negative externality of ride-hailing drop-offs on commuters who drive. Our model captures the fact that the curbside space dedicated to commuter drop-offs is limited, potentially causing delays during drop-offs for ride-hailing commuters. Such delays are illustrated in a simulation study by Overtoom et al. (2020). The ride-hailing drop-offs also affect commuters who drive, as ride-hailing vehicles leave and rejoin the traffic stream frequently. This can disturb the traffic flow and induce extra delays for commuters who pass through the drop-off zone to go to the parking areas. Although the impact of ride-hailing drop-offs depends on road properties and traffic conditions, empirical evidence suggests that it is generally negative and significant (Goodchild et al. 2019). In addition, our model considers dynamic parking fees and ride-hailing fares, as both of these charges might change dynamically based on the time of the day and the level of congestion. Finally, we offer insight into both short-term and long-term measures to reduce the total system cost a social planner can implement in anticipation of mass usage of ride-hailing by commuters. In short, our paper offers a unique perspective into the future of smart cities by investigating the impact of ride-hailing on parking and congestion downtown.

3.3 Model

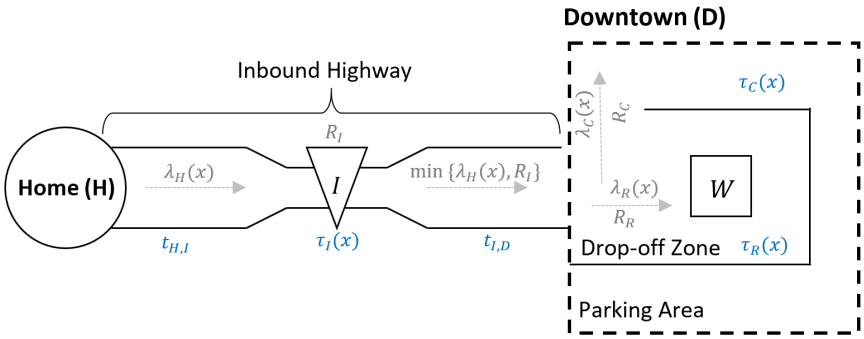
We study the problem of N morning commuters who travel from home (H) to work (W) that is located downtown (D). There are two modes of transportation available to these commuters: driving and ride-hailing. In §3.3.1 we describe the commuters' model, and in §3.3.2 we discuss the model for the transportation network company (TNC). Appendix A provides a table for the summary of our notation.

3.3.1 The Commuters

We describe the model for individual commuters who travel from home to work during morning rush hour. We call the commuters who drive their personal vehicles conventional (C) commuters and those who use ride-hailing services ride-hailing (R) commuters. Figure 3.1(a) depicts the travel route of conventional and ride-hailing commuters. As shown in this figure, to go from H to D, all commuters travel on a highway which has an inbound bottleneck (I) with the capacity of R_I vehicles per hour (where $0 < R_I < \infty$). This means that if more than R_I commuters arrive at this bottleneck per hour, these commuters experience a delay on their way to D. Upon arriving at D, the two types of commuters, i.e., C and R, follow different paths. The ride-hailing commuters are dropped off in the drop-off zone surrounding W, and walk a short distance from the drop-off zone to W. Due to limited curbside space, at most R_R (> 0) drop-offs per hour are possible. If more than R_R ride-hailing vehicles per hour arrive at W to drop off their commuters, these commuters experience a drop-off delay. The conventional commuters travel through the drop-off zone to go to the downtown parking areas, which consists of all downtown parking spots surrounding W. The capacity of downtown roads and streets is limited, hence the number of conventional commuters that can travel from the drop-off zone to the parking area per hour, denoted as R_C, is limited, potentially causing delay for these commuters. In addition, the ride-hailing commuters awaiting drop-off might add to the congestion that the conventional commuters who are traveling to the downtown parking areas experience. The price of downtown parking is denoted as p.

Every morning a commuter makes two decisions: departure time x ($\in [0, x_{max}]$) from H (where the latest departure time x_{max} is determined endogenously), and mode of transportation j ($\in \{C, R\}$). The departure rate of commuters from H at x is denoted as $\lambda_H(x)$. For $j \in \{C, R\}$, we

Figure 3.1: An illustration of the morning commute route for conventional and ride-hailing commuters.



let $\lambda_j(x)$ represent the rate of commuters who choose transportation mode j and leave their homes at x. As such, the total departure rate from H, $\lambda_H(x)$, is equal to $\lambda_C(x) + \lambda_R(x)$ for $x \in [0, x_{max}]$. Following prior literature (e.g., Arnott et al. (1991), Qian et al. (2011), Qian & Rajagopal (2015), and Liu (2018)), the official start time at work is denoted by T for all commuters. According to the U.S. Bureau of Labor Statistics (2015), most workers in the U.S. are on the job between 8 a.m. and 5 p.m, so it is justified to assume that the majority of companies start their workday at a specific time (e.g., 8 am).

As illustrated in Figure 3.1, travel time of commuters who leave H at x is divided into: (1) free-flow travel time, (2) delay in the inbound bottleneck, (3) delay caused by drop-off congestion at W for ride-hailing commuters, (4) delay caused by congestion within D for conventional commuters, and (5) walking time. We elaborate on each of these travel times.

(1) Free-flow travel time: This is the travel time of vehicles when they are able to move freely on segments of highways and roads without a capacity constraint. Specifically, the free-flow travel time occurs on the following two segments: between home H and inbound bottleneck I, and between inbound bottleneck I and the boundary of downtown D. Free-flow travel time in each of these segments, which has a fixed duration, is denoted respectively by $t_{H,I}$, and $t_{I,D}$. Since all commuters experience $t_{H,I}$ and $t_{I,D}$, without loss of generality, we normalize them to zero.

(2) Inbound queueing delay: If the rate of commuters who arrive to the inbound bottleneck exceeds the capacity of this bottleneck, R_I, these commuters experience a delay, which is called the inbound queueing delay, $\tau_I(x)$.

(3) Drop-off congestion time for ride-hailing commuters: Before being dropped off at W,

the ride-hailing commuters may experience delays due to the limited curbside space at W. We call this delay the drop-off congestion time for ride-hailing commuters and denote it by $\tau_R(x)$.

(4) Downtown congestion time for conventional commuters: Before finding an empty parking spot downtown, the conventional commuters may experience delays due to the limited downtown streets capacity and the externality of the ride-hailing drop-offs. We call this delay the downtown congestion time for conventional commuters and denote it by $\tau_C(x)$.

(5) Walking time: Both types of commuters walk to W: The conventional commuters walk from the parking area to W after parking their vehicles. We denote this walking time, which is a fixed value, by t_C. The ride-hailing vehicles also walk from the drop-off zone to W after being dropped off. This walking time is assumed to have a fixed value and is denoted as t_R ($< t_C$).

Taken together, the total travel time of conventional commuters is $\tau_I(x) + \tau_C(x) + t_C$, and that of ride-hailing commuters is $\tau_I(x) + \tau_R(x) + t_R$.

We make the following assumptions throughout the paper:

A1. The downtown parking fee, $p(x)$, is an increasing function of x ($\in [0, x_{max}]$), because the commuters choose the cheapest parking spots available, so the parking fee goes up as a commuter's departure time from H increases.

A2. The capacity of the inbound highway is higher than the capacity of downtown roads and the curbside drop-off capacity, i.e., $R_I > R_C$ and $R_I > R_R$. If any of these inequalities does not hold, the congestion time associated with that inequality disappears: when the inbound bottleneck capacity R_I is lower than the downtown roads capacity R_C (resp., the drop-off rate R_R), there is no downtown congestion for conventional commuters (resp., drop-off congestion for ride-hailing commuters). Our analysis can be easily extended to those simpler cases with no congestion.

A3. We model each segment on the way from H to the parking areas (inbound bottleneck, drop-off area, and downtown) as a queue with deterministic time-varying arrival and service rates. For example, the inbound bottleneck is a queue with arrival rate $\lambda_H(x)$ for $x \in [0, x_{max}]$ and service rate R_I vehicles per hour. As discussed in Kim & Whitt (2013), calculating the exact wait time (i.e., congestion time in a segment) for each individual commuter that arrives to that segment is complex. Thus, we estimate the individual wait time of a commuter that

leaves H at x as the number of commuters that are present in the queue divided by the average throughput of the queue. We divide by the average throughput because commuters that are present in the queue, but have left H before x, might leave the queue at different rates depending on their arrival time to the queue. Specifically, the congestion time $\tau_I(x)$ that a commuter traversing the I bottleneck at x experiences is equal to $\frac{\int_0^x [\lambda_H(u)-R_I]^+ du}{(\int_0^x \min\{\lambda_H(u),R_I\}du)/x}$. The numerator of this expression is the total number of commuters present in the bottleneck at x, and the denominator is the average departure rate from I between time zero and x. Similarly, we calculate the drop-off congestion time for ride-hailing commuters as $\tau_R(x) = \frac{\int_0^{x+\tau_I(x)} [\lambda_R(u)R_I/\lambda_H(u)-R_R]^+ du}{\int_0^{x+\tau_I(x)} \min\{\lambda_R(u)R_I/\lambda_H(u),R_R\}du/[x+\tau_I(x)]}$.[1]

A4. The drop-off process interrupts the flow of conventional vehicles when the number of ride-hailing vehicles waiting to drop-off their commuters, i.e., $\int_0^{x+\tau_I(x)} [\lambda_R(u) - R_R]^+ du$, exceeds $\Gamma(\geq 0)$. By A3, when $\int_0^{x+\tau_I(x)} [\lambda_R(u) - R_R]^+ du \geq \Gamma$, we state the downtown congestion for conventional commuters as $\tau_C(x) = \frac{\int_0^{x+\tau_I(x)} [\lambda_C(u)R_I/\lambda_H(u)-R_C]^+ du + \int_0^{x+\tau_I(x)} \delta[\lambda_R(u)R_I/\lambda_H(u)-R_R]^+ du}{\int_0^{x+\tau_I(x)} \min\{\lambda_C(u)R_I/\lambda_H(u),R_C\}du/[x+\tau_I(x)]}$, and as $\tau_C(x) = \frac{\int_0^{x+\tau_I(x)} [\lambda_C(u)R_I/\lambda_H(u)-R_C]^+ du}{\int_0^{x+\tau_I(x)} \min\{\lambda_C(u)R_I/\lambda_H(u),R_C\}du/[x+\tau_I(x)]}$, otherwise. The case of $\Gamma = 0$ leads to a simplified model, as in this case any drop-off spillover into travel lanes causes negative externality for conventional vehicles.[2]

Now that we have modeled the travel time of commuters, we next consider the costs associated with their commutes. The cost that a commuter incurs consists of three elements: travel time cost, work schedule penalty, and parking cost or ride-hailing fare. First, to define the travel time cost, let α and α' (where $0 < \alpha < \alpha'$) represent the monetary value of one unit of travel time in a vehicle, which includes the vehicle usage costs (e.g., gas/electricity, depreciation, mileage, etc.) and the value of commuters' time, and the monetary value of one unit of walking time, respectively. The travel time cost is then equal to $\alpha[\tau_I(x) + \tau_C(x)] + \alpha' t_C$ and $\alpha[\tau_I(x) + \tau_R(x)] + \alpha' t_R(x)$ for conventional and ride-hailing commuters, respectively.

[1] For the drop-off congestion time $\tau_R(x)$, the numerator is the total number of ride-hailing commuters that are present at the drop-off zone at $x + \tau_I(x)$ (which is the time when commuters who leave H at x arrive at the drop-off zone in W), and the denominator of $\tau_R(x)$ is the average drop-off rate at $x + \tau_I(x)$.

[2] For the downtown congestion time $\tau_C(x)$, the numerator can have two components. The first component, i.e., $\int_0^{x+\tau_I(x)} [\lambda_C(u)R_I/\lambda_H(u) - R_C]^+ du$, is the total number of conventional commuters that are present in the downtown area and head to the parking area at $x + \tau_I(x)$ (which is the time when these commuters that leave H at x arrive at W). The second component, i.e., $\int_0^{x+\tau_I(x)} \delta[\lambda_R(u)R_I/\lambda_H(u) - R_R]^+ du$, where $0 \leq \delta \leq 1$, captures the negative externality of the ride-hailing commuters that are present in the drop-off area on the travel time of the conventional vehicles. Lastly, the denominator of $\tau_C(x)$ is the average travel rate of conventional commuters in D at $x + \tau_I(x)$.

The second cost item is a work schedule penalty, which is a penalty that a commuter incurs when he arrives before time T. Following the prior literature (e.g., Arnott et al. (1991), Qian et al. (2011), and Liu (2018)), we define a work schedule penalty for the type $j \in \{C, R\}$ commuter as the difference between the actual arrival time, $x + \tau_I(x) + \tau_j(x) + t_j$, and the official start time at work, T. Let β represent the monetary cost of early start of work. Then, the work schedule penalty for commuters who arrive at W early is equal to $\beta\{T - [x + \tau_I(x) + \tau_j(x) + t_j]\}$. Finally, a conventional commuter who leaves H at x pays parking fee $p(x)$, while a ride-hailing commuter who leaves H at x pays a ride-hailing fare $u(x)$.

Before defining the total cost of commuters, we state our final assumptions:

A5. The penalty of starting work early is lower than the monetary value of passengered travel time, i.e., $0 < \beta < \alpha$. This is the standard assumption in the literature, and it is empirically supported (e.g., Small (1982)). In addition, similar to Liu (2018), the marginal increase in the parking fee is lower than the marginal decrease in the work schedule penalty, i.e., $p'(x) \leq \beta\ (< \alpha)$ for $x \in [0, x_{max}]$. This means that a commuter prefers to arrive to work closer to time T, because if he delays his arrival time at W by one unit of time, his marginal saving in the work schedule penalty (β) is higher than his marginal increase in the cost of parking downtown ($p'(x)$).

A6. No commuter intentionally decides to arrive late at work. In other words, the monetary value of arriving late at work is so high that all commuters arrive at W before the work start time T.

Putting the three cost elements together, we can express the total cost of a conventional commuter and a ride-hailing commuter who depart H at x, respectively, as follows:

$$C_C(x) = \alpha[\tau_I(x) + \tau_C(x)] + \alpha' t_C + \beta\{T - [\tau_I(x) + \tau_C(x) + t_C]\} + p(x), \tag{3.1}$$

$$C_R(x) = \alpha[\tau_I(x) + \tau_R(x)] + \alpha' t_R + \beta\{T - [\tau_I(x) + \tau_R(x) + t_R]\} + u(x). \tag{3.2}$$

Lastly, we define the total system cost (also known as the social cost) as follows.

$$\int_0^{x_{max}} \lambda_C(x)[C_C(x) - p(x)] + \lambda_R(x) C_R(x) dx = \tag{3.3}$$
$$\int_0^{x_{max}} \lambda_H(x)[(\alpha - \beta)\tau_I(x) + \beta(T - x)] + \lambda_C(x)[(\alpha - \beta)\tau_C(x) + (\alpha' - \beta)t_C] + \lambda_R(x)[(\alpha - \beta)\tau_R(x)$$
$$+ [(\alpha' - \beta)t_R + u(x)]dx.$$

The cost in (3.3) consists of three terms: The first term ($\lambda_H(x)[(\alpha-\beta)\tau_I(x)+\beta T]$) is the travel cost that all commuters incur, regardless of their mode of transportation. The second term ($\lambda_C(x)[(\alpha-\beta)\tau_C(x)+(\alpha'-\beta)t_C]$) is the sum of downtown congestion cost and walking time for conventional commuters. The last term ($\lambda_R(x)[(\alpha-\beta)\tau_R(x)+[(\alpha'-\beta)t_R+u(x)]$) is the sum of ride-hailing fare, drop-off congestion cost, and walking time for ride-hailing commuters. The total system cost does not include parking fees $p(x)$, because parking fees paid by the commuters who park downtown cancel out parking fees collected by the social planner. In other words, parking revenues collected by the social planner are considered as a part of social welfare, so they are not counted towards the total system cost.

In our subsequent analyses, we illustrate our analytical results using the parameter values estimated from the Pittsburgh Metropolitan Area. We summarize our calibrated parameter values in Table 1 while referring the readers to Pi et al. (2019) and Mirzaeian et al. (2021).

Table 3.1: Summary of the calibrated model parameters

Parameter	Value	Parameter	Value	Parameter	Value
N	$20,000$ commuters	t_C	15 minutes	$p(0)$	\$13
R_I	4,600 vehicles per hour	t_R	2 minutes	$p(x_{max})$	\$20
R_C	$4,600$ vehicles per hour	α	\$9 per hour	$p(x)$	$7x/x_{max}+13$ dollars
R_R	$3,600$ drop-offs per hour	α'	\$12 per hour	δ	0.8
K	$10,000$ vehicles	β	\$3.90 per hour	Γ	$(R_R+R_C)/60$

Notes. By A1, we calculate $p(x)$ as $7x/x_{max}+13$, because, according to Pittsburgh Parking Authority (2019), the minimum parking rate in downtown Pittsburgh is \$13 and the maximum rate is \$20.

3.3.2 The TNC

There is a TNC that offers ride-hailing services to the morning commuters. The goal of the TNC is two-fold. On the one hand, the TNC wants to maximize its profit by determining the fares $u(x)$ for $x \in [0, x_{max}]$. In fact, $u(x)$ represents the profit margin of the TNC, as we assume that the TNC directly charges the commuters for the travel cost associated with each trip (i.e., the inbound congestion and drop-off congestion costs). Since the ride-hailing demand is elastic, i.e., $\lambda_R(x)$ depends on $u(x)$, the TNC faces a trade-off between charging a higher fare and attracting more commuters. On the other hand, the TNC wants to minimize the total travel time of ride-hailing trips, as ride-hailing drivers also experience congestion, which might deter them from joining the

ride-hailing platform. We state the objective of the TNC service as follows:

$$\max_{0\leq u(x)} \int_0^{x_{max}} \lambda_R(x)u(x) - \alpha[\tau_I(x) + \tau_R(x)]dx. \tag{3.4}$$

3.4 Equilibrium Analysis

In this section, we present a joint equilibrium for the commuters and the TNC. In this equilibrium, individual commuters decide on their departure times and transportation modes to minimize their costs, and the TNC decides on the ride-hailing fares to maximize its profit. We define an equilibrium as when the following three conditions are satisfied:

Condition 1: $C_C(x) = C_R(x)$ for any $x \in [0, x_{max}]$ such that $\lambda_C(x) \neq 0$ and $\lambda_R(x) \neq 0$.

Condition 2: For each $j \in \{C, R\}$, $\frac{\partial C_j(x)}{\partial x} = 0$ for any $x \in [0, x_{max}]$ such that $\lambda_j(x) \neq 0$.

Condition 3: $u(x) = u^*(x)$ for any $x \in [0, x_{max}]$ such that $u^*(x)$ satisfies (3.4).

Conditions 1 and 2 guarantee that all commuters incur the same total cost, regardless of their choices of transportation mode and departure time, respectively. This means that, under equilibrium, no commuter can unilaterally change his departure time and/or transportation mode to reduce his cost. Condition 3 guarantees that the TNC's profit is maximized. This means that, given the ride-hailing demand under equilibrium, the TNC cannot change its fares to achieve a higher profit.

We next characterize the equilibrium solutions. Proposition 3.1 indicates that there are two possible forms of equilibrium. We use the following additional notation to describe these equilibria: x_j for $j \in \{C, R\}$ denotes the earliest departure time of commuters who choose transportation mode j. Proofs are provided in Appendix C.2.

Proposition 3.1. *There exists a unique equilibrium which is presented in Table 3.2, where* $x_C = \Big[\delta R_R\{(\alpha R_C - \beta\delta R_R)[R_R(\alpha-\beta)+R_C\beta/\delta]u_0 - [R_C[p(x_{max})-p(0)]+N(\alpha-\beta)](R_C-\delta R_R)\alpha\beta/\delta\}/\{\beta(R_C-\delta R_R)[(\alpha-\beta)R_R + R_C\beta/\delta][\alpha R_C - (\alpha+\delta\beta)R_R]\}\Big]^+$, $x_R = \Big[\{R_C[p(x_{max}) - p(0)] + N(\alpha-\beta) - \delta(\alpha-\beta)\Gamma\}\{\alpha(\alpha-\beta)R_R - \beta[\alpha R_C + (\alpha-\delta\beta)R_R] - (\alpha-\beta)\alpha R_C[(\alpha'-\beta)(t_C-t_R)+p(0)]/\beta\}/[(\alpha-2\beta)(\alpha R_C)^2]\Big]^+$, $\bar{x}_R = \frac{\Gamma(\alpha-\beta)}{\alpha R_R} + x_R$ *for case (i) and* $\bar{x}_R = \frac{\Gamma(\alpha-\beta)-\beta(R_C/\delta-R_R)x_C}{\alpha R_R}$ *for case (ii),* $A = \frac{\alpha-p'(x)}{\alpha-\beta}R_C$, $B = \frac{\alpha-p'(x)}{\alpha-\beta}R_C - \frac{\delta\beta}{\alpha-\beta}R_R$, $C = R_R + \frac{\beta/\delta}{\alpha-\beta}R_C$, $D = \frac{\alpha}{\alpha-\beta}R_R$, $u_0 = (\alpha'-\beta)(t_C-t_R)+p(0)$, $u_0 = \lceil 100u_0 - 1\rceil/100$, *and* $M = \alpha x_R + u_0$.

Table 3.2: A characterization of the equilibrium.

Condition	$\lambda_C(x)$	$\lambda_R(y)$	$u(x)$	x_{max}
(i) $x_C = 0$	A for $0 \leq x < \bar{x}_R$ B for $x_R \leq x \leq x_{max}$	0 for $0 \leq x < x_R$ D for $x_R \leq x \leq x_{max}$	M for $0 \leq x < x_R$ $u_0 + \beta x_R$ for $x_R \leq x \leq x_{max}$	$\{N(\alpha - \beta) + R_C[p(x_{max}) - p(0)] + \alpha R_R x_R - \beta\delta R_R \bar{x}_R\}/\{\alpha R_C + (\alpha - \beta\delta)R_R\}$
(ii) $x_R = 0$	0 for $0 \leq x < x_C$ A for $x_C \leq x < \bar{x}_R$ B for $\bar{x}_R \leq x \leq x_{max}$	C for $0 \leq x < x_C$ D for $x_C \leq x \leq x_{max}$	$\beta(1 - \frac{R_C}{\delta R_R})x + u_0$ for $0 \leq x \leq x_C$ $u(x_C)$ for $x_C \leq x \leq x_{max}$	$x_C + \{N(\alpha - \beta) + R_C[p(x_{max}) - p(0)] - (\beta/\delta)R_C x_C\}/\{\alpha R_C + (\alpha - \beta\delta)R_R\}$

Proposition 3.1 shows that there exist two different cases of equilibrium, depending on when the TNC decides to offer rides during the morning commute: case (i) and case (ii). In determining when to offer rides, the TNC faces a trade-off between the total number of commuters it serves and the fare it charges. On the one hand, if the TNC offers rides early during the morning commute period, it has to reduce its fares to be competitive with the low downtown parking fees. On the other hand, due to the low fares, more commuters choose ride-hailing. In determining their mode of transportation, commuters also face a trade-off between the amount of congestion they experience and the ride-hailing fare. By choosing ride-hailing, a commuter does not need to pay $p(x)$ for parking downtown or experience any downtown congestion $\tau_C(x)$. However, the commuter incurs a fare $u(x)$ for using the ride-hailing service, and might experience some congestion at the drop-off zone, $\tau_R(x)$.

In case (i), described in Table 3.2, $x_C = 0$ and $x_R \geq 0$, which means that the TNC may not offer rides to early commuters. In this case, all commuters who leave early (i.e., $x < x_R$) choose to drive, i.e., $\lambda_C = \frac{\alpha - p'(x)}{\alpha - \beta} R_C > 0$ and $\lambda_R(x) = 0$. This happens because the downtown parking fees are low and it is not profitable for the TNC to offer rides, so it sets fares so high that no commuter chooses ride-hailing, i.e., $u(x)$ is equal to $M = \alpha x_R + u_0$ which is the maximum cost associated with choosing to drive during this time interval. At x_R, the parking fee and cost of downtown congestion become so high that the TNC adjusts its fare so some commuters choose ride-hailing, i.e, $\lambda_R(x) = \frac{\alpha}{\alpha - \beta} R_R > 0$. The TNC's profit margin remains constant for $x \geq x_R$, this guarantees a steady demand for the TNC, i.e., $\lambda_R(x)$ remains constant. The reason is that, if it increases its margins, then more commuters choose to drive, and if it decreases its margins, more commuters choose ride-hailing at a lower profit margin for the TNC. The fact that the TNC's profit margin remains unchanged does not mean that the price commuters pay for ride-hailing remains constant too. As mentioned in §3.3.2, $u(x)$ represents the profit margin of the TNC for a commuter who

leaves at x, but the total ride-hailing cost for this commuter, $\alpha\tau_I(x) + \alpha\tau_R(x) + u(x)$, includes the costs of inbound congestion $\tau_I(x)$ and drop-off congestion $\tau_R(x)$, which increase as more commuters leave H. At $\bar{x}_R$, the number of ride-hailing vehicles waiting to drop off their commuters becomes so high (i.e., $\int_{x_R}^{\bar{x}_R} \lambda_R(x) - R_R dx \geq \Gamma$) that these vehicles interrupt the flow of conventional vehicles and increase the amount of congestion they experience. Due to this extra congestion, for $x \geq \bar{x}_R$, the departure rate of conventional commuters from H decreases to $\lambda_C = \frac{\alpha - p'(x)}{\alpha - \beta} R_C - \frac{\beta\delta}{\alpha - \beta} R_R$ (which is less than $\lambda_C(x)$ for early commuters).

In case (ii), $x_C \geq 0$ and $x_R = 0$, which means that the TNC offers rides that are less expensive than driving for early commuters. In this case, all commuters who leave early (i.e., $x < x_C$) choose ride-hailing, i.e., $\lambda_C = 0$ and $\lambda_R(x) = R_R + \frac{\beta/\delta}{\alpha - \beta} R_C$. The TNC sets its profit margin just below the cost of driving, i.e., $u(x) = u_0$, so while it deters commuters from driving, it maximizes the TNC's profit. As the drop-off congestion cost increases, the TNC decreases its profit margin to guarantee that no commuter chooses to drive. At x_C, it is no longer profitable for the TNC to lower its profit margin, hence, the TNC sets its fare such that some commuters choose to drive, i.e, $\lambda_C = \frac{\alpha - p'(x)}{\alpha - \beta} R_C$ and $\lambda_R(x) = \frac{\alpha}{\alpha - \beta} R_R > 0$ (which is less than $\lambda_R(x)$ for early commuters). Similar to case (i) when both modes of transportation are used, the TNC's profit margin remains constant for $x \geq x_C$. Similar to case (i), at $\bar{x}_R$, due to the extra negativity caused by the drop-off process, the departure rate of conventional commuters decreases, i.e., $\lambda_C = \frac{\alpha - p'(x)}{\alpha - \beta} R_C - \frac{\beta\delta}{\alpha - \beta} R_R$ for $x \in [\bar{x}_R, x_{max}]$.

Case (i) and Case (ii) coincide when $x_C = x_R = 0$. In this scenario, both modes of transportation are chosen by commuters during the entire morning commute window $[0, x_{max}]$. We observe this case in our calibrated model presented in Table 3.1. Since the profit margin of the TNC is constant ($u(x) = u_0 = \$14.755$), the departure rate of commuters who choose ride-hailing remains constant over time, i.e., $\lambda_R(x) = 105.88$ commuters per minute for $x \in [0, x_{max} = 127.85]$. Similarly, by condition A1 the downtown parking fee is assumed to be linearly increasing in departure time x (i.e., $p(x) = 13 + 0.054x$ and $p'(x)$ is constant), the departure rate of commuters who choose to drive remains constant, i.e., $\lambda_C(x) = 85.91$ commuters per minute for $x \in [0, \bar{x}_R = 4.68]$. At $\bar{x}_R$, the negative externality of the drop-off process increases downtown congestion for conventional commuters, resulting in a decrease in the departure rate of conventional commuters to $\lambda_C(x) = 49.21$ commuters per minute for $x \in [\bar{x}_R, x_{max}]$. The daily individual commuter cost, total system

cost, and the optimal TNC profit are equal to \$29.82, \$490,352.40, and \$100,477.84, respectively. Another important insight derived from the equilibrium is that as the negative externality caused by ride-hailing drop-offs increases (i.e., δ increase), more commuters switch from driving to ride-hailing, i.e., $\lambda_C(x)$ is decreasing in δ. In addition, as more commuters choose ride-hailing, the total congestion time for all commuters can increase. This sheds light on the importance of managing the ride-hailing vehicles' operations by either providing designated drop-off zones that do not interfere with the flow of the non-ride-hailing vehicles or imposing drop-off fees on ride-hailing commuters such that they internalize some of this externality.

3.5 Social Optimum

We analyze the case in which a social planner dictates departure times from H and transportation mode, i.e., $\lambda_j(x)$ for $x \in [0, x_{max}]$ and $j \in \{C, R\}$, and the ride-hailing fare, $u(x)$ for $x \in [0, x_{max}]$, that minimize the total system cost in (3.3).[3] This is different from the equilibrium case, under which the goal of the TNC is to maximize its profit, and the goal of a commuter is to minimize his own travel cost regardless of how his decision affects other commuters. Under equilibrium, the commuters' and the TNC's decisions may lead to congestion, and hence increase the total system cost. This is not a desirable social outcome, so we find the *social optimum* (SO) that minimizes the total system cost. We can state this problem as follows:

$$\min_{\lambda_C(x),\lambda_R(x),u(x)} \int_0^{x_{max}} \lambda_H(x)[(\alpha-\beta)\tau_I(x)+\beta(T-x)] + \lambda_C(x)[(\alpha-\beta)\tau_C(x)+(\alpha'-\beta)t_C] \tag{3.5}$$
$$+ \lambda_R(x)[(\alpha-\beta)\tau_R(x)+[(\alpha'-\beta)t_R+u(x)]dx.$$

$$\text{subject to: } \lambda_H(x) = \lambda_C(x) + \lambda_R(x)$$
$$0 \le \lambda_C(x), \lambda_R(x) \le N$$
$$0 \le u(x).$$

In the following proposition, we describe the values of $\lambda_C(x)$, $\lambda_R(x)$ and $u(x)$ that minimize the total system cost in (3.5) and satisfy the constraints.[4] Proposition 3.2 demonstrates that there are

[3] The social planner dictates the lowest fares that keep the TNC in the market. Otherwise, the social planner, who minimizes the total system cost, sets the ride-hailing fare to zero, and the TNC stops offering rides.

[4] As presented earlier in §3.3.1, the first constraint describes the overall departure rate from H, $\lambda_H(x)$, is equal

two forms of SOs: SO1 and SO2.

Proposition 3.2. *(a) [SO1] Suppose* $\frac{\beta N R_C}{2(\alpha' R_C + \beta R_R) R_R} - (t_C - t_R) \geq 0$. *There exists an SO such that* $\lambda_C(x) = R_C$, $\lambda_R(x) = R_R$, $u(x) = (\alpha' - \beta)(t_C - t_R)$ *for* $x \in [0, x_{max}]$, $x_{max} = \frac{N}{R_R + R_C}$, *and* $T = x_{max} + t_C$.

(b) [SO2] Suppose $\frac{\beta N R_C}{2(\alpha' R_C + \beta R_R) R_R} - (t_C - t_R) < 0$. *There exists an SO such that* $\lambda_C(x) = 0$, $\lambda_R(x) = R_R$, $u(x) = \frac{R_R}{R_R + R_C}(\alpha' - \beta)(t_C - t_R)$ *for* $x \in [0, x_{max}]$, $x_{max} = \frac{N}{R_R}$, *and* $T = x_{max} + t_R$.

Proposition 3.2 states that the social planner's decisions on commuters' departure times and transportation mode, and the ride-hailing fares follow either SO1 or SO2. The social planner decides between these two SOs based on the total system costs associated with them. In particular, when the total system cost of SO1 is lower than that of SO2, which can be stated as when $\frac{\beta N R_C}{2(\alpha' R_C + \beta R_R) R_R} - (t_C - t_R) \geq 0$, the social planner wants commuters and the TNC to follow SO1. Otherwise, SO2 is the optimal solution. We first discuss SO1, which pertains to our calibrated model, and then SO2.

In our calibrated model, $\frac{\beta N R_C}{2(\alpha' R_C + \beta R_R) R_R} - (t_C - t_R) = 30.18 > 0$, so we observe SO1. Under SO1, from the social planner's perspective, the cost of driving, i.e., $(\alpha' - \beta) t_C = \$2.03$, is equal to the cost of using rid-hailing, i.e., $u(x) + (\alpha' - \beta) t_R = (\alpha' - \beta)(t_C - t_R) + (\alpha' - \beta) t_R = (\alpha' - \beta) t_C = \$2.03$, so she assigns the maximum number of commuters per minute that does not create any congestion to use each of the transportation modes, i.e., $\lambda_C(x) = R_C = 76.67$ commuters per minute and $\lambda_R(x) = R_R = 60$ commuters per minute which result in $\tau_C(x) = \tau_R(x) = 0$ for $x \in [0, x_{max}]$.

For the calibrated model, the SO pattern observed is different from the equilibrium pattern described in §3.4. In fact, the daily total system cost under SO ($\$111,439.98$) is less than a quarter of that under equilibrium ($\$490,352.40$). This discrepancy stems from high drop-off and downtown congestion times under equilibrium. Under SO, the departure rates from H for for conventional and ride-hailing commuters are equal to the capacities of most downstream bottlenecks they go through, i.e., $\lambda_C(x) = R_C$ and $\lambda_R(x) = R_R$, so commuters do not experience any congestion. In contrast, under equilibrium, the departure rate $\lambda_R(x) = 105.88$ commuters per minute is always higher than the drop-off rate $R_R = 60$ commuters per minute, so all ride-hailing commuters incur a positive drop-off congestion cost. In addition, under equilibrium the conventional commuters experience a

to the sum of the departure rates of conventional and ride-hailing commuters, i.e., $\lambda_C(x) + \lambda_R(x)$. The second constraint guarantees that the departure rates are positive and do not exceed the total number of commuters, N. The last constraint guarantees the non-negativity of the ride-hailing fare $u(x)$.

positive downtown congestion cost, because of the negative externality of the ride-hailing drop-offs even though the departure rate of conventional commuters is lower than the capacity of roads downtown, i.e., $\lambda_C(x) = 49.53 < 76.67 = R_C$. The low total system cost under SO also means a lower profit for the TNC. In fact, the TNC's profit reduces substantially from $\$100,477.84$ under equilibrium to $\$15,409.76$ under SO. This shows that the TNC's high profit under equilibrium comes at the expense of creating a significant amount of congestion that can be eliminated if the social planner can persuade commuters to follow the SO solution.

We compare SO1 with SO in the model where all commuters drive. We observe that having ride-hailing as an alternative to driving during the morning commute can improve social welfare. In fact, there does not exist an SO solution such that all commuters are assigned to drive, i.e., $\lambda_C(x) > 0$ and $\lambda_R(x) = 0$ for $x \in [0, x_{max}]$, because the total system cost of this scenario is always higher than that of SO1. Under this scenario, the cost associated with driving, $(\alpha' - \beta)t_C$, is equal to that under SO1, but the total duration of the morning commute, i.e., x_{max} minutes, is longer because the total number of commuters who leave H per minute, $\lambda_H(x) = \lambda_C(x) = R_C$, is lower than that under SO1, i.e., $\lambda_H(x) = \lambda_C(x) + \lambda_R(x) = R_C + R_R$. Thus, ride-hailing leads to a lower total system cost by shortening the duration of the morning rush hour.

Finally, under SO2, from the social planner's perspective, the cost of driving per commuter, i.e., $(\alpha' - \beta)t_C = \$2.03$, is higher than that of using ride-hailing, i.e., $u(x) + (\alpha' - \beta)t_R = \frac{R_R}{R_R+R_C}(\alpha' - \beta)(t_C - t_R) + (\alpha' - \beta)t_R = (\alpha' - \beta)t_C - \frac{R_C}{R_R+R_C}(\alpha' - \beta)(t_C - t_R) = \$1.04$, so she assigns the maximum number of commuters per minute that does not create any congestion to use ride-hailing, i.e., $\lambda_C(x) = 0$ and $\lambda_R(x) = R_R = 60$ commuters per minute for $x \in [0, x_{max}]$.[5] This result is particularly interesting as it represents the possibility of moving toward a world where no one uses (or even owns) personal vehicles.

3.6 Reducing the Total System Cost of the Morning Commute

As discussed in §3.5, commuters' decisions under equilibrium are different from the SO decisions. In this section, we examine solutions that the social planner may adopt in order to close the

[5]The reason why SO2 results in a higher total system cost is that it extends the duration of the morning commute from $x_{max} = \frac{N}{R_C+R_R} = 20,000/[(4600+3600)/60] = 146.34$ minutes under SO1 to $x_{max} = \frac{N}{R_R} = 20,000/(3600/60) = 333.33$ minutes under SO2.

gap between the total system cost under equilibrium and that under SO: a short-term solution of regulating parking fees and curbside drop-off tolls in §3.6.1 and a long-term solution of adjusting parking and curbside drop-off capacities in §3.6.2.

3.6.1 Pricing and Tolling Schemes

To reduce the total system cost, the social planner can use two levers that are commonly used in practice (Federal Highway Administration (2020) and Meiszner (2019)): parking fees and curbside drop-off tolls. We consider a dynamic parking pricing scheme that enables the social planner to regulate downtown parking fee $p(x)$ based on departure time x. A drop-off tolling scheme, denoted as $\pi(x)$, is also imposed on commuters who choose ride-hailing based on their departure time x to balance downtown congestion $\tau_C(x)$ and drop-off congestion $\tau_R(x)$. Proposition 3.3 describes these pricing and tolling schemes.

Proposition 3.3. *(a) Suppose* $\frac{\beta N R_C}{2(\alpha' R_C + \beta R_R) R_R} - (t_C - t_R) \geq 0$. *Under equilibrium,* $\lambda_C(x)$, $\lambda_R(x)$ *and* $u(x)$ *follow those in SO1, when* $p(x) = \pi(x) = \beta x$ *for* $0 \leq x \leq x_{max}$.
(b) Suppose $\frac{\beta N R_C}{2(\alpha' R_C + \beta R_R) R_R} - (t_C - t_R) < 0$. *Under equilibrium,* $\lambda_C(x)$, $\lambda_R(x)$ *and* $u(x)$ *follow those in SO2, when* $p(x) > \beta x - (\alpha' - \beta)(t_C - t_R)\frac{R_C}{R_R + R_C}$ *and* $\pi(x) = \beta x$ *for* $0 \leq x \leq x_{max}$.

Proposition 3.3 indicates that there exist a parking fee scheme $p(x)$ and a drop-off toll scheme $\pi(x)$ such that the SO presented in Proposition 3.2 results in the same travel cost for all commuters regardless of their departure time or transportation mode. In other words, when these parking fees and congestion tolls are imposed, decisions of commuters under equilibrium match those of SO. This short-term solution is particularly important to the social planner, as it enables her to influence commuters' decisions and reduce aggregate congestion and travel costs of the morning commute. We first discuss the pricing and tolling scheme for SO1, which pertains to our calibrated model, and then those for SO2. Finally, we discuss demand elasticity under SO.

As mentioned in §3.5, under SO1 the social planner wants commuters who leave at $x \in [0, x_{max}]$ to use both modes of transportation. Since there is no congestion under SO, i.e., $\tau_I(x) = \tau_C(x) = \tau_R(x) = 0$, and the ride-hailing fare $u(x)$ is a constant value, the only time-varying element of the travel cost of a commuter who leaves H at x, presented in (3.1) and (3.2), is the work schedule penalty, i.e., $\beta(T - x)$. Hence, to guarantee an equal travel cost for all commuters, the parking fee

or drop-off toll that a commuter who leaves H at x pays must offset the time-varying portion of the work schedule penalty, i.e., $p(x) = \pi(x) = \beta x$. In the calibrated model, under SO1 the duration of the morning commute window is $x_{max} = 146.34$ minutes, so the parking fee and drop-off toll range from zero (for commuters who leave at time zero) to $\beta x_{max} = \frac{3.9}{60} \times 146.34 = \$9.51$ (for commuters who leave at x_{max}). This indicates that under SO the downtown parking fee for conventional commuters decreases (from at least \$13 under equilibrium to at most \$9.51 under SO), while a curbside drop-off fee is imposed on ride-hailing commuters to account for the negative externality they create. The decrease in parking fee does not mean that the city's total revenue decreases. In fact, the city's daily revenue does not change significantly: it increases by 0.06% from $\$106,055.65$ in parking revenue under equilibrium to $\$106,722.19$ in parking and toll revenue under SO. So instead of collecting high parking fees from only 32% of commuters under equilibrium, the city can increase its revenue by collecting lower parking fees or drop-off tolls from all commuters under SO. These pricing schemes are also beneficial for commuters, as their daily travel cost decreases from \$29.82 under equilibrium to \$12.51 under SO. This shows that, without regulating parking fees and drop-off tolls, commuters overpay due to congestion and expensive parking fees. In fact, the TNC takes advantage of this unregulated market and makes a profit of $\$100,477.84$, while after imposing the pricing schemes, its profit significantly decreases to $\$15,409.76$. However, there is a risk associated with regulating parking fees and imposing drop-off tolls: the TNC might stop offering rides due to its low profit. One recommendation for solving this issue is for the city to (partially) subsidize the drop-off tolls to incentivize the TNC to continue its operations.

Under SO2, similar to SO1, the drop-off toll is set such that all commuters incur the same travel cost, i.e., $\pi(x) = \beta x$ for commuters who leave at x. However, since the social planner does not want any commuters to drive, she sets the parking fee $p(x)$ higher than the minimum amount that leads to the same travel cost for driving, i.e., $\beta x - (\alpha' - \beta)(t_C - t_R)\frac{R_C}{R_R + R_C}$.

As mentioned, the daily individual travel cost of a commuter decreases significantly under SO. This might incentivize some commuters who currently use alternative modes of transportation, such as public transportation, walking, and biking, to switch to driving or using ride-hailing. Corollary 3.1 describes the relationship between the total number of commuters, N, and the daily travel cost under SO, denoted by C.

Corollary 3.1. $N = \frac{R_R+R_C}{\beta}(C - \alpha' t_C)$.

Corollary 3.1 shows that the number of daily commuters, N, increases until the maximum travel cost that commuters are willing to pay is reached. For example, if the commuters are willing to have a daily travel cost of up to \$15, then the number of daily commuters increases from its current value of $20,000$ to $N = \frac{R_R+R_C}{\beta}(C - \alpha' t_C) = \frac{3600+4600}{3.9}(15 - 12 \times (15/60)) = 27,333$. The increase in the number of daily commuters prolongs the duration of the morning commute (from 146.34 minutes in the calibrated model to $x_{max} = N/(R_R + R_C) = 200$ minutes in this example) as there are more commuters who need to park or be dropped off.

3.6.2 Improving the Infrastructure

Although parking pricing and drop-off tolling are practical tools for reducing the cost of the morning commute, the social planner should also explore a long-term sustainable plan. In particular, since in both SOs the social planner wants commuters to use ride-hailing, there is a need for more drop-off spots. However, the social planner faces a trade-off. On the one hand, increasing the drop-off capacity reduces the drop-off congestion time experienced by ride-hailing commuters, and the externality of ride-hailing drop-offs on downtown congestion experienced by conventional commuters. On the other hand, since drop-off congestion decreases, ride-hailing demand increases, which can lead to higher ride-hailing fares.

The following corollary derives the optimal drop-off capacity, denoted by R_R^*, that minimizes the total system cost of the morning commute under SO.[6]

Corollary 3.2. *(a) If* $\frac{\beta N}{2R_C} > (\alpha' - \beta)(t_C - t_R)$, *then* $R_R^* = R_I$.
(b) Otherwise, $R_R^* = \min\{R_I, \max\{\ \frac{R_C(\beta N+\sqrt{2\beta N R_C(\alpha'-\beta)(t_C-t_R)})}{2R_C(\alpha'-\beta)(t_C-t_R)-\beta N}, \frac{\alpha' R_C+\sqrt{(\alpha' R_C)^2+2\beta^2 N R_C/(t_C-t_R)}}{2\beta}\}\}$.

Corollary 3.2 shows that the optimal value of curbside drop-off capacity, R_R^*, can at most be equal to the inbound bottleneck capacity R_I. This happens because increasing the drop-off capacity beyond the inbound bottleneck capacity does not reduce the cost, as commuters will experience congestion in the inbound bottleneck. To effectively improve the total system cost, the social planner should increase the inbound bottleneck capacity simultaneously with the drop-off capacity, because otherwise, one of them becomes the bottleneck and the other one has some

[6] We assume there is no cost associated with converting a parking space to a drop-off space. However, such a cost can easily be incorporated in our model, as it is linear in the number of converted parking spots.

underutilized capacity. In addition, when the ride-hailing fare under SO is low, i.e., part (a) where $\frac{\beta N}{2R_C} > (\alpha' - \beta)(t_C - t_R)$, we recommend the social planner to increase drop-off capacity to its maximum value of R_I.[7] In this case the positive effect of reducing congestion dominates the negative effect of increased ride-hailing fares, as the ride-hailing fare is low and there is room for a surge. We observe this scenario in our calibrated model, where $\frac{\beta N}{2R_C} = 1300 > 269.1 = (\alpha' - \beta)(t_C - t_R)$. Hence, the value of the drop-off capacity should increase from its current value of $3,600$ drop-offs per hour to the value of the inbound bottleneck capacity, which is $4,600$ drop-offs per hour. In this scenario, we still observe SO1, and the total system cost decreases by 8% from $\$111,439.98$ to $\$102,634.17$. When the ride-hailing fare under SO is high, i.e., part (b) where $\frac{\beta N}{2R_C} \leq (\alpha' - \beta)(t_C - t_R)$, then it might not be optimal to increase drop-off capacity all the way to its maximum value of R_I. In particular, the optimal value of R_R becomes equal to either the maximum value that guarantees that SO1 is observed and driving is a viable option, i.e., $\frac{\alpha' R_C + \sqrt{(\alpha' R_C)^2 + 2\beta^2 N R_C/(t_C - t_R)}}{2\beta}$, or the value that balances the ride-hailing fare and congestion under SO2 when no commuter chooses to drive, i.e., $\frac{R_C(\beta N + \sqrt{2\beta N R_C(\alpha' - \beta)(t_C - t_R)})}{2R_C(\alpha' - \beta)(t_C - t_R) - \beta N}$.

3.7 Conclusion

In this paper we investigate the effect of ride-hailing on morning rush hour congestion and traffic patterns. We characterize the departure time and mode of transportation for commuters and ride-hailing fares for the TNC under equilibrium. Our model also takes into account the capacity of roadways and the monetary value of travel time (both while in a car and walking).

We show that in an unregulated market the TNC, that maximizes its profit, can create a significant amount of drop-off and downtown congestion. As more commuters use ride-hailing, the negative externality of ride-hailing drop-offs increases, hence, some commuters switch from driving (and parking) their individual vehicles to using ride-hailing. However, if regulated properly, ride-hailing services can not only reduce the daily travel costs for commuters, but also mitigate downtown congestion.

To understand the type of regulatory actions needed, we compare the equilibrium patterns against those determined by a social planner who minimizes the total system cost. In fact, the

[7] The ride-hailing fare under both types of SO is a function of $(\alpha' - \beta)(t_C - t_R)$, which is the right hand side of the condition for part (a) of Corollary 3.2.

socially optimal decisions are significantly different from those of commuters and the TNC under equilibrium. Under equilibrium commuters' departure rate can exceed highway and road capacities, creating a substantial amount of congestion, whereas under SO the departure rates of each type of commuter is equal to the most downstream bottleneck they go through such that there is no congestion. The TNC profit margin per commuter is lower under SO than that under equilibrium, because the social planner directs commuters to drive (by adjusting downtown parking fees) if ride-hailing fares are high, so the TNC is forced to reduce its fares. This shows that the TNC's high profit under equilibrium comes at the expense of creating downtown and drop-off congestion.

To regulate the market, we propose two levers that are commonly used in practice: dynamic parking fees for commuters who drive and dynamic curbside drop-off tolls for commuters who use ride-hailing. These two levers influence commuters to choose the transportation mode option that the social planner intends. By imposing optimal parking fees and drop-off tolls, not only does the social planner's revenue increase, the commuters' daily travel cost also decreases. In fact, the drop-off toll plays an important role in reducing total system cost, as it results in internalizing some of the drop-off process's negative externalities by ride-hailing commuters. In addition, since ride-hailing commuters pay the drop-off toll, they are no longer willing to pay high ride-hailing fares, resulting in a lower total profit for the TNC. There is, however, a risk associated with this strategy: the TNC might cease its operations due to the decline in profits. Since having ride-hailing as an option for commuting to work leads to a lower total system cost (there is no SO solution where all commuters are assigned to drive), the social planner can allocate some of its revenue as subsidy to the TNC to encourage the TNC to continue offering rides.

To further reduce total system cost, we propose creating dedicated drop-off zones downtown. We derive the optimal increase in curbside drop-off capacity under SO. This increase in capacity accelerates the morning commute and shortens the duration of the morning rush hour. This happens because the departure rate of ride-hailing commuters is equal to the drop-off capacity (to avoid causing drop-off congestion), so more commuters can leave their residences per hour.

To the best of our knowledge, this is the first attempt to characterize the impact of ride-hailing on downtown congestion patterns with endogenous departure time and transportation mode decisions. Our paper sheds light on the importance of jointly regulating parking and drop-off fees as well as improving infrastructure to reap the full benefit of ride-hailing as an alternative to

driving for morning commuters. There are various ways to explore other extensions to our model. For example, one could consider stochastic capacities for the inbound highway, drop-off zone and downtown roads. In addition, our model can be extended to include public transportation as a mode of transportation. Our model and analysis provide a foundation for understanding how ride-hailing affects traffic and congestion patterns during the morning commute. We hope that our work helps pave the road for future research on this matter.

Chapter 4

Conclusions

In this dissertation, we study three effects of innovative transportation technologies on city operations. For each of these topics, we provide mathematical models and analysis, and calibrate our models to data to offer practical guidance for city planners. Our results indicate that, despite the discernible benefits of these technologies, they might result in unintended negative consequences (e.g., more congestion). As such, we recommend policies that can mitigate these consequences and accentuate the benefits of these innovative technologies.

In chapter 1, we investigate the effects of AVs on highway congestion by modeling a segment of a highway as a queueing system. We analyze two policies for a mixed fleet of HVs and AVs: the D policy, and the I policy. Using the queueing model, we compare the mean travel time of a single vehicle as well as the throughput of the highway under each of the two policies, and also against a benchmark case in which all vehicles are HVs. The queueing model captures the potential benefits of AVs by taking into account platoons and headway, which are explicitly modeled using a Markovian arrival process (MAP). The difference in the mix of vehicle fleets leads to different vehicle headways and speeds (hence, different service rates) under different policies.

Our analysis shows that, contrary to industry experts predictions, the D policy should not be employed when the AV penetration rate is low. As such, it would be beneficial for U.S. states that are considering designating a lane to AVs in the near future to reevaluate this policy. In fact, when the AV penetration rate is low, the D policy is not only outperformed by the I policy both in terms of mean travel time and throughput, but also it is inferior to the benchmark case which represents the status quo. The I policy, however, performs surprisingly well, especially for highly congested

highways: under this policy, a moderate number of AVs can make a substantial improvement.

In chapter 2, we study the effects of AVs on morning rush hour congestion and parking patterns. For this problem, a continuous-time game theoretic model is used to analytically characterize commuters' daily decisions on departure time and parking location. This model is calibrated to data from Pittsburgh to provide practical recommendations for city planners on parking and tolls pricing, as well as adjusting infrastructure to the needs and characteristics of AVs.

The results of chapter 2 shed light on the potential negative consequences of AVs on the morning commute, should appropriate policies not be put in place. Industry experts believe that the ability of AVs to drop-off commuters at work and park outside downtown will mitigate downtown parking demand and congestion. However, under equilibrium, the outward flow of AVs from downtown to the outside downtown parking area and the drop-off process may create a significant amount of congestion, resulting in a higher total system cost than in the case with no AVs. By imposing optimal parking and toll prices, that are analytically derived in this chapter, a social planner can reduce the total system cost to its minimum value. This is, however, a short-term solution, as the social planner might choose to direct AVs to park downtown in order to reduce congestion. To reduce both congestion and downtown parking demand, we propose converting some downtown parking spots to drop-off spots. This not only encourages AVs to park outside downtown, but also reduces drop-off congestion.

Lastly, in chapter 3, we investigate the effect of ride-hailing drop-offs on morning rush hour congestion and commuters' choice with respect to mode of transportation. Similar to chapter 2, a continuous time game theoretic model is used to derive commuters' departure time and transportation mode decisions. By calibrating the model to traffic and parking data from Pittsburgh, we provide pragmatic short- and long-term recommendations for city planners.

Our results show that ride-hailing can be beneficial if proper policies are put in place. Ride-hailing services reduce downtown parking demand and might also mitigate downtown congestion, as ride-hailing commuters are dropped off at work without the need to search for finding an empty parking spot. However, as ride-hailing vehicles accumulate to drop off their commuters, they can interrupt the flow of conventional commuters who head to downtown parking areas. In fact, as the negative externality from ride-hailing drop-offs increases, more conventional commuters may decide to switch to ride-hailing, creating even more congestion. In addition, under equilibrium, by

strategically deciding on its ride-hailing fare and when to offer rides, the TNC benefits from the lack of regulations. This results in a higher total system cost than in the case with no ride-hailing. We find that a social planner is able to reduce the total system cost as well as the individual commuter cost by imposing dynamic parking fees and drop-off tolls. Since it is not socially optimal to completely remove ride-hailing as a transportation mode available to morning commuters, the social planner should ensure that the infrastructure is ready for these type of services by creating more curbside drop-off spots.

There exist a number of interesting questions for future research that can be derived from this dissertation. In chapter 1, the potential behavioral effects of either of the AV policies is not discussed. Future research could examine how implementing the D policy might affect commuters adoption of AVs, and how these effects differ from those of implementing the I policy. In addition, as AVs become more prevalent, more data can be gathered to yield more precise results.

Chapter 2 provides a good starting point for understanding the potential impacts of AVs on morning rush hour traffic and parking patterns, but future research could continue to explore this topic. For example, AVs might have the option to circle around downtown instead of parking. It would be important to investigate the effect of this option on downtown congestion and parking. In addition, future research is needed to determine the sustainability implications of this option. Another avenue to explore is to consider a mixed fleet of AVs and HVs.

For chapter 3, which investigates the effect of ride-hailing on the travel time of commuters who drive to work, further attempts could prove quite beneficial to the literature. For example, one could asses how the interaction between commuters (e.g., example carpooling) might play a role in commuters' decisions and morning rush hour traffic patterns. Adding public transportation as a third transportation option could be another interesting topic to pursue. Finally, after the COVID-19 pandemic, the possibility of work-from-home warrants further investigation.

Appendix A

Additional Material for Chapter 1

A.1 Summary of Notation

Notation	Definition	Notation	Definition
λ	Arrival rate to the highway	N	Number of lanes
L	Length of the highway segment	J	Jam density of the highway
c	Highway capacity	$\boldsymbol{C}$	Irreducible generator matrix of the MAP
$\boldsymbol{C}^0$	Transition rate matrix to the non-absorbing states of the MAP	$\boldsymbol{C}^1$	Transition rate matrix to the absorbing state of the MAP
$\boldsymbol{\delta}^0$	Initial distribution of the non-absorbing states of the platoon size distribution	δ	Initial distribution of the absorbing state of the platoon size distribution
$\boldsymbol{G}^0$	Probability transition matrix associated with the non-absorbing states of the platoon size distribution	$\boldsymbol{G}$	Probability transition matrix associated with the absorbing state of the platoon size distribution
δ^i	Parameter of the platoon size distribution (or equivalently the reciprocal of the mean platoon size) in the model $i \in \{B, DA, DH, I\}$	$\boldsymbol{M}$	Transition probability matrix of the platoon size distribution
$\boldsymbol{\alpha}_n^0(1)$	Initial distribution of the non-absorbing states of the state-dependent intraplatoon headway distribution	$\alpha_n(1)$	Initial distribution of the absorbing state of the state-dependent intraplatoon headway distribution
$\boldsymbol{Q}_n^0(1)$	Transition rate matrix associated with the non-absorbing states of the state-dependent intraplatoon headway distribution	$\boldsymbol{Q}_n(1)$	Transition rate matrix associated with the absorbing state of the state-dependent intraplatoon headway distribution
$\boldsymbol{\alpha}_n^0(2)$	Initial distribution of the non-absorbing states of the state-dependent interplatoon headway distribution	$\alpha_n(2)$	Initial distribution of the absorbing state of the state-dependent interplatoon headway distribution
$\boldsymbol{Q}_n^0(2)$	Transition rate matrix associated with the non-absorbing states of the state-dependent interplatoon headway distribution	$\boldsymbol{Q}_n(2)$	Transition rate matrix associated with the absorbing state of the state-dependent interplatoon headway distribution
$1/\xi_n^i$	Mean intraplatoon headway when there are n vehicles on the highway in the model $i \in \{B, DA, DH, IA, IR, IHA, IAH, I\}$	$1/\psi_n^i$	Mean interplatoon headway when there are n vehicles on the highway in the model $i \in \{B, DA, DH, IA, IR, IHA, IAH, I\}$
μ_n^i	Service rate of a single vehicle when there are n vehicles on the highway in the model $i \in \{B, DA, DH, I\}$	$V_n^I(p)$	Speed of a single vehicle when there are n vehicles on the highway in the model $i \in \{B, DA, DH, I\}$, and the AV proportion is p
k	Highway density	q	Highway flow
h_n	Mean headway when there are n vehicles on the highway	d_n	Stopping distance when there are n vehicles on the highway
$\tilde{\boldsymbol{\pi}}$	Steady state distribution of the MAP	$\boldsymbol{\pi}$	Steady state distribution of an $M/G_n/c/c$ queueing model
p	Proportion of AVs	π_c	Blocking probability
W	Mean travel time of vehicles	θ	Throughput of the highway

Table A1: Table of Notation

Throughout this paper, for model parameters, we use a superscript $i \in \{B, D, DH, DA, I\}$ and a subscript $n \in \{1, 2, \cdots, Nc\}$, where B, D, DH, DA, and I represent the benchmark case, the D policy, the HV queue of the D policy, the AV queue of the D policy, and the I policy, respectively, and n is the number of vehicles on highway. Moreover, for thresholds of λ and p, we use superscripts

(i, j) for $i \in \{D, I, DI\}$ and $j \in \{W, \theta, W\theta\}$, where D, I, and DI represent the D policy, the I policy, and comparison between D and I policies, respectively; and W, θ, and $W\theta$ indicate thresholds for W, θ, and comparison between W and θ, respectively. We also use underscore and overscore to indicate smaller and larger thresholds, respectively.

Notation	Definition	Reference
$\underline{\lambda}^{(D,W)}$	The arrival rate threshold before which the D policy has a higher W than the benchmark case.	Proposition 1.1
$\bar{\lambda}^{(D,W)}$	The arrival rate threshold after which the D policy may have a lower W than the benchmark case.	Proposition 1.1
$\lambda^{(D,\theta)}$	The arrival rate threshold after which the D policy may have a higher θ than the benchmark case.	Proposition 1.1
$\bar{\lambda}^{(D,W\theta)}$	The arrival rate threshold after which there exists an interval of p such that the D policy tends to increase θ, but it does not decrease W over those of the benchmark case.	Corollary 1.1
$\underline{\lambda}^{(D,W\theta)}$	The arrival rate threshold after which there exists an interval of p such that the D policy tends to increase θ and decrease W over those of the benchmark case.	Corollary 1.1
$\lambda^{(I,W)}$	The arrival rate threshold after which the I policy may have a lower W than the benchmark case.	Proposition 1.2
$\lambda^{(I,\theta)}$	The arrival rate threshold after which the I policy may have a higher θ than the benchmark case.	Proposition 1.2
$p^{(D,W)}$	The AV proportion threshold after which the D policy has a lower W than the benchmark case for a highly loaded highway.	Proposition 1.1
$\underline{p}^{(D,\theta)} \leq p \leq \bar{p}^{(D,\theta)}$	The AV proportion interval in which the D policy has a higher θ than the benchmark case for a highly loaded highway.	Proposition 1.1
$p^{(DI,W)}$	The AV proportion threshold after which the D policy has a lower W than the I policy for a highly loaded highway.	Proposition 1.3
$\underline{p}^{(DI,\theta)} \leq p \leq \bar{p}^{(DI,\theta)}$	The AV proportion interval in which the D policy has a higher θ than the I policy for a highly loaded highway.	Proposition 1.3

Table A2: Table of Thresholds

A.2 MAP Characterization

As mentioned in §1.3.1.2, to model platooning on a highway using a MAP, one needs to specify the distributions of the following three elements: (1) the size of each platoon, (2) the time gap between two consecutive vehicles traveling in the same platoon ("intraplatoon headway"), and (3) the time gap between the last vehicle of one platoon and the first vehicle of the following platoon ("interplatoon headway"). We model the size of a platoon using a discrete phase type (PH-type) distribution of order l. A distribution on $1, 2, \cdots, l$ is a discrete phase-type distribution if it is the distribution of the first passage time to the absorbing state of a Markov chain with l states,

such that state l is absorbing and the rest of the states are transient. Any discrete distribution can be written as a PH-type distribution. This distribution is represented by $(\boldsymbol{\delta}^0, \boldsymbol{G}^0)$. The vector $\boldsymbol{\delta}^0$ corresponds to the probability of starting at the non-absorbing states $1, 2, \cdots, l-1$. Similarly, δ is the probability of starting at the absorbing state l. The vector $(\delta, \boldsymbol{\delta}^0)$ represents the initial distribution of states, where $\delta + \boldsymbol{\delta}^0\mathbf{1} = 1$. The $(l-1) \times (l-1)$ matrix $\boldsymbol{G}^0$ is the probability matrix associated with non-absorbing transitions (transitions among non-absorbing states). Analogously, $\boldsymbol{G}$ corresponds to transitions to the absorbing state, satisfying $\boldsymbol{G}^0\mathbf{1} + \boldsymbol{G} = \mathbf{1}$, where $\mathbf{1}$ is a vector of ones. The transition probability matrix of this discrete-time Markov chain (DTMC) is then $M = \left[\begin{smallmatrix} \boldsymbol{G}^0 & \boldsymbol{G} \\ \mathbf{0} & 1 \end{smallmatrix}\right]$. We present the example of a uniform distribution on $1, 2, \cdots, l$ in Example 1, while presenting another example of a geometric distribution in Example 2 at the end of this appendix.

We model the intraplatoon headway, when there are n vehicles in the system, using a continuous PH-type distribution of order m_1 represented by $(\boldsymbol{\alpha}_n^0(1), \boldsymbol{Q}_n^0(1))$, having mean $1/\xi_n$. The vector $\boldsymbol{\alpha}_n^0(1)$ (resp., $\alpha_n(1)$) demonstrates the initial distribution of the non-absorbing states (resp., the absorbing state). The matrices $\boldsymbol{Q}_n^0$ and $\boldsymbol{Q}_n$ represent a non-absorbing transition rate matrix and an absorbing transition rate matrix, respectively.

Lastly, we also model the interplatoon headway, when there are n vehicles in the system, using a continuous PH-type distribution of order m_2 represented by $(\boldsymbol{\alpha}_n^0(2), \boldsymbol{Q}_n^0(2))$, and with mean $1/\psi_n$.

Having these three elements specified, the platooning process of a *single* lane can be characterized as a MAP with the following matrices:

$$\boldsymbol{C}_n^0 = \begin{bmatrix} \boldsymbol{Q}_n^0(2) & 0 \\ 0 & \boldsymbol{I} \otimes \boldsymbol{Q}_n^0(1) \end{bmatrix}, \text{ and } \boldsymbol{C}_n^1 = \begin{bmatrix} \delta \boldsymbol{Q}_n(2)\boldsymbol{\alpha}_n^0(2) & \boldsymbol{\delta}^0 \otimes \boldsymbol{Q}_n(2)\boldsymbol{\alpha}_n^0(1) \\ \boldsymbol{G} \otimes \boldsymbol{Q}_n(1)\boldsymbol{\alpha}_n^0(2) & \boldsymbol{G}^0 \otimes \boldsymbol{Q}_n(1)\boldsymbol{\alpha}_n^0(1) \end{bmatrix}. \tag{A.1}$$

The size of these matrices is $m \times m$, where $m = m_2 + m_1(l-1)$. See Example 2 in Appendix A.2.

To model platooning on a highway with $N > 1$ lanes, we consider a MAP with matrices $(\boldsymbol{C}_n^0, \boldsymbol{C}_n^1)$, and assume that a vehicle joins one of the lanes with probability $1/N$. The platoon formation process on a specific lane itself is a MAP with matrices $(\boldsymbol{C}_n^0 + (1 - \frac{1}{N})\boldsymbol{C}_n^1, \frac{1}{N}\boldsymbol{C}_n^1)$. This is called the random thinning of a MAP (see Proposition 2.2.3 in He (2014) for more details).

Example 1. (Discrete PH-type distribution) A uniform distribution on $1, 2, \cdots, l$ can be

[-¿, ¿=stealth', auto, semithick, node distance=2.5cm] every state=[fill=white,draw=black,thick,text=black,scale=1,inner sep=2pt] [state] (1) 1; [state] (2)[right of=1] 2; [state] (3)[right of=2] 3; (6)[right of= 3] ...; [state] (4)[right of= 6] $l-1$; [state] (5)[right of= 4] l;
(1) edge[bend left] node $\frac{l-2}{l-1}$ (2) edge[bend right, pos=0.1, below] node$\frac{1}{l-1}$ (5) (2) edge[bend left] node $\frac{l-3}{l-2}$ (3) edge[bend right, pos=0.1, below] node$\frac{1}{l-2}$ (5) (3) edge[bend left] node $\frac{l-4}{l-3}$ (6) edge[bend right,pos=0.1, below] node$\frac{1}{l-3}$ (5) (6) edge[bend left] node $\frac{1}{2}$ (4) (4) edge[bend left] node 1 (5) ;

Figure A.1: The Markov chain associated with a uniform distribution on $1, 2, \cdots, l$.

represented as: $\boldsymbol{G}^0 = \begin{bmatrix} 0 & \frac{l-2}{l-1} & 0 & \cdots & 0 \\ 0 & 0 & \frac{l-3}{l-2} & \cdots & 0 \\ & & & \ddots & \\ 0 & 0 & 0 & \cdots & \frac{1}{2} \\ 0 & 0 & 0 & \cdots & 0 \end{bmatrix}$, and $\boldsymbol{\delta}^0 = [\frac{l-1}{l}, 0, \cdots, 0]$. Note that for this Markov chain the first passage time to the absorbing state l happens with probability $\frac{1}{l}$ in i steps, where $i = 1, 2, \cdots, l$. The Markov chain associated with this uniform distribution is illustrated in Figure A.1. In this Markov chain, transition to l happens in one step, only if we start at l, and this happens with probability $\frac{1}{l}$. If we start at state one and then go to l, it happens in two steps with the probability $\frac{l-1}{l} \times \frac{1}{l-1} = \frac{1}{l}$. Repeating this process, one can observe that transition to the absorbing state happens in any i ($\in \{1, 2, \cdots, l\}$) steps with the same probability $\frac{1}{l}$, representing a uniform distribution on $1, 2, \cdots, l$. Note that uniform distribution is only a special case of the PH-type distributions, and other distributions have different $\boldsymbol{\delta}^0$ and $\boldsymbol{G}^0$.

Example 2. (MAP) Suppose the platoon size follows a geometric distribution with parameters $\delta = G$ and $\delta^0 = G^0 = 1 - \delta$. Assuming the interplatoon and intraplatoon headways follow $exp(\psi)$ and $exp(\xi)$, respectively, we have

$$\boldsymbol{C}^0 = \begin{bmatrix} -\psi & 0 \\ 0 & -\xi \end{bmatrix} \text{ and } \boldsymbol{C}^1 = \begin{bmatrix} -(1-\delta)\psi & (1-\delta)\psi \\ \delta\xi & -\delta\xi \end{bmatrix}.$$

The generator matrix of the MAP is then

$$\boldsymbol{C} = \boldsymbol{C}^0 + \boldsymbol{C}^1 = \begin{bmatrix} -(1-\delta)\psi & (1-\delta)\psi \\ \delta\xi & -\delta\xi \end{bmatrix}.$$

The stationary probability vector of this MAP, $\tilde{\boldsymbol{\pi}}$, is given by

$$\tilde{\boldsymbol{\pi}}\boldsymbol{C} = \mathbf{0}, \text{ and } \tilde{\boldsymbol{\pi}} = [\frac{\delta\xi}{\delta\xi + (1-\delta)\psi}, \frac{(1-\delta)\psi}{\delta\xi + (1-\delta)\psi}].$$

Finally, $1/h = \tilde{\boldsymbol{\pi}}\boldsymbol{C}^1\mathbf{1} = \frac{\xi\psi}{\delta\xi+(1-\delta)\psi}$ is the arrival rate of the MAP.

A.3 Proofs

Lemma A.1. *For an $M/G_n/c/c$ queueing model, the mean travel time W is increasing in arrival rate $\lambda \geq 0$.*

Proof. By (1.12) and (1.13), we can state the mean travel time as follows:

$$\begin{aligned} W = \frac{\sum_{n=0}^{c} n\pi_n}{\lambda(1-\pi_c)} &= \frac{\sum_{n=1}^{c}\{[n(L\lambda)^n/(n!V_n\cdots V_1)]/[1+\sum_{i=1}^{c}(L\lambda)^i/(i!V_i\cdots V_1)]\}}{\lambda\{1-[(L\lambda)^c/(c!V_c\cdots V_1)]/[1+\sum_{i=1}^{c}(L\lambda)^i/(i!V_i\cdots V_1)]\}} \\ &= \frac{L\sum_{n=0}^{c-1}(L\lambda)^n/(n!V_{n+1}V_n\cdots V_1)}{1+\sum_{n=1}^{c-1}(L\lambda)^n/(n!V_n\cdots V_1)}. \end{aligned} \tag{A.2}$$

Taking the derivative with respect to λ, we get the following:

$$\begin{aligned} W' &= \frac{\partial W}{\partial \lambda} \\ &= L^2 \frac{\{\sum_{n=0}^{c-2}(L\lambda)^n/[n!V_{n+2}V_{n+1}\cdots V_1]\}\{1+\sum_{n=1}^{c-1}(L\lambda)^n/[n!V_n\cdots V_1]\} - \{\sum_{n=0}^{c-2}(L\lambda)^n/[n!V_{n+1}V_n\cdots V_1]\}\{\sum_{n=0}^{c-1}(L\lambda)^n/[n!V_{n+1}V_n\cdots V_1]\}}{\{1+\sum_{n=1}^{c-1}(L\lambda)^n/[n!V_n\cdots V_1]\}^2}. \end{aligned}$$

Let $\alpha_0 = 1$ and $\alpha_n = (L\lambda)^n/[n!V_n\cdots V_1]$ for $n = 1, 2, \cdots, c$, then we have:

$$W' = L^2 \frac{[\sum_{n=0}^{c-2}\alpha_n/\big(V_{n+2}V_{n+1}\big)]\big(1+\sum_{n=1}^{c-1}\alpha_n\big) - \big(\sum_{n=0}^{c-2}\alpha_n/V_{n+1}\big)\big(\sum_{n=0}^{c-1}\alpha_n/V_{n+1}\big)}{\big(1+\sum_{n=1}^{c-1}\alpha_n\big)^2}.$$

Using induction, we show W is increasing in λ for $c = 1, 2, 3, \cdots$. For $c = 1$, $W = 1/V_1$ is weakly increasing in λ (it is independent of λ). As the induction hypothesis, we assume $W' \geq 0$ when the

capacity of the system is c. Since the denominator of W' is positive, we have:

$$[\sum_{n=0}^{c-2} \alpha_n/\big(V_{n+2}V_{n+1}\big)]\big(1+\sum_{n=1}^{c-1}\alpha_n\big) - \big(\sum_{n=0}^{c-2}\alpha_n/V_{n+1}\big)\big(\sum_{n=0}^{c-1}\alpha_n/V_{n+1}\big) \geq 0.$$

Assume the capacity of the system is $c+1$, then W' is as follows:

$$\begin{aligned} W' &= \frac{[\sum_{n=0}^{c-1} \alpha_n/(V_{n+2}V_{n+1})](1+\sum_{n=1}^{c}\alpha_n) - (\sum_{n=0}^{c-1}\alpha_n/V_{n+1})(\sum_{n=0}^{c}\alpha_n/V_{n+1})}{(1+\sum_{n=1}^{c}\alpha_n)^2} \\ &= \frac{[\sum_{n=0}^{c-2} \alpha_n/(V_{n+2}V_{n+1})](1+\sum_{n=1}^{c-1}\alpha_n) - (\sum_{n=0}^{c-2}\alpha_n/V_{n+1})(\sum_{n=0}^{c-1}\alpha_n/V_{n+1}) + \frac{\alpha_{c-1}}{V_c}(\frac{1}{V_{c+1}}-\frac{1}{V_1}) + \sum_{n=0}^{c-2}\frac{\lambda\alpha_{c-1}\alpha_n}{V_cV_{n+1}}[\frac{1}{n+1}-\frac{1}{c}]}{(1+\sum_{n=1}^{c}\alpha_n)^2}. \end{aligned}$$

Since $\frac{1}{n+1}-\frac{1}{c} \geq 0$, $\frac{1}{V_{c+1}}-\frac{1}{V_1} > 0$, and by the induction hypothesis, the numerator of W' is positive. The denominator is also positive. Thus, $W' \geq 0$, when the capacity of the system is $c+1$.

By the principle of mathematical induction W is increasing in $\lambda \geq 0$. □

Lemma A.2. *For an $M/G_n/c/c$ queueing model, the throughput θ is an increasing function of the arrival rate λ, if at least one of the following conditions holds:*

$$\lambda \geq \frac{V_1(c+1-1/\pi_c) - c\pi_c(V_1-V_c)}{L(1-\pi_c)} \tag{A.3}$$

or

$$nV_n \text{ is increasing in } n. \tag{A.4}$$

Proof. proof We first prove the sufficiency of condition (A.3) as follows.

By 1.12, θ for an $M/G_n/c/c$ queueing model can be expressed as:

$$\theta = \lambda(1-\pi_c) = \lambda\left[1 - \frac{(L\lambda)^c/(c!V_c\cdots V_1)}{1+\sum_{n=1}^{c}(L\lambda)^n/(n!V_n\cdots V_1)}\right] = \lambda\left[\frac{1+\sum_{n=1}^{c-1}(L\lambda)^n/(n!V_n\cdots V_1)}{1+\sum_{n=1}^{c}(L\lambda)^n/(n!V_n\cdots V_1)}\right].$$

Letting $\alpha_0 = 1$ and $\alpha_n = (L\lambda)^n/[n!V_n \cdots V_1]$ for $n = 1, 2, \cdots, c$, we get $\theta = \dfrac{\lambda + \lambda \sum_{n=1}^{c-1} \alpha_n}{1 + \sum_{n=1}^{c} \alpha_n}$, and $\pi_c = \alpha_c/(1 + \sum_{n=1}^{c} \alpha_n)$. The derivative of θ with respect to λ is as follows:

$$\theta' = \frac{[1 + \sum_{n=1}^{c-1}(n+1)\alpha_n]\big(1 + \sum_{n=1}^{c} \alpha_n\big) - \big(1 + \sum_{n=1}^{c-1} \alpha_n\big)\big(\sum_{n=1}^{c} n\alpha_n\big)}{\big(1 + \sum_{n=1}^{c} \alpha_n\big)^2} = \frac{\alpha_c[c - (c+1)/\pi_c + 1/\pi_c^2] + \sum_{n=1}^{c-1} n\alpha_n}{\alpha_c/\pi_c^2}. \tag{A.5}$$

In the numerator of the above expression, $c - (c+1)/\pi_c + 1/\pi_c^2$ is non-negative, if $0 \le \pi_c \le \frac{1}{c}$ or $\pi_c = 1$. For $\frac{1}{c} < \pi_c < 1$, we restate the numerator as follows:

$$\begin{aligned}\alpha_c[c - (c+1)/\pi_c + 1/\pi_c^2] + \sum_{n=1}^{c-1} n\alpha_n &\ge \alpha_c[c - (c+1)/\pi_c + 1/\pi_c^2] + \sum_{n=0}^{c-2} \frac{L\lambda}{V_1}\alpha_n \\ &= \alpha_c\Big[c - (c+1)/\pi_c + 1/\pi_c^2 + \frac{L\lambda}{V_1}(\frac{1}{\pi_c} - 1) - \frac{cV_c}{V_1}\Big].\end{aligned}$$

The above expression is positive if $\lambda \ge \dfrac{V_1(c + 1 - 1/\pi_c) - c\pi_c(V_1 - V_c)}{L(1 - \pi_c)}$. Thus, $\theta' \ge 0$, when $\lambda \ge \dfrac{V_1(c + 1 - 1/\pi_c) - c\pi_c(V_1 - V_c)}{L(1 - \pi_c)}$.

Next, we use induction to prove the sufficiency of condition (A.4) for $c = 1, 2, 3, \cdots$. For $c = 1$, $\pi_0 = 1 - \pi_1 = \frac{V_1}{L\lambda + V_1}$, so $\theta = \frac{\lambda V_1}{L\lambda + V_1}$ and $\theta' = \frac{V_1^2}{(L\lambda + V_1)^2} \ge 0$. Thus θ is increasing in λ. Suppose $\theta' \ge 0$ when the capacity of the system is c and $cV_c \ge (c-1)V_{c-1}$. Since in (A.5) the denominator of θ' is positive, we have:

$$[1 + \sum_{n=1}^{c-1}(n+1)\alpha_n]\big(1 + \sum_{n=1}^{c} \alpha_n\big) - \big(1 + \sum_{n=1}^{c-1} \alpha_n\big)\big(\sum_{n=1}^{c} n\alpha_n\big) \ge 0.$$

Assume the capacity of the system is $c+1$, then θ' is as follows:

$$\theta' = \frac{[1 + \sum_{n=1}^{c-1}(n+1)\alpha_n](1 + \sum_{n=1}^{c} \alpha_n) - (1 + \sum_{n=1}^{c-1} \alpha_n)(\sum_{n=1}^{c} n\alpha_n)}{(1 + \sum_{n=1}^{c+1} \alpha_n)^2} + \frac{(c-n)\sum_{n=1}^{c-1}\Big[\alpha_c\alpha_{n+1} - \alpha_{c+1}\alpha_n\Big] + c(\alpha_c\alpha_1 - \alpha_{c+1}) + (c+1)\alpha_c}{(1 + \sum_{n=1}^{c+1} \alpha_n)^2}$$

Due to the induction hypothesis, the first fraction is positive. To show that the second fraction is positive, it suffices to show that each $\alpha_c\alpha_{n+1} - \alpha_{c+1}\alpha_n$ for $n = 0, 1, \cdots, c-1$ is positive. since the procedure is similar, we show this only for $n = c-1$.

$$\alpha_c\alpha_c - \alpha_{c+1}\alpha_{c-1} = \frac{\lambda^{2c}}{(c!V_c\cdots V_1)\big[(c-1)!V_{c-1}\cdots V_1\big]}\Big(\frac{1}{cV_c} - \frac{1}{(c+1)V_{c+1}}\Big).$$

This expression is positive, if $cV_c \le (c+1)V_{c+1}$.

By the principle of mathematical induction θ is increasing in λ, if nV_n is increasing in n. □

Proof of Proposition 1.1. a) By (A.2), for an $M/G_n/c/c$ queueing system, the mean travel time is calculated as follows:

$$W = \frac{\sum_{n=0}^{c} n\pi_n}{\lambda(1-\pi_c)} = \frac{L\sum_{n=0}^{c-1}(L\lambda)^n/[n!V_{n+1}V_n\cdots V_1]}{1+\sum_{n=1}^{c-1}(L\lambda)^n/[n!V_n\cdots V_1]}.$$

Taking the limit of W as the arrival rate λ approaches 0, we get the following:

$$W_0 = \lim_{\lambda\to 0} W = \lim_{\lambda\to 0}\frac{L\sum_{n=0}^{c-1}(L\lambda)^n/[n!V_{n+1}V_n\cdots V_1]}{1+\sum_{n=1}^{c-1}(L\lambda)^n/[n!V_n\cdots V_1]} = L/V_1.$$

In terms of our queueing models, we have $W_0^B = L/V_1^B$, and $W_0^D(p) = pW_0^{DA} + (1-p)W_0^{DH} = (pL)/V_1^{DA} + [(1-p)L]/V_1^{DH}$. Since, $V_1^{DA} = V_1^B$ and $V_1^{DH} \le V_1^B$, $W_0^D(p) \ge W_0^B$. Moreover, according to Lemma A.1, W^B and $W^D(p)$ are both increasing in λ. Thus, for every p there exists a $\underline{\lambda}^{(D,W)}(p)$ such that for $\lambda \le \underline{\lambda}^{(D,W)}(p)$, $W_\infty^D(p) \ge W_\infty^B$. Let $\underline{\lambda}^{(D,W)} = \min_p \underline{\lambda}^{(D,W)}(p)$, then when $\lambda \le \underline{\lambda}^{(D,W)}$, $W^D(p) \ge W^B$ for $p \in [0,1]$.

Taking the limit of W as the arrival rate λ approaches ∞, we get the following expression:

$$W_\infty = \lim_{\lambda\to\infty} W = \lim_{\lambda\to\infty}\frac{L\sum_{n=0}^{c-1}(L\lambda)^n/[n!V_{n+1}V_n\cdots V_1]}{1+\sum_{n=1}^{c-1}(L\lambda)^n/[n!V_n\cdots V_1]} = L/V_c. \tag{A.6}$$

In terms of our queueing models, we have $W_\infty^B = L/V_{Nc}^B$, and $W_\infty^D(p) = pW_\infty^{DA} + (1-p)W_\infty^{DH} = (pL)/V_c^{DA} + [(1-p)L]/V_{(N-1)c}^{DH}$. As a result, $W_\infty^D(p) \le W_\infty^B$ if and only if $p \ge p^{(D,W)} = \frac{V_c^{DA}V_{Nc}^B - V_c^{DA}V_{(N-1)c}^{DH}}{V_{Nc}^B V_c^{DA} - V_{Nc}^B V_{(N-1)c}^{DH}}$ (since $V_{(N-1)c}^{DH} - V_c^{DA} \le 0$ the direction of inequality changes). Moreover, according to Lemma A.1, W^B and $W^D(p)$ are both increasing in λ. Thus, for a given $p \ge p^{(D,W)}$, there exists a $\bar{\lambda}^{(D,W)}(p)$ such that when $\lambda \ge \bar{\lambda}^{(D,W)}(p)$, $W_\infty^D(p) \le W_\infty^B$. Let $\bar{\lambda}^{(D,W)} = \max_p \bar{\lambda}^{(D,W)}(p)$, then when $\lambda \ge \bar{\lambda}^{(D,W)}$,

$W^D_\infty(p) \geq W^B_\infty$ for $p \geq p^{(D,W)}$. It is immediately clear that, $\underline{\lambda}^{(D,W)} \leq \bar{\lambda}^{(D,W)}$.

b) For an $M/G_n/c/c$ queueing system, the throughput θ is equal to $\lambda(1-\pi_c)$; in terms of our queueing models we have: $\theta^B = \lambda(1-\pi^B_{Nc})$ and $\theta^D = \theta^{DA} + \theta^{DH} = p\lambda(1-\pi^{DA}_c) + (1-p)\lambda(1-\pi^{DH}_{(N-1)c})$. Thus, $\theta^D(p) \geq \theta^B$ if and only if $p\pi^{DA}_c + (1-p)\pi^{DH}_{(N-1)c} \leq \pi^B_{Nc}$. Taking the limit of θ for any $M/G_n/c/c$ queue as the arrival rate λ approaches ∞, we get the following:

$$\theta_\infty = \lim_{\lambda\to\infty}\theta = \lim_{\lambda\to\infty}\lambda(1-\pi_c) = \lim_{\lambda\to\infty}\lambda\left[1 - \frac{(L\lambda)^c/(c!V_c\cdots V_1)}{1+\sum_{n=1}^{c}(L\lambda)^n/(n!V_n\cdots V_1)}\right]$$
$$= \lim_{\lambda\to\infty}\left\{\frac{L^{c-1}\lambda^c/[(c-1)!V_{c-1}\cdots V_1]}{(L\lambda)^c/(c!V_c\cdots V_1)}\right\} = (cV_c)/L = NJV_c. \quad \text{(A.7)}$$

Suppose $\lambda \geq \lambda^{(D,\theta)} = \min_p \max\{\lambda^{DA}_{jam}/p, \lambda^{DH}_{jam}/(1-p)\}$. According to Lemma A.2 and our assumption in the statement of the proposition that θ^{DA} and θ^{DH} are increasing in λ, θ^{DA} and θ^{DH} are increasing and decreasing in p, respectively. Let p^* be the smallest value of the AV proportion such that $p^*\lambda \geq \lambda^{DA}_{jam}$ and $(1-p^*)\lambda \geq \lambda^{DH}_{jam}$. For $p \geq p^*$ (resp., $p \leq p^*$), θ^{DA} (resp., θ^{DH}) is equal to its jam value, i.e., $\theta^{DA}_\infty = cV^{DA}_c/L$ (resp., $\theta^{DH}_\infty = (N-1)cV^{DH}_{(N-1)c}/L$). Since $V^{DA}_c + (N-1)cV^{DH}_{(N-1)c} \geq NcV^B_{Nc}$, there exists $\underline{p}^{(D,\theta)} \leq p^*$, such that $(1-\pi^{DA}_c)\underline{p}^{(D,\theta)}\lambda + (N-1)cV^{DH}_{(N-1)c}/L = NcV^B_{Nc}/L$, and there exists $\bar{p}^{(D,\theta)} \geq p^*$, such that $(1-\pi^{DH}_{(N-1)c})(1-\bar{p}^{(D,\theta)})\lambda + cV^{DA}_c/L = NcV^B_{Nc}/L$. □

Proof of Corollary 1.1. a) We want $\underline{p}^{(D,\theta)} \leq p^{(D,W)}$. Substituting $\underline{p}^{(D,\theta)} = \frac{NcV^B_{Nc} - (N-1)cV^{DH}_{(N-1)c}}{L\lambda(1-\pi^{DA}_c)}$, and $p^{(D,W)} = \dfrac{V^{DA}_c V^B_{Nc} - V^{DA}_c V^{DH}_{(N-1)c}}{V^{DA}_c V^B_{Nc} - V^B_{Nc} V^{DH}_{(N-1)c}}$, we get $\lambda \geq \bar{\lambda}^{(D,W\theta)} = \frac{V^B_{Nc}(V^{DA}_c - V^{DH}_{(N-1)c})[NcV^B_{Nc} - (N-1)cV^{DH}_{(N-1)c}]}{LV^{DA}_c(V^B_{Nc} - V^{DH}_{(N-1)c})(1-\pi^{DA}_c)}$. According to Proposition 1.1, $p^{(D,W)}$ and $\underline{p}^{(D,\theta)}$ exist when $\lambda \geq \bar{\lambda}^{(D,W)}$ and $\lambda \geq \lambda^{(D,\theta)}$, respectively, $\underline{p}^{(D,\theta)} \leq p^{(D,W)}$ if $\lambda \geq \max\{\bar{\lambda}^{(D,W\theta)}, \bar{\lambda}^{(D,W)}, \lambda^{(D,\theta)}\}$.

b) $\bar{p}^{(D,\theta)} \geq (1-\pi^{DH}_{(N-1)c})\bar{p}^{(D,\theta)} = (1-\pi^{DH}_{(N-1)c}) - (\frac{NcV^B_{Nc} - cV^{DA}_c}{L\lambda})$, and $p^{(D,W)} = \dfrac{V^{DA}_c V^B_{Nc} - V^{DA}_c V^{DH}_{(N-1)c}}{V^{DA}_c V^B_{Nc} - V^B_{Nc} V^{DH}_{(N-1)c}}$, so $\bar{p}^{(D,\theta)} \geq p^{(D,W)}$ if $\lambda \geq \underline{\lambda}^{(D,W\theta)} = \frac{V^B_{Nc}(V^{DA}_c - V^{DH}_{(N-1)c})[NcV^B_{Nc} - cV^{DA}_c]}{L\{\pi^{DH}_{(N-1)c}V^B_{Nc}(V^{DA}_c - V^{DH}_{(N-1)c}) + V^{DH}_{(N-1)c}(V^{DA}_c - V^B_{Nc})\}}$. According to Proposition 1.1, $p^{(D,W)}$ and $\bar{p}^{(D,\theta)}$ exist when $\lambda \geq \bar{\lambda}^{(D,W)}$ and $\lambda \geq \lambda^{(D,\theta)}$, respectively, $\bar{p}^{(D,\theta)} \geq p^{(D,W)}$ if $\lambda \geq \max\{\underline{\lambda}^{(D,W\theta)}, \bar{\lambda}^{(D,W)}, \lambda^{(D,\theta)}\}$. □

Proof of Proposition 1.2. a) As mentioned in Proof of Proposition 1.1, $W_\infty = L/V_c$, so $W^I_\infty(p) = L/V^I_{Nc}(p) \leq W^B_\infty = L/V^B_{Nc}$, if and only if $V^I_{Nc}(p) \geq V^B_{Nc}$. Moreover, according to Lemma A.1, W^B and $W^I(p)$ are both increasing in λ. Thus, there exists a $\lambda^{(I,W)}$ such that for $\lambda \geq \lambda^{(I,W)}$, $W^I_\infty(p) \leq W^B_\infty$ if and only if $V^I_{Nc}(p) \geq V^B_{Nc}$.

Proof of (1.14): By (1.2), $V^I_{Nc}(p) \geq V^B_{Nc}$ if and only if $h^I_{Nc}(p) \leq h^B_{Nc}$. Substituting h^B_{Nc} with $\frac{\delta^B}{\psi^{IHH}_{Nc}} + \frac{1-\delta^B}{\xi^{IHH}_{Nc}}$ and $h^I_{Nc}(p)$ with $\delta^I\big(\frac{(1-p)^2}{\psi^{IHH}_{Nc}} + \frac{p(1-p)}{\psi^{IHA}_{Nc}} + \frac{p(1-p)}{\psi^{IAH}_{Nc}} + \frac{p^2}{\psi^{IAA}_{Nc}}\big) + (1-\delta^I)\big(\frac{(1-p)^2}{\xi^{IHH}_{Nc}} + \frac{p(1-p)}{\xi^{IHA}_{Nc}} + \frac{p(1-p)}{\xi^{IAH}_{Nc}} + \frac{p^2}{\xi^{IAA}_{Nc}}\big)$, we get the following:

$$\begin{aligned}\frac{\delta^B}{\psi^{IHH}_{Nc}} + \frac{1-\delta^B}{\xi^{IHH}_{Nc}} &\geq \delta^I\big(\frac{(1-p)^2}{\psi^{IHH}_{Nc}} + \frac{p(1-p)}{\psi^{IHA}_{Nc}} + \frac{p(1-p)}{\psi^{IAH}_{Nc}} + \frac{p^2}{\psi^{IAA}_{Nc}}\big) + (1-\delta^I)\big(\frac{(1-p)^2}{\xi^{IHH}_{Nc}} + \frac{p(1-p)}{\xi^{IHA}_{Nc}} + \frac{p(1-p)}{\xi^{IAH}_{Nc}} + \frac{p^2}{\xi^{IAA}_{Nc}}\big)\\ &= p\delta^I\big(\frac{p-2}{\psi^{IHH}_{Nc}} + \frac{(1-p)}{\psi^{IHA}_{Nc}} + \frac{(1-p)}{\psi^{IAH}_{Nc}} + \frac{p}{\psi^{IAA}_{Nc}}\big) + \frac{\delta^I}{\psi^{IHH}_{Nc}} + p(1-\delta^I)\big(\frac{p-2}{\xi^{IHH}_{Nc}} + \frac{(1-p)}{\xi^{IHA}_{Nc}} + \frac{(1-p)}{\xi^{IAH}_{Nc}} + \frac{p}{\xi^{IAA}_{Nc}}\big) + \frac{1-\delta^I}{\xi^{IHH}_{Nc}}\end{aligned}$$

Rearranging the inequality, we have the following:

$$\begin{aligned}p\delta^I\big(\frac{2-p}{\psi^{IHH}_{Nc}} - \frac{(1-p)}{\psi^{IHA}_{Nc}} - \frac{(1-p)}{\psi^{IAH}_{Nc}} - \frac{p}{\psi^{IAA}_{Nc}}\big) + p(1-\delta^I)\big(\frac{2-p}{\xi^{IHH}_{Nc}} - \frac{(1-p)}{\xi^{IHA}_{Nc}} - \frac{(1-p)}{\xi^{IAH}_{Nc}} - \frac{p}{\xi^{IAA}_{Nc}}\big) &\geq \frac{\delta^I-\delta^B}{\psi^{IHH}_{Nc}} + \frac{1-\delta^I-1+\delta^B}{\xi^{IHH}_{Nc}}\\ &= \frac{p(\delta^{IAA}-\delta^{IHH})}{\psi^{IHH}_{Nc}} + \frac{p(-\delta^{IAA}+\delta^{IHH})}{\xi^{IHH}_{Nc}}\end{aligned}$$

Canceling p's out, and rearranging the right hand side, we get the following expression:

$$\delta^I\big(\frac{2-p}{\psi^{IHH}_{Nc}} - \frac{1-p}{\psi^{IHA}_{Nc}} - \frac{1-p}{\psi^{IAH}_{Nc}} - \frac{p}{\psi^{IAA}_{Nc}}\big) + (1-\delta^I)\big(\frac{2-p}{\xi^{IHH}_{Nc}} - \frac{1-p}{\xi^{IHA}_{Nc}} - \frac{1-p}{\xi^{IAH}_{Nc}} - \frac{p}{\xi^{IAA}_{Nc}}\big) \geq \big(\frac{\delta^{IAA}}{\psi^{IHH}_{Nc}} + \frac{1-\delta^{IAA}}{\xi^{IHH}_{Nc}}\big) - \big(\frac{\delta^{IHH}}{\psi^{IHH}_{Nc}} + \frac{1-\delta^{IHH}}{\xi^{IHH}_{Nc}}\big).$$

b) As mentioned in Proof of Proposition 1.1, $\theta_\infty = cV_c/L$, so $\theta^I_\infty(p) = NcV^I_{Nc}(p)/L \geq \theta^B_\infty = NcV^B_{Nc}/L$, if and only if $V^I_{Nc}(p) \geq V^B_{Nc}$. Moreover, according to Lemma A.2, θ^B and $\theta^I(p)$ are both increasing in λ. Thus, there exists a $\lambda^{(I,\theta)}$ such that for $\lambda \geq \lambda^{(I,\theta)}$, $\theta^I_\infty(p) \geq \theta^B_\infty$ if and only if $V^I_{Nc}(p) \geq V^B_{Nc}$. □

Proof of Proposition 1.3. In this proof we assume that the conditions stated in Propositions 1.1 and 1.2 for the existence of thresholds of p hold. Next, we discuss each column of the table as follows.

Column 1: As p approaches zero, $V^I_n(p) \to V^B_n$, so $W^I(p) \to W^B$ and $\theta^I(p) \to \theta^B$. Moreover, $\lim_{p\to 0} W^D(p) = \lim_{p\to 0} pW^{DA} + (1-p)W^{DH} = W^{DH} = L/V^{DH}_{(N-1)c} \geq L/V^B_{Nc} = W^B$, and $\lim_{p\to 0}\theta^D(p) = \lim_{p\to 0} p\theta^{DA} + (1-p)\theta^{DH} = \theta^{DH} = (N-1)cV^{DH}_{(N-1)c}/L \leq NcV^B_{Nc}/L = \theta^B$.

Column 2: As p approaches one, $V^I_{Nc}(p) \to V^{DA}_c$, so $W^I(p) \to W^D$. Also, since $V^I_{Nc}(p)$

is increasing in p, $W^D(p) = L/V^I_{Nc}(1) \leq W^B = L/V^B_{Nc}$. Similarly, $\theta^I(p) = NcV^I_{Nc}(1)/L \geq \theta^B = NcV^B_{Nc}/L$. However, $\theta^I(p) \to NcV^I_{Nc}(p)/L \geq cV^I_{Nc}(p)/L = cV^{DA}_c/L = \theta^{DA} = \lim_{p\to 1}\theta^D$. Moreover, $\lim_{p\to 1}\theta^D = \theta^{DA} = cV^{DA}_c/L \geq NcV^B_{Nc}/L = \theta^B$ if and only if $\frac{V^{DA}_c}{cV^B_{Nc}} \geq N$.

Column 3: We first analyze W. Because we assume that $V^I_{Nc}(p)$ is increasing in p, $W^I(p) = L/V^I_{Nc}(p) \leq L/V^I_{Nc}(0) = L/V^B_{Nc} = W^B$. For a given $p \in [0,1]$, $W^D(p) = \frac{pL}{V^{DA}_c} + \frac{(1-p)L}{V^{DH}_{(N-1c)}} \leq L/V^I_{Nc}(p) = W^I(p)$, if and only if $-\frac{V^{DA}_c V^{DH}_{(N-1)c}}{V^{DA}_c - V^{DH}_{(N-1)c}} \leq V^I_{Nc}(p)\left(p - \frac{V^{DA}_c}{V^{DA}_c - V^{DH}_{(N-1)c}}\right)$; derived by rearranging the first inequality. Note that in this inequality $V^{DA}_c - V^{DH}_{(N-1)c} \geq 0$, because $V^{DA}_c = V^I_{Nc}(1) \geq V^B_{Nc} \geq V^{DH}_{(N-1)c}$. The right hand side of this inequality, i.e., $V^I_{Nc}(p)\left(p - \frac{V^{DA}_c}{V^{DA}_c - V^{DH}_{(N-1)c}}\right)$, is increasing in p. Thus, $W^D(p) \leq W^I(p)$, if and only if $p \geq p^{(DI,W)}$, where $p^{(DI,W)}$ is the smallest p such that $-\frac{V^{DA}_c V^{DH}_{(N-1)c}}{V^{DA}_c - V^{DH}_{(N-1)c}} \leq V^I_{Nc}(p)\left(p - \frac{V^{DA}_c}{V^{DA}_c - V^{DH}_{(N-1)c}}\right)$.

We need to show that $p^{(DI,W)} \geq p^{(D,W)}$. If $p \leq p^{(D,W)}$, $W^D(p) \geq W^B$, i.e., $\frac{pL}{V^{DA}_c} + \frac{(1-p)L}{V^{DH}_{(N-1c)}} \geq L/V^B_{Nc} \geq L/V^I_{Nc}(p) = W^I(p)$. The last inequality holds because we assume that $V^I_{Nc}(p)$ is increasing in p, and $V^I_n(0) = V^B_n$. Thus, for $p \leq p^{(D,W)}$, $W^I(p) \leq W^D(p)$, and $p^{(DI,W)} \geq p^{(D,W)}$.

Next, we analyze θ similar to the proof of Proposition 1.1. Three possible scenarios can occur. First, if there is no $p \in [0,1]$ that satisfies $p\lambda(1-\pi^{DA}_c) + (1-p)\lambda(1-\pi^{DH}_{(N-1)c}) = NcV^I_{Nc}(p)/L$, then $\theta^I(p) > \theta^D(p)$, and $\underline{p}^{(DI,\theta)}$ and $\bar{p}^{(DI,\theta)}$ do not exist. Second, if $p\lambda(1-\pi^{DA}_c) + (N-1)cV^{DH}_{(N-1)c}/L = NcV^I_{Nc}(p)/L$ has two roots, then $\underline{p}^{(DI,\theta)}$ (resp., $\bar{p}^{(DI,\theta)}$) is the smallest (resp., largest) p that satisfies this equality. Lastly, if $p\lambda(1-\pi^{DA}_c) + (N-1)cV^{DH}_{(N-1)c}/L = NcV^I_{Nc}(p)/L$ has a unique root (i.e., $\underline{p}^{(DI,\theta)}$), then $cV^{DA}_c/L + (1-p)\lambda(1-\pi^{DH}_{(N-1)c}) \geq NcV^I_{Nc}(p)/L = \theta^I(p)$ has a unique root ($\bar{p}^{(DI,\theta)}$) as well. In this case, $\underline{p}^{(DI,\theta)} \leq \bar{p}^{(DI,\theta)}$ holds, because at $\bar{p}^{(DI,\theta)}$ the AV queue is jammed, so this value of p is higher than the value of p at which the AV queue is not jammed, i.e., $\underline{p}^{(DI,\theta)}$.

Since we assume that $V^I_{Nc}(p)$ is increasing in p and it is concave (or convex) everywhere, $\theta^I(p)$ is also increasing in p, and $\theta^D(p) \geq \theta^I(p)$, if and only if $\underline{p}^{(DI,\theta)} \leq p \leq \bar{p}^{(DI,\theta)}$. Lastly, $\underline{p}^{(D,\theta)} \leq \underline{p}^{(DI,\theta)} \leq \bar{p}^{(DI,\theta)} \leq \bar{p}^{(D,\theta)}$, because $V^I_{Nc}(p) \geq V^B_{Nc}$.

Remark. The conditions on V can be expressed in terms of parameters as follows: $V^I_{Nc}(p)$ is increasing in p if $\frac{\partial[\delta^I(1/\xi^I_{Nc} + 1/\psi^I_{Nc}) + 1/\xi^I_{Nc}]}{\partial p} \geq 0$, and $V^I_{Nc}(p)$ is concave everywhere or convex everywhere if $[\delta^I(1/\xi^I_{Nc}+1/\psi^I_{Nc})+1/\xi^I_{Nc}]\frac{\partial^2[\delta^I(1/\xi^I_{Nc} + 1/\psi^I_{Nc}) + 1/\xi^I_{Nc}]}{\partial p^2} - 2\frac{\partial[\delta^I(1/\xi^I_{Nc} + 1/\psi^I_{Nc}) + 1/\xi^I_{Nc}]}{\partial p}$

≤ 0 or $[\delta^I(1/\xi^I_{Nc}+1/\psi^I_{Nc})+1/\xi^I_{Nc}]\dfrac{\partial^2[\delta^I(1/\xi^I_{Nc}+1/\psi^I_{Nc})+1/\xi^I_{Nc}]}{\partial p^2}-2\dfrac{\partial[\delta^I(1/\xi^I_{Nc}+1/\psi^I_{Nc})+1/\xi^I_{Nc}]}{\partial p}\geq 0$ for all $p\in[0,1]$. □

Proof of Corollary 1.2. As we discussed in the Proof of Proposition 1.3, $\underline{p}^{(DI,\theta)}\leq\bar{p}^{(DI,\theta)}$. To show that $\underline{p}^{(DI,\theta)}\leq p^{(DI,W)}$, we consider two case: $p^{(DI,W)}\leq\dfrac{V_c^{DA}}{V_c^{DA}+(N-1)V^{DH}_{(N-1)c}}$ and $p^{(DI,W)}\geq\dfrac{V_c^{DA}}{V_c^{DA}+(N-1)V^{DH}_{(N-1)c}}$.

Case 1 ($p^{(DI,W)}\leq\dfrac{V_c^{DA}}{V_c^{DA}+(N-1)V^{DH}_{(N-1)c}}$): Suppose $p\leq p^{(DI,W)}$, i.e., $L/V^I_{Nc}(p)\leq pL/V_c^{DA}+(1-p)L/V^{DH}_{(N-1)c}$. Let $\lambda p(1-\pi_c^{DA})=c\tilde{V}_c^{DA}/L$. By definition $p\leq\underline{p}^{(DI,\theta)}$, if $\dfrac{c\tilde{V}_c^{DA}/L+(N-1)cV^{DH}_{(N-1)c}/L}{Nc}\leq V^I_{NC}(p)/L$, or equivalently $\dfrac{LNc}{c\tilde{V}_c^{DA}+(N-1)cV^{DH}_{(N-1)c}}\geq L/V^I_{NC}(p)$. Thus, if $\dfrac{LNc}{c\tilde{V}_c^{DA}+(N-1)cV^{DH}_{(N-1)c}}\geq pL/V_c^{DA}+(1-p)L/V^{DH}_{(N-1)c}$, then $p\leq\underline{p}^{(DI,\theta)}$. Rearranging this inequality we get: if $p\geq\dfrac{(V^{DH}_{(N-1)c}-\tilde{V}_c^{DA})V_c^{DA}}{(V^{DH}_{(N-1)c}-V_c^{DA})[\tilde{V}_c^{DA}+(N-1)V^{DH}_{(N-1)c}]}$, then $p\leq\underline{p}^{(DI,\theta)}$. Note that the numerator of the right hand side of this inequality is positive, because if $V^{DH}_{(N-1)c}\leq\tilde{V}_c^{DA}$, $p\geq\dfrac{V_c^{DA}}{\tilde{V}_c^{DA}+(N-1)V^{DH}_{(N-1)c}}\geq\dfrac{V_c^{DA}}{V_c^{DA}+(N-1)V^{DH}_{(N-1)c}}$, and this is a contradiction. If $V^{DH}_{(N-1)c}\geq\tilde{V}_c^{DA}$, $p\geq 0\geq[(V^{DH}_{(N-1)c}-\tilde{V}_c^{DA})V_c^{DA}]\{(V^{DH}_{(N-1)c}-V_c^{DA})[\tilde{V}_c^{DA}+(N-1)V^{DH}_{(N-1)c}]\}$, so in this case $p\leq\underline{p}^{(DI,\theta)}$. Thus, $\underline{p}^{(DI,\theta)}\leq p^{(DI,W)}$.

Case 2 ($p^{(DI,W)}\geq\dfrac{V_c^{DA}}{V_c^{DA}+(N-1)V^{DH}_{(N-1)c}}$): Suppose $p\leq p^{(DI,W)}$, i.e., $L/V^I_{Nc}(p)\leq pL/V_c^{DA}+(1-p)L/V^{DH}_{(N-1)c}$. In this case, $L/V^I_{Nc}(p)\leq pL/V_c^{DA}+(1-p)L/V^{DH}_{(N-1)c}\leq\dfrac{LN}{V_c^{DA}+(N-1)V^{DH}_{(N-1)c}}$, if and only if $p\geq\dfrac{V_c^{DA}}{V_c^{DA}+(N-1)V^{DH}_{(N-1)c}}$. In other words, if $p\geq\dfrac{V_c^{DA}}{V_c^{DA}+(N-1)V^{DH}_{(N-1)c}}$, $\theta^I(p)$ is higher than the jam throughput of the D policy, i.e., $cV_c^{DA}/L+(N-1)cV^{DH}_{(N-1)c}/L$. Thus $\theta^I(p)\geq\theta^D(p)$ when $p\geq\dfrac{V_c^{DA}}{V_c^{DA}+(N-1)V^{DH}_{(N-1)c}}$. Therefore, $\underline{p}^{(DI,\theta)}$, the smallest value of p at which $\theta^D(p)$ becomes higher than $\theta^I(p)$, cannot be higher than $\dfrac{V_c^{DA}}{V_c^{DA}+(N-1)V^{DH}_{(N-1)c}}$. Moreover, $p^{(DI,W)}\geq\dfrac{V_c^{DA}}{V_c^{DA}+(N-1)V^{DH}_{(N-1)c}}$, thus $\underline{p}^{(DI,\theta)}\leq p^{(DI,W)}$. □

A.4 Parameter Estimation

A.4.1 State-dependent Speed Curve

We follow the transportation literature to fit a function to our speed-volume data from Arizona. As explained in Del Castillo & Benitez (1995) and Jain & Smith (1997), there exist various functions to represent the relationship between speed and volume. Tiwari & Marsani (2014) provide a summary of these functions. We fit all different functions mentioned in Tiwari & Marsani (2014) to our data. Specifically, several linear, logarithmic, polynomial, and exponential functions were tested, and the functional form $y = ae^{-x^2/b} + c$ gave us the highest coefficient of determination which is 85%.

A.4.2 HV Mean Platoon Size

To estimate the mean platoon size for HVs we use data. The data from Arizona Department of Transportation do not include the headway between vehicles, so we use another data set from the Institute for Transportation of Iowa State University. This data set consists of more than $314,000$ instances of headways between vehicles, ranging from milliseconds to hundred of seconds, and is collected in 2015 from several highways in Iowa, including I-74 and I-80. In order to distinguish between interplatoon and intraplatoon headways in this data set, we use the mixtools library in R to divide the headways into two clusters: one for the smaller headway values corresponding to the intraplatoon headways and the other for the larger headway values corresponding to the interplatoon headways. Based on the posterior distribution of the clusters, if a headway value is larger than 2.355 seconds, the probability that it belongs to the intraplatoon headway cluster is less than 10^{-7}. Thus, when the headway between two consecutive vehicles is less than 2.355 seconds, we assume they belong to the same platoon; otherwise they are in two separate ones.[1] Counting the number of consecutive vehicles in the same platoon, we get a sample of platoon size values. Among different discrete distributions, a geometric distribution with parameter 0.667 fits this sample well (see Figure (A.2)). Thus, we set the mean platoon size of HVs equal to $1/0.667 = 1.5$ vehicles.

[1] We are not able to use this data in estimating the mean interplatoon and intraplatoon headways, since these parameters depend on the number of vehicles present on the highway, n, and the data from Iowa State University do not include n.

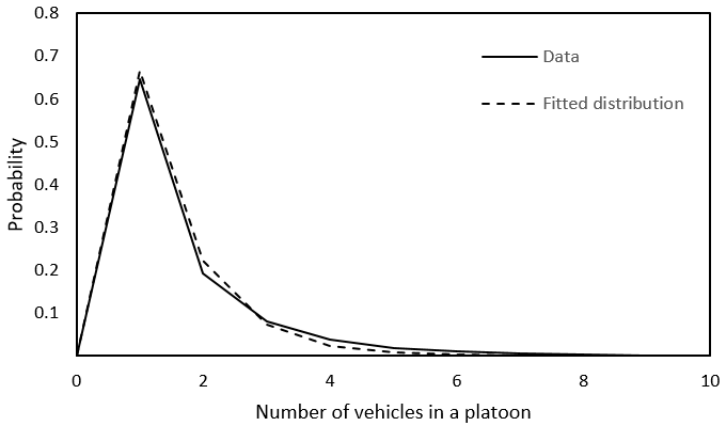


Figure A.2: Platoon size data and the fitted curve with $R^2 = 99.5\%$.

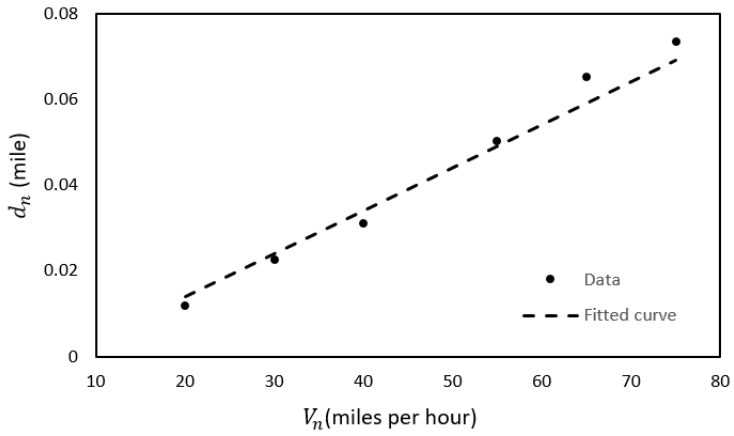


Figure A.3: Safe stopping distance data and the fitted curve with $R^2 = 97.6\%$.

A.4.3 Safe Stopping Time

To estimate the mean interplatoon headway of AVs, we use data from National Highway Traffic Safety Administration (NHTSA 2015). This dataset provides a set of (d_n, V_n) pairs, where d_n is the safe distance (in miles) from the vehicle in front to stop a vehicle when driving at speed V_n (in miles per hour); this applies to both human-driven and autonomous vehicles. Figure A.3 shows the pairs of (d_n, V_n) provided by NHTSA, and the fitted line to this data. The value of coefficient of determination, R^2, for this curve is 97.6%. The linear regression model fitted to these data points is $d_n = 0.001V_n - 0.006$ (miles). Noting that safe stopping time is $\frac{d_n}{V_n}$, we obtain the mean interplatoon headway in the AV queue, $1/\psi_n^{DA}$, in (9).

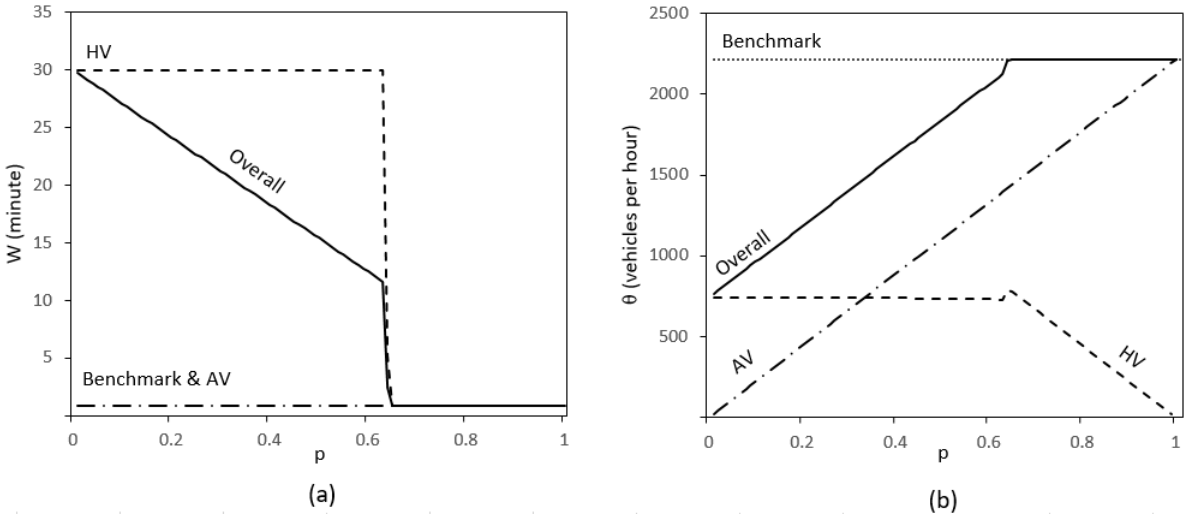


Figure A.4: QoS measures for the D policy when $\lambda = 2,217$ vehicles per hour: (a) mean travel time, and (b) throughput

A.5 Additional Results

A.5.1 Performance of Lightly Loaded Highways with AVs

For a lightly loaded highway, there is no congestion in the benchmark case, and vehicles are able to drive freely on the highway. However, since the D policy divides the highway into two queueing systems, when p is low, the HV queue becomes congested. In this case, the large W of this queue increases the overall W of the D policy beyond W of the benchmark case. As shown in Figure A.4(a), for a lightly loaded highway with $\lambda = 2,217$ vehicles per hour, when $p \leq 0.64$, the mean travel time W under the D policy is higher than in the benchmark case. When $p > 0.64$, this policy works as well, but not better than the benchmark case, because at this value of λ the average speed of vehicles on the highway in the benchmark case is already high, so that adding AVs does not improve the system further. The same argument holds for the throughput θ; see Figure A.4(b).

Figure A.5(a) compares the mean travel time W between the benchmark and those policies with AVs when $\lambda = 2,217$ vehicles per hour. Under the I policy, one can observe from Figure A.5(a) that the maximum improvement in W is only two seconds (0.036 minutes). Also, as shown in Figure A.5(b), at this value of λ the throughput of the benchmark case is equal to the arrival rate which is the maximum achievable throughput, and thus adding AVs does not improve the throughput.

Comparing the D policy with the I policy, Figure A.5 also shows that the latter performs

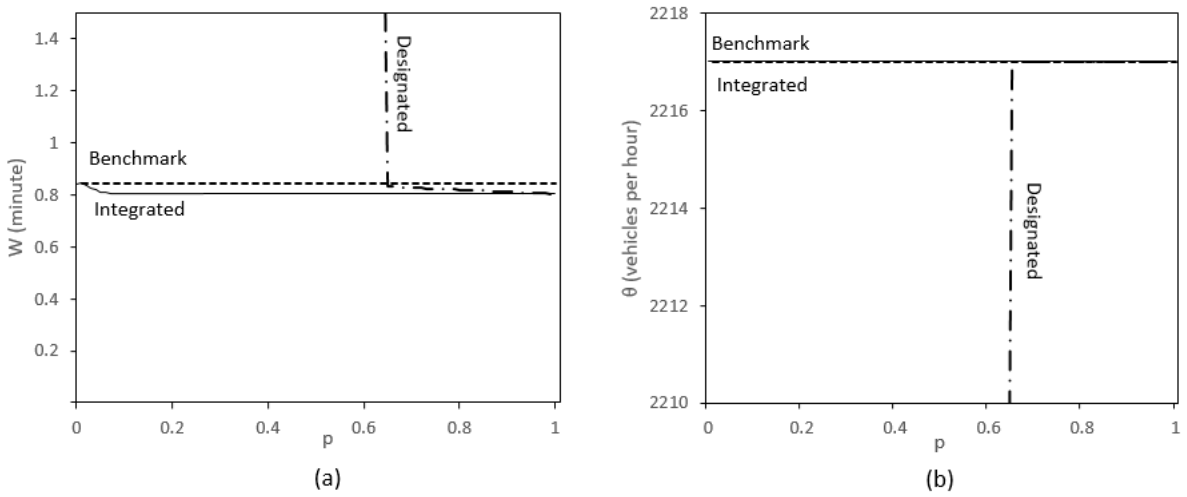


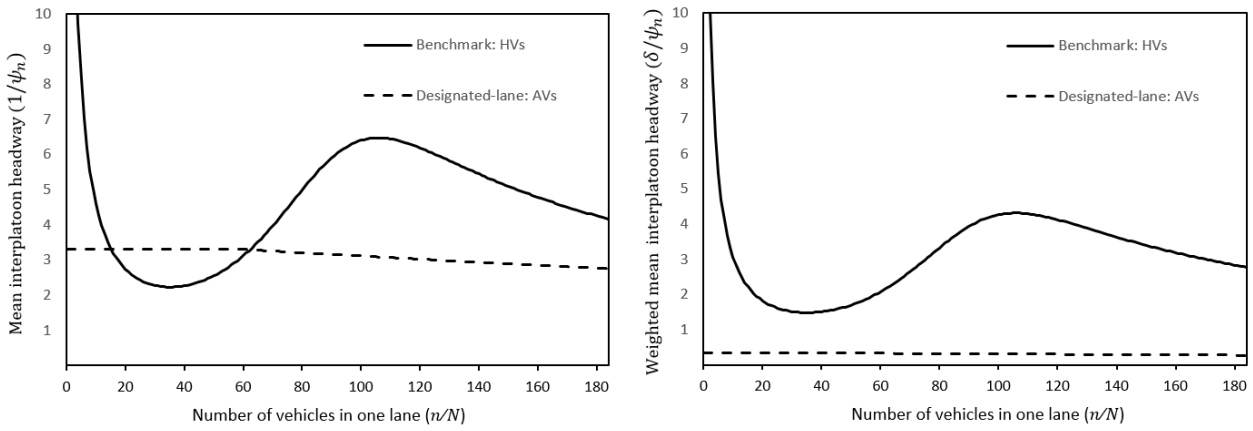
Figure A.5: A comparison between the benchmark case, the D policy and the I policy when $\lambda = 2,217$ vehicles per hour: (a) mean travel time, and (b) throughput.

better than the former. However, under the I policy, AVs improve traffic flow only marginally, and therefore a choice between these two policies does not have a significant impact on congestion. Table A3 shows the maximum percentage of improvement in W and θ over the benchmark case, for several values of $\lambda \leq \underline{\lambda}^{(D,W)} = \lambda^{(D,\theta)} = 2,510$. As this table shows, under the I policy, the increase in θ is almost zero, and W decreases by at most 7%. Since, at all these values of λ, $W^B < 1$ minute, a 7% improvement in W is insignificant.

λ (vehicles per hour)	W^B (min.)	θ^B (vehicles per hour)	Max increase in W (%)	Max increase in θ (%)
500	0.810	500	0.29	0
1000	0.811	1000	0.99	0
1500	0.821	1500	2.13	0
2000	0.835	2000	3.76	0
2500	0.863	2499.97	6.93	0.001

Table A3: Effect of the I policy on the performance measures for lightly loaded highways

In general, when λ is low, unlike the I policy that performs at least as well as the benchmark case, the D policy may not have as good performance. In this case, HVs are capable of driving at the free-flow speed in the benchmark case, and assigning one lane to AVs slows HVs down, while AVs do not improve the speed significantly over the benchmark case.



(a) (b)

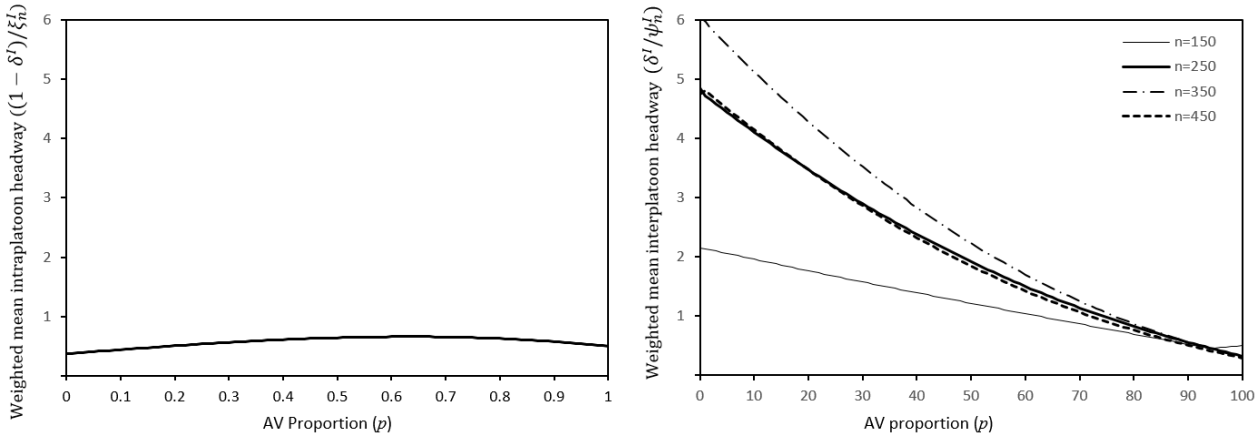
Figure A.6: Comparison between the benchmark case and the AV queue of the D policy: (a) the mean interplatoon headway, and (b) the weighted mean interplatoon headway.

A.5.2 State-Dependent Speed under the D Policy

Under the D policy, the speed of AVs is higher than that of HVs for any given number of vehicles in one lane. To understand this result, recall from equation (3) that speed is a function of mean headway $h = \frac{1-\delta}{\xi} + \frac{\delta}{\psi}$ which is a weighted average of the mean intraplatoon headway $1/\xi$, and the mean interplatoon headway $1/\psi$, with weights determined by the mean platoon size $1/\delta$. In the following, we first compare $\frac{1-\delta}{\xi}$ between HVs and AVs, and then $\frac{\delta}{\psi}$.

First, we compare the weighted mean intraplatoon headways, $\frac{1-\delta}{\xi}$. For the benchmark case, the mean platoon size is $1/\delta^B = 1.5$ vehicles (see Appendix A.4.2 for details of this estimation), and the mean intraplatoon headway is $1/\xi_n^B = 1.1$ seconds (Tientrakool et al. 2011). Although the mean intraplatoon headway of vehicles in the AV queue (0.55 seconds) is shorter than that of HVs in the benchmark case (1.1 seconds), due to the large mean platoon size of AVs (10), the *weighted* mean intraplatoon headway of vehicles in the AV queue, $\frac{1-\delta^{DA}}{\xi_n^{DA}} = 0.495$ seconds, is longer than that in the benchmark case, $\frac{1-\delta^B}{\xi_n^B} = 0.367$ seconds. These values are independent of n, and their difference is in the order of milliseconds.

Next, we compare the weighted mean interplatoon headways, $\frac{\delta}{\psi}$. We can derive the mean interplatoon headway of HVs by rearranging equation (4) as $1/\psi_n^B = \frac{\xi_n^B ND - (1-\delta^B) n V_n^B}{n V_n^B \delta^B \xi_n^B}$. We plot $1/\psi_n^B$ and $1/\psi_n^{DA}$ in equation (9) in Figure A.6(a), which shows that the mean interplatoon headway of vehicles in the AV queue is not always lower than that in the benchmark model. However, Figure



(a) (b)

Figure A.7: Headways under the I policy as a function of the proportion of AVs (p): (a) weighted mean intraplatoon headway, and (b) weighted mean interplatoon headway.

A.6(b) illustrates that, since the mean size of AV platoons is much higher than that of HV platoons, the weighted mean interplatoon headway of vehicles in the AV queue (δ^{DA}/ψ_n^{DA}) is always smaller than that in the benchmark case (δ^B/ψ_n^B), and their minimum difference is 1.15 seconds.

Finally, we combine $\frac{1-\delta}{\xi}$ and $\frac{\delta}{\psi}$. Note that the difference in the weighted mean intraplatoon headway ($= 0.495 - 0.367$ seconds) is always smaller than the difference in the weighted mean interplatoon headway (≥ 1.15 seconds). Therefore, AVs maintain a lower mean headway than HVs, and the speed of vehicles in the AV queue is higher than the speed of vehicles in the benchmark case as well as in the HV queue for any n/N.

A.5.3 State-Dependent Speed under the I Policy

We observe that, for very small values of p (i.e., $p < 0.08$), the speed in the benchmark case can be higher than that under the I policy. In other words, for such p, there exist values of n such that $V_n^I(p) \leq V_n^B$. This occurs because for small p and n the mean intraplatoon headway and the mean interplatoon headway can be both increasing in p. As p increases further, $V_n^I(p) > V_n^B$ for all $n = 1, 2, \cdots, 555$. As depicted in Figure 3(b), when n is high ($n = 250, 350$, and 450), $V_n^I(p)$ increases with p, but for lower values of n ($n = 150$), $V_n^I(p)$ first decreases with p and then increases. As mentioned in §4.2, the mean headway h is the determining factor of speed at each value of n, and h is the weighted average of the mean intraplatoon headway ($(1-\delta)/\xi_n^I$) and the mean interplatoon headway (δ/ψ_n^I). Thus, to understand this result, we first discuss each of these

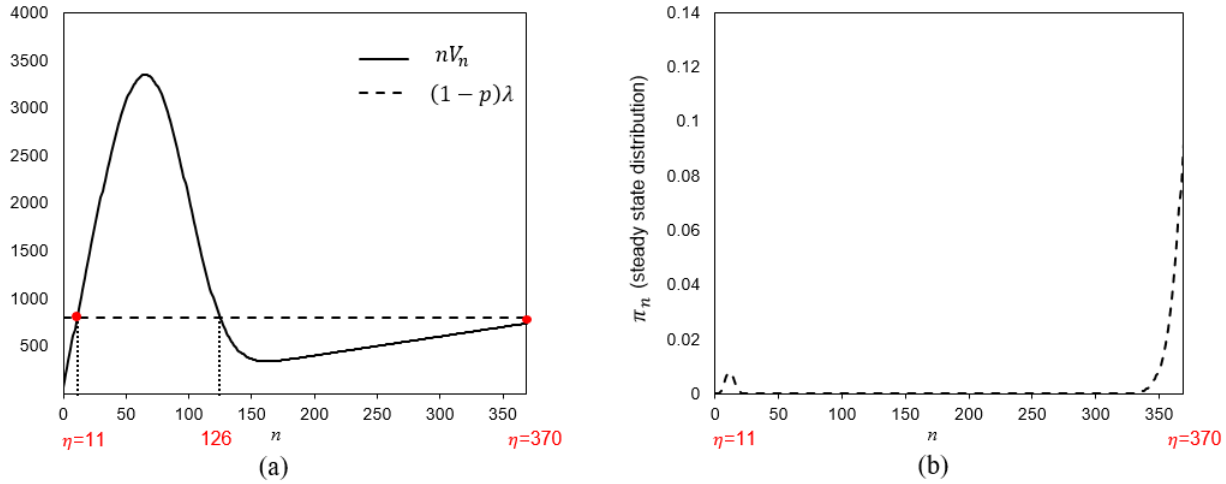


Figure A.8: The HV queue of the D policy when $p = 0.93$: (a) the state-dependent service rate, and (b) the steady state distribution.

headways and then combine them.

First, Figure A.7(a) depicts that the weighted mean intraplatoon headway, which is independent of n, is first increasing and then decreasing in p. The intraplatoon headway of the HV-HV pair (1.1 seconds) is higher than that of the AV-AV pair (0.55 seconds), and lower than that of the HV-AV pair (1.4 seconds). When the proportion of AVs (p) is small (high), HV-AV pairs are more (less) prevalent than AV-AV pairs; hence, as p increases, the weighted mean intraplatoon headway increases (decreases).

Next, Figure A.7(b) illustrates the weighted mean interplatoon headway, which is the product of δ^I and $1/\psi_n^I$. While δ^I is decreasing in p, $1/\psi_n^I$ is not always decreasing in p. According to (11), $1/\psi_n^I$ is the weighted average of the mean interplatoon headway of AVs and that of HVs. There exist values of n such that the mean interplatoon headway of AVs is higher than that of HVs (n/N is between 20 and 60), so $1/\psi_n^I$ increases with p at these values of n, and decreases otherwise. When n is high (e.g., $n = 250, 350$, or 450), as p increases, the number of platoons decreases (δ^I decreasing in p), and the probability of having an AV as the leader of a platoon increases. Because the interplatoon headway maintained by an AV is set to the safe stopping time, which is lower than what a HV maintains for these values of n, the mean interplatoon headway ($1/\psi_n^I$) decreases with p. Thus, for large values of n, δ^I/ψ_n^I, is a decreasing function of p. When n is low (e.g., $n = 150$), δ^I is decreasing in p, but $1/\psi_n^I$ can be increasing in p. For small values of p, the effect of $1/\psi_n^I$ outweighs the effect of δ^I, and for higher values the opposite is true. As a result, δ^I/ψ_n^I is

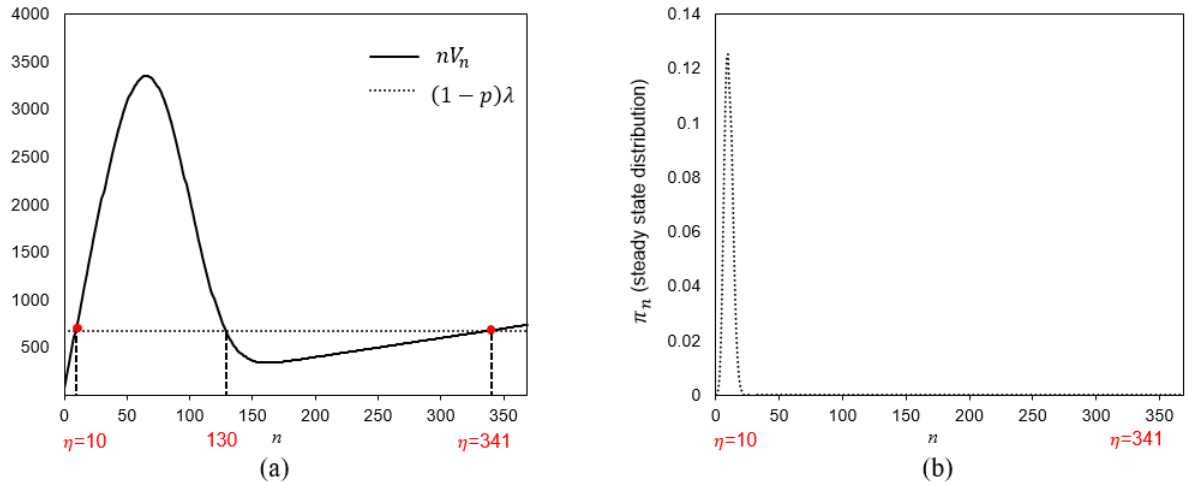


Figure A.9: The HV queue of the D policy when $p = 0.94$: (a) the state-dependent service rate, and (b) the steady state distribution.

increasing in p when p is small enough (e.g., at $p = 0.01$ and $n = 150$), and it is decreasing in p for higher values of p.

Finally, when at least one of n or p is high, since the rate of reduction in the weighted mean interplatoon headway is higher than the rate of increase in the weighted mean intraplatoon headway, the mean headway of vehicles is decreasing in p, and AVs increase the speed of vehicles by reducing the weighted mean interplatoon headway. In contrast, when p and n are both low, AVs decrease the speed of vehicles by forming larger platoons and increasing the mean platoon headway.

A.5.4 Intuition behind the Sharp Decrease in W^{DH}

The behavior of mean travel time as a function of the arrival rate depends on the steady state speed of vehicles. Since speed depends on the number n of vehicles on the highway, to calculate the steady state speed, one needs to know the steady state probability distribution of n, π_n, which depends on the arrival rate λ and the service rate nV_n. Intuitively, when there are n vehicles on the highway such that $nV_n \geq \lambda$ (resp., $nV_n \leq \lambda$), vehicles exit the highway faster (resp., slower) than they enter, and therefore the number of vehicles decreases (resp., increases) over time. As a result, for any given n, if $nV_n \neq \lambda$, the probability of having n vehicles on the highway is very low. This implies that with a high probability there are $[\eta]$ vehicles on the highway in steady state such that $\eta V_\eta = \lambda$ (where η may not be unique).

In the HV queue of the D policy, as shown in Figure 4(a), the mean travel time, W^{DH}, decreases

sharply at $p = 0.93$. This sharp decrease in W^{DH} happens due to the significant difference in the steady state speed of vehicles, $1/\sum_{n=1}^{c} \pi_n(L/V_n)$, when $p = 0.93$ and that when $p = 0.94$. Let us first consider $p = 0.93$. For this p, Figure A.8(a) shows that, for $n < 11$ and $126 < n < 370$, $(1-p)\lambda > nV_n$, and therefore the number of vehicles tends to increase over time to the next point such that $(1-p)\lambda = nV_n$, i.e., 11 and 370, respectively. For $11 < n < 126$, $(1-p)\lambda < nV_n$, and the number of vehicles tends to decrease over time to the last point where $(1-p)\lambda = nV_n$, i.e., 11. As a result, $[\eta] \in \{11, 370\}$, and with a high probability there are either 11 or 370 vehicles on the highway in steady state. Figure A.8(b) illustrates that the probability of having 370 vehicles is about 9 times the probability of having 11 vehicles. As a result, in steady state, there is a high chance that there are 370 vehicles on the highway driving at $V_{370}^{DH} = 2$ miles per hour, and the mean travel time W^{DH} is primarily determined by this low speed of 2 miles per hour. Next, consider $p = 0.94$. For this p, Figure A.9(a) displays that $[\eta]$ is equal to either 10 or 341, but the probability of having 10 vehicles is much higher than that of 341; see Figure A.9(b). As a result, with a high chance vehicles drive at $V_{10}^{DH} = 68$ miles per hour, and W^{DH} is primarily determined by this high speed of 68 miles per hour.

It is interesting to observe the value of $[\eta]$ at which the maximum π_n is attained changes dramatically from $[\eta] = 370$ at $p = 0.93$ to $[\eta] = 10$ at $p = 0.94$. We can understand this better by inspecting (13) closely as follows. As p increases from 0.93 to 0.94, λ is the only parameter that is changed in π_n given in (13). π_n consists of two components: $\pi_0 = (1 + \sum_{n=1}^{c} \frac{(\lambda L)^n}{n! V_n \cdots V_1})^{-1}$, and $\frac{(\lambda L)^n}{n! V_n \cdots V_1}$. Due to the decrease in λ from $(1-0.93)11,342 = 794$ at $p = 0.93$ to $(1-0.94)11,342 = 681$ at $p = 0.94$, π_0 is 80 times greater at $p = 0.94$ than that at $p = 0.93$. Furthermore, when $n = 10$ (resp., 370), the $\frac{(\lambda L)^n}{n! V_n \cdots V_1}$ term of π_n is 80 (resp., 10^{24}) times lower at $p = 0.94$ than that at $p = 0.93$. As a result, $\pi_{10} = 0.007$ at $p = 0.93$ is significantly lower than $\pi_{10} = 0.125$ at $p = 0.94$, whereas $\pi_{370} = 0.1$ at $p = 0.93$ is significantly higher than $\pi_{370} = 1.39 \times 10^{-24}$ at $p = 0.94$.

Formation of a spontaneous jam (having an abrupt decrease in W) for HVs has been observed in prior literature: Bando et al. (1995) and Treiber et al. (2000) show that there exists a critical traffic density at which the highway becomes jammed. However, analyzing AVs, we observe that this is not a universal behavior. Figure 4(a) illustrates that W^{DA} increases fairly smoothly from less than one minute to 2.5 minutes. As depicted in Figure A.10(a), in this case, nV_n is strictly increasing in n, so for each value of the arrival rate $p\lambda$, η is unique; for example, at $\lambda = 681$ vehicles

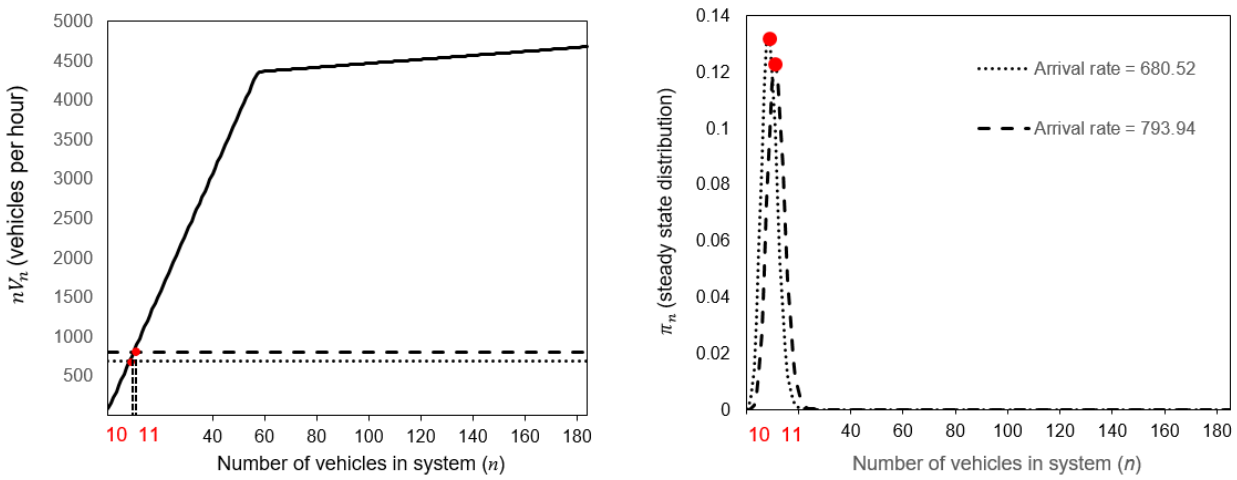


Figure A.10: The AV queue of the D policy: (a) the state dependent service rate, and (b) the steady state mean speed of vehicles.

per hour, $\eta = 10$, and at $\lambda = 794$ vehicles per hour, $\eta = 11$. Therefore, unlike the HV queueing system, a small increase in p does not result in a substantial increase in η, and hence π_η.

A.5.5 D Policy with Two AV Lanes

We numerically analyze a designated-lane policy with two AV lanes ("D2 policy"). As mentioned in §5.3, when p is high, i.e., $p \geq \bar{p}^{(DI,\theta)}$, since the D policy dedicates only one lane to the majority of vehicles, $\theta^D(p)$ becomes lower than $\theta^I(p)$. For high values of p, it seems intuitive to increase the number of designated lanes to AVs under the D policy.

Figure A.11(a) illustrates mean travel time under different policies for $\lambda = 11,342$ vehicles per hour.[2] We observe that the D2 policy reduces W slightly over the D policy, but the I policy is still superior to both of these policies. First, comparing $W^{D2}(p)$ with W^B, we observe that, similar to the D policy, $W^{D2}(p) \leq W^B$ only p is higher than a threshold, i.e., $p \geq 0.58$. Second, although $W^{D2}(p)$ shows a similar pattern to $W^D(p)$, the sharp decrease in $W^{D2}(p)$ happens at a higher p (i.e., $p = 0.96$) than it does in $W^D(p)$ (i.e., $p = 0.93$). As discussed in §5.1, this sharp decrease happens when the HV queue is not highly loaded anymore. Since the number of HVs lanes under the D2 policy is one fewer than that under the D policy, the HV queue under the D2 policy becomes lightly loaded at a higher p than it does under the D policy. Lastly, $W^{D2}(p)$ is never lower than

[2]When p is low, we expect $W^D(p)$ to be lower than $W^{D2}(p)$, but in our numerical analysis, they are equal. Due to a lack of data for a one-lane highway, we set V_n^{DH2} equal to V_{2n}^{DH}, which is higher than the actual speed of vehicles on a one-lane highway.

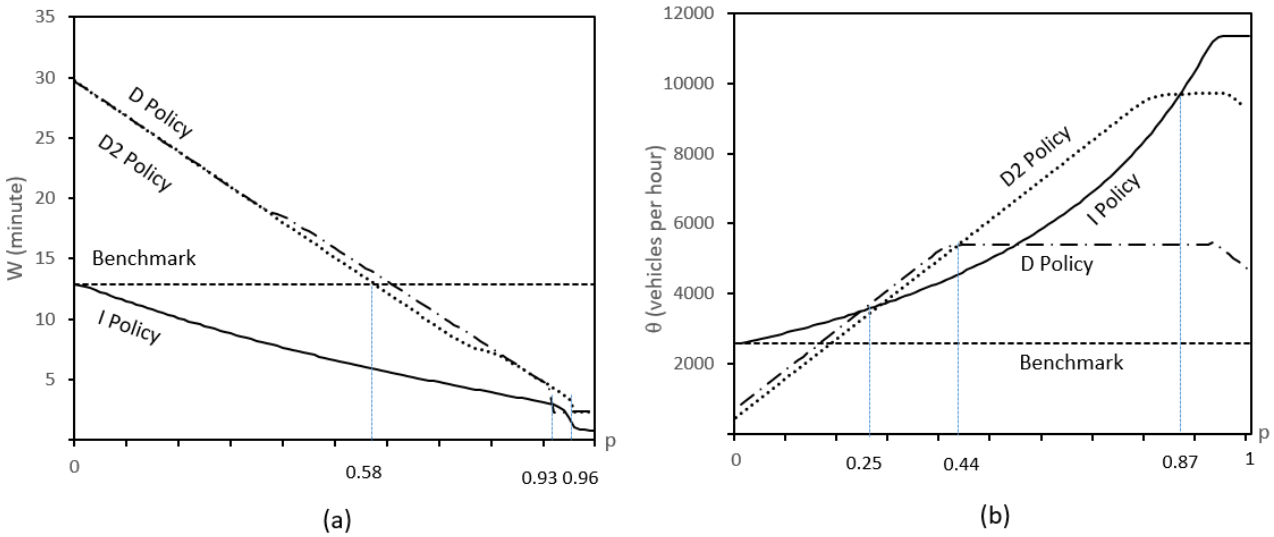


Figure A.11: A comparison among D2 policy, the D policy and the I policy when $\lambda = 11,342$ vehicles per hour: (a) mean travel time, and (b) throughput.

$W^I(p)$. This happens because the HV queue under the D2 policy stays heavily loaded unless p is very high, i.e., $p \geq 0.96$, and the high W of this queue negatively affects $W^{D2}(p)$.

Figure A.11(b) compares throughput under different policies. In summary, for highly loaded highways, adding more lanes helps improve throughput when a proportion of AVs is moderately high, but it has a minimal impact on mean travel time. When p is very low or very high (i.e., $p \leq 0.25$ or $p \geq 0.87$), the I policy performs the best; when p is moderately low (i.e., $0.25 \leq p \leq 0.44$), the D policy performs the best; and when p is moderately high (i.e., $0.44 \leq p \leq 0.87$), the D2 policy performs the best. This is intuitive, because the I policy always improve θ over the benchmark case, but in order for the D and D2 policies to increase θ, p should be moderate, so that both AV and HV queues are well utilized.

A.6 Robustness Checks

A.6.1 Generalization of Analytical Results

We show analytically that our main results hold for any choice of platoon size distribution, any hyperexponential or hypoexponential (also called generalized Erlang) distribution for intraplatoon headway, and any hyperexponential or hypoexponential distribution for interplatoon headway. Note that the majority of continuous distributions can be approximated with either a hyperexponential

distribution or a hypoexponential distribution (Harchol-Balter 2013): a hyperexponential distribution can be used to approximate almost all distributions with a coefficient of variation (CV) greater than one, and a hypoexponential distribution is useful for approximating distributions with a CV less than one (including the deterministic case).

[1] Platoon Size Distribution: We first show that, for any discrete platoon size distribution with mean $1/\delta$, when intraplatoon and interplatoon headways follow exponential distributions, the mean headway is equal to $h = \delta/\psi + (1-\delta)/\xi$. According to Alfa and Neuts (1995), every discrete size distribution with finite support can be represented by

$$(\boldsymbol{\delta}^0, \boldsymbol{G}^0), \text{ where } \boldsymbol{\delta}^0 = [1, 0, \cdots, 0], \text{ and } \boldsymbol{G}^0 = \begin{bmatrix} 0 & \phi_1 & 0 & \cdots & 0 \\ 0 & 0 & \phi_2 & \cdots & 0 \\ \vdots & \vdots & & \ddots & \\ 0 & 0 & 0 & \cdots & \phi_m \\ 0 & 0 & 0 & \cdots & 0 \end{bmatrix}.$$

The mean of this distribution is equal to $1/\delta = \boldsymbol{\delta}^0(\boldsymbol{I} - \boldsymbol{G}^0)^{-1}\mathbf{1} = 1 + \phi_1 + \phi_1\phi_2 + \cdots + \phi_1\phi_2\cdots\phi_m$. Assuming the intraplatoon and interplatoon headways follow $exp(\xi)$ and $exp(\psi)$, respectively, we have

$$\boldsymbol{C}^0 = \begin{bmatrix} -\psi & 0 & \cdots & 0 \\ 0 & -\xi & \cdots & 0 \\ \vdots & \vdots & \ddots & \\ 0 & 0 & \cdots & -\xi \end{bmatrix}, \text{ and } \boldsymbol{C}^1 = \begin{bmatrix} 0 & \psi & 0 & \cdots & 0 \\ (1-\phi_1)\xi & 0 & \phi_1\xi & \cdots & 0 \\ \vdots & \vdots & & \ddots & \\ (1-\phi_m)\xi & 0 & 0 & \cdots & \phi_m\xi \\ \xi & 0 & 0 & \cdots & 0 \end{bmatrix}.$$

Note that these matrices are of size $(m+2) \times (m+2)$. The generator matrix of the MAP is then equal to

$$\boldsymbol{C} = \boldsymbol{C}^0 + \boldsymbol{C}^1 = \begin{bmatrix} -\psi & \psi & 0 & \cdots & 0 \\ (1-\phi_1)\xi & -\xi & \phi_1\xi & \cdots & 0 \\ \vdots & \vdots & & \ddots & \\ (1-\phi_m)\xi & 0 & 0 & \cdots & \phi_m\xi \\ \xi & 0 & 0 & \cdots & -\xi \end{bmatrix}.$$

The stationary probability vector of this MAP, $\tilde{\boldsymbol{\pi}}$, is given by $\tilde{\boldsymbol{\pi}}\boldsymbol{C} = \mathbf{0}$, and $\tilde{\boldsymbol{\pi}} = [\frac{\delta\xi}{\delta\xi+(1-\delta)\psi}, \frac{\delta\psi}{\delta\xi+(1-\delta)\psi}, \frac{\phi_1\delta\psi}{\delta\xi+(1-\delta)\psi}, \cdots,$
Finally, $1/h = \tilde{\boldsymbol{\pi}}\boldsymbol{C}^1\mathbf{1} = \xi\psi/[\delta\xi+(1-\delta)\psi]$ is the arrival rate of the MAP. Hence, $h = \delta/\psi+(1-\delta)/\xi$.

[2] Hyperexponential Distributions for Intraplatoon Headway and Interplatoon Headway: We show that, for any hyperexponential intraplatoon distribution and any hyperexponential interplatoon distribution with means $1/\xi$ and $1/\psi$, respectively, the mean headway is equal to $h = \delta/\psi + (1-\delta)/\xi$. Suppose the platoon size follows a geometric distribution with parameters $\delta = G$ and $\delta^0 = G^0 = 1-\delta$. Assuming the intraplatoon headway follows a hyperexponential distribution with m_1 phases and mean $1/\xi$, we have

$$\boldsymbol{Q}^0(1) = \begin{bmatrix} -\xi_1 & & \\ & \ddots & \\ & & -\xi_{m_1} \end{bmatrix}, \boldsymbol{Q}(1) = \begin{bmatrix} \xi_1 \\ \vdots \\ \xi_{m_1} \end{bmatrix}, \text{ and } \boldsymbol{\alpha}^0(1) = (\alpha_{11}, \cdots, \alpha_{1m_1}).$$

The mean of this distribution is equal to $1/\xi = -\boldsymbol{Q}^0(1)^{-1}\boldsymbol{\alpha}^0(1) = \sum_{i=1}^{m_1} \alpha_{1i}/\xi_i$. Similarly, suppose the interplatoon headway follows a hyperexponential distribution with m_2 phases and mean $1/\psi$. In this case, we have

$$\boldsymbol{Q}^0(2) = \begin{bmatrix} -\psi_1 & & \\ & \ddots & \\ & & -\psi_{m_2} \end{bmatrix}, \boldsymbol{Q}(2) = \begin{bmatrix} \psi_1 \\ \vdots \\ \psi_{m_2} \end{bmatrix}, \text{ and } \boldsymbol{\alpha}^0(2) = (\alpha_{21}, \cdots, \alpha_{2m_2}).$$

The mean of this distribution is equal to $1/\psi = -\boldsymbol{Q}^0(2)^{-1}\boldsymbol{\alpha}^0(2) = \sum_{i=1}^{m_2} \alpha_{2i}/\psi_i$.

By substituting these matrices into (2), we can characterize the MAP with the following matrices:

$$\boldsymbol{C}^0 = \begin{bmatrix} -\psi_1 & & & & & & \\ & \ddots & & & & & \\ & & -\psi_{m_2} & & & & \\ & & & -\xi_1 & & & \\ & & & & \ddots & & \\ & & & & & -\xi_{m_2} \end{bmatrix}, \text{ and}$$

$$\boldsymbol{C}^1 = \begin{bmatrix} \delta\alpha_{21}\psi_1 & \cdots & \delta\alpha_{2m_2}\psi_1 & (1-\delta)\alpha_{11}\psi_1 & \cdots & (1-\delta)\alpha_{1m_1}\psi_1 \\ \vdots & & \vdots & \vdots & & \vdots \\ \delta\alpha_{21}\psi_{m_2} & & \delta\alpha_{2m_2}\psi_{m_2} & (1-\delta)\alpha_{11}\psi_{m_2} & \cdots & (1-\delta)\alpha_{1m_1}\psi_{m_2} \\ \delta\alpha_{21}\xi_1 & & \delta\alpha_{2m_2}\xi_1 & (1-\delta)\alpha_{11}\xi_1 & \cdots & (1-\delta)\alpha_{1m_1}\xi_1 \\ \vdots & & \vdots & \vdots & & \vdots \\ \delta\alpha_{21}\xi_{m_1} & & \delta\alpha_{2m_2}\xi_{m_1} & (1-\delta)\alpha_{11}\xi_{m_1} & \cdots & (1-\delta)\alpha_{1m_1}\xi_{m_1} \end{bmatrix}.$$

The generator matrix of the MAP is then

$$\boldsymbol{C} = \boldsymbol{C}^0 + \boldsymbol{C}^1 = \begin{bmatrix} (\delta\alpha_{21}-1)\psi_1 & \cdots & \delta\alpha_{2m_2}\psi_1 & (1-\delta)\alpha_{11}\psi_1 & \cdots & (1-\delta)\alpha_{1m_1}\psi_1 \\ & \ddots & & \vdots & & \vdots \\ \delta\alpha_{21}\psi_{m_2} & & (\delta\alpha_{2m_2}-1)\psi_{m_2} & (1-\delta)\alpha_{11}\psi_{m_2} & \cdots & (1-\delta)\alpha_{1m_1}\psi_{m_2} \\ \delta\alpha_{21}\xi_1 & & \delta\alpha_{2m_2}\xi_1 & [(1-\delta)\alpha_{11}-1]\xi_1 & \cdots & (1-\delta)\alpha_{1m_1}\xi_1 \\ & & & & \ddots & \\ \delta\alpha_{21}\xi_{m_1} & & \delta\alpha_{2m_2}\xi_{m_1} & (1-\delta)\alpha_{11}\xi_{m_1} & \cdots & [(1-\delta)\alpha_{1m_1}-1]\xi_{m_1} \end{bmatrix}.$$

The stationary probability vector of this MAP, $\tilde{\boldsymbol{\pi}}$, is given by $\tilde{\boldsymbol{\pi}}\boldsymbol{C} = \mathbf{0}$, and

$$\tilde{\boldsymbol{\pi}} = [\frac{\delta\alpha_{21}\psi_{-1}\prod_{i=1}^{m_1}\xi_i}{\gamma}, \cdots, \frac{\delta\alpha_{2m_2}\psi_{-m_2}\prod_{i=1}^{m_1}\xi_i}{\gamma}, \frac{(1-\delta)\alpha_{11}\xi_{-1}\prod_{i=1}^{m_2}\psi_i}{\gamma}, \cdots, \frac{(1-\delta)\alpha_{1m_1}\xi_{-m_1}\prod_{i=1}^{m_2}\psi_i}{\gamma}],$$

where $\gamma = \delta\prod_{i=1}^{m_1}\xi_i(\alpha_{21}\psi_{-1} + \cdots + \alpha_{2m_2}\psi_{-m_2}) + (1-\delta)\prod_{i=1}^{m_2}\psi_i(\alpha_{11}\xi_{-1} + \cdots + \alpha_{1m_1}\xi_{-m_1})$, $\xi_{-i} = \xi_1\cdots\xi_{i-1}\xi_{i+1}\cdots\xi_{m_1}$ for $i \in \{1, \cdots, m_1\}$ and $\psi_{-j} = \psi_1\cdots\psi_{j-1}\psi_{j+1}\cdots\psi_{m_1}$ for $j \in \{1, \cdots, m_2\}$.

Finally, after some simplifications, the mean headway of vehicles is calculated as $h = 1/(\tilde{\boldsymbol{\pi}}\boldsymbol{C}^1\mathbf{1}) = \frac{\delta}{\psi} + \frac{1-\delta}{\xi}$.

[3] Hypoexponential Distributions for Intraplatoon Headway and Interplatoon Headway: We show that, for any hypoexponential intraplatoon distribution and any hypoexponential interplatoon distribution with means $1/\xi$ and $1/\psi$, respectively, the mean headway is equal to $h = \delta/\psi + (1-\delta)/\xi$.[3] Suppose the platoon size follows a geometric distribution with parameters $\delta = G$ and $\delta^0 = G^0 = 1-\delta$. Assuming the intraplatoon headway follows a hypoexponential

[3] A deterministic distribution (or constant) can be approximated as an Erlang distribution. This Erlang distribution is a special case of a hypoexponential distribution for when the number of phases grows large while the time spent in each state goes to zero.

distribution with m_1 phases and mean $1/\xi$, we have

$$\boldsymbol{Q}^0(1) = \begin{bmatrix} -\xi_1 & \xi_1 & 0 & \cdots & 0 \\ 0 & -\xi_2 & \xi_2 & \cdots & 0 \\ & & \ddots & & \\ 0 & 0 & 0 & \cdots & -\xi_{m_1} \end{bmatrix}, \boldsymbol{Q}(1) = \begin{bmatrix} 0 \\ \vdots \\ 0 \\ \xi_{m_1} \end{bmatrix}, \text{ and } \boldsymbol{\alpha}^0(1) = (1, 0, \cdots, 0).$$

The mean of this distribution is equal to $1/\xi = -\boldsymbol{Q}^0(1)^{-1}\boldsymbol{\alpha}^0(1) = \sum_{i=1}^{m_1} 1/\xi_i$. Similarly, suppose the interplatoon headway follows a hyperexponential distribution with m_2 phases and mean $1/\psi$. In this case, we have

$$\boldsymbol{Q}^0(2) = \begin{bmatrix} -\psi_1 & \psi_1 & 0 & \cdots & 0 \\ 0 & -\psi_2 & \psi_2 & \cdots & 0 \\ & & \ddots & & \\ 0 & 0 & 0 & \cdots & -\psi_{m_2} \end{bmatrix}, \boldsymbol{Q}(2) = \begin{bmatrix} 0 \\ \vdots \\ 0 \\ \psi_{m_2} \end{bmatrix}, \text{ and } \boldsymbol{\alpha}^0(2) = (1, 0, \cdots, 0).$$

The mean of this distribution is equal to $1/\psi = -\boldsymbol{Q}^0(2)^{-1}\boldsymbol{\alpha}^0(2) = \sum_{i=1}^{m_2} 1/\psi_i$.

By substituting these matrices into (2), we can characterize the MAP with the following matrices:

$$\boldsymbol{C}^0 = \begin{bmatrix} -\psi_1 & \psi_1 & 0 & \cdots & 0 & & & & & \\ 0 & -\psi_2 & \psi_2 & \cdots & 0 & & & & & \\ & & \ddots & & & & & & & \\ 0 & 0 & 0 & \cdots & -\psi_{m_2} & & & & & \\ & & & & & -\xi_1 & \xi_1 & 0 & \cdots & 0 \\ & & & & & 0 & -\xi_2 & \xi_2 & \cdots & 0 \\ & & & & & & & \ddots & & \\ & & & & & 0 & 0 & 0 & \cdots & -\xi_{m_1} \end{bmatrix}, \text{ and}$$

$$\boldsymbol{C}^1 = \begin{bmatrix} 0 & 0 & \cdots & 0 & 0 & 0 & \cdots & 0 \\ \vdots & \vdots & & \vdots & \vdots & \vdots & & \vdots \\ 0 & 0 & \cdots & 0 & 0 & 0 & \cdots & 0 \\ \delta\psi_{m_2} & 0 & \cdots & 0 & (1-\delta)\psi_{m_2} & 0 & \cdots & 0 \\ 0 & 0 & \cdots & 0 & 0 & 0 & \cdots & 0 \\ \vdots & \vdots & & \vdots & \vdots & \vdots & & \vdots \\ 0 & 0 & \cdots & 0 & 0 & 0 & \cdots & 0 \\ \delta\xi_{m_1} & 0 & \cdots & 0 & (1-\delta)\xi_{m_1} & 0 & \cdots & 0 \end{bmatrix}.$$

The generator matrix of the MAP is then

$$\boldsymbol{C} = \boldsymbol{C}^0 + \boldsymbol{C}^1 = \begin{bmatrix} -\psi_1 & \psi_1 & 0 & \cdots & 0 & 0 & 0 & 0 & \cdots & 0 \\ 0 & -\psi_2 & \psi_2 & \cdots & 0 & 0 & 0 & 0 & \cdots & 0 \\ & & \ddots & & & 0 & 0 & 0 & \cdots & 0 \\ \delta\psi_{m_2} & 0 & 0 & \cdots & -\psi_{m_2} & (1-\delta)\psi_{m_2} & 0 & 0 & \cdots & 0 \\ 0 & 0 & 0 & \cdots & 0 & -\xi_1 & \xi_1 & 0 & \cdots & 0 \\ 0 & 0 & 0 & \cdots & 0 & 0 & -\xi_2 & \xi_2 & \cdots & 0 \\ & & & & & & & \ddots & & \\ \delta\xi_{m_1} & 0 & 0 & \cdots & 0 & (1-\delta)\xi_{m_1} & 0 & 0 & \cdots & -\xi_{m_1} \end{bmatrix}.$$

The stationary probability vector of this MAP, $\tilde{\boldsymbol{\pi}}$, is given by $\tilde{\boldsymbol{\pi}}\boldsymbol{C} = \mathbf{0}$, and

$$\tilde{\boldsymbol{\pi}} = [\frac{\delta\xi\psi}{\psi_1[\delta\xi + (1-\delta)\psi]}, \cdots, \frac{\delta\xi\psi}{\psi_{m_2}[\delta\xi + (1-\delta)\psi]}, \frac{(1-\delta)\xi\psi}{\xi_1[\delta\xi + (1-\delta)\psi]}, \cdots, \frac{(1-\delta)\xi\psi}{\xi_{m_1}[\delta\xi + (1-\delta)\psi]}].$$

Finally, after some simplifications, the mean headway of vehicles is calculated as $h = 1/(\tilde{\boldsymbol{\pi}}\boldsymbol{C}^1\mathbf{1}) = \frac{\delta}{\psi} + \frac{1-\delta}{\xi}$.

A.6.2 Sensitivity Analysis

We perform sensitivity analyses on several model parameters used in our numerical analysis in §4. We observe that, although different model parameters may affect the performance of our policies, the insights provided by our calibrated model as well as our recommended policies remain

unchanged: if the performance metric considered by policy makers is mean travel time W, we recommend the I policy; otherwise, we recommend the D policy in a moderate region of p, and the I policy in all other regions. Next, we present details of 10 possible cases we consider for sensitivity analyses.

Case 1. (AVs maintain a fixed headway of 0.55 seconds) As the AV technology improves, these vehicles might become capable of maintaining a fixed headway, improving their efficiency with respect to traffic flow. We perform a sensitivity analysis on this feature as follows. Suppose platoon size, intraplatoon headway, and interplatoon headway of AVs are deterministic, and an AV maintains an intraplatoon headway and an interplatoon headway of 0.55 seconds from the vehicle immediately in front, whether this vehicle is an AV or an HV.[4] Recall that, in our base model, an AV maintains an intraplatoon headway of 1.4 seconds from an HV, and interplatoon headway equal to safe stopping time. These changes affect the state-dependent speed in the AV queue of the D policy, V_n^{DA}, as well as the state-dependent speed under the I policy, $V_n^I(p)$. For the AV queue, by substituting $h_n^{DA} = 0.55$ sec in $V_n^{DA} = 1/(nh_n^{DA})$, we get $V_n^{DA} = \min\{74.7, \frac{6{,}545.46}{n}\}$, where 74.7 is the free-flow speed of the highway. Under the I policy, by substituting $1/\xi^{IAA} = 1/\xi^{IHA} = 1/\psi^{IAA} = 1/\psi^{IHA} = 0.55$ sec in (10) and (11), and after some simplifications, we obtain $V_n^I(p)$ for $n = 1, 2, \cdots, 555$ and $p \in [0, 1]$ as follows:

$$V_n^I(p) = \min\{74.7, \frac{3,600N/n}{0.55p + (1-p)\frac{(1.1+1.87p)}{3} + 2.17p[10,800 - 0.55n(46.67e^{-\frac{n^2}{21,049}} + 3.13)]/[3n(46.67e^{-\frac{n^2}{21,049}} + 3.13)]}\}.$$

Figure A.13 shows that our results are reasonably robust to this change, and the overall intuition provided by the original calibrated model holds: if the performance metric considered by policy makers is mean travel time W, we recommend the I policy; otherwise, we recommend the D policy in a moderate region of p, and the I policy in other regions. We observe that this case improves the performance of both D policy and I policies over the base case presented in §5, but the amount of improvement in the performance of the I policy is more significant than that of the D policy, because the I policy benefits from this change in all three lanes of the highway. In terms of W, this change reduces W^I by at most 3.11 minutes (79.19%) compared to the original W^I, and reduces

[4]Since all consecutive AVs form one platoon, we consider this scenario only when there is a sufficient load of AVs that can maintain a fixed headway of 0.55 seconds.

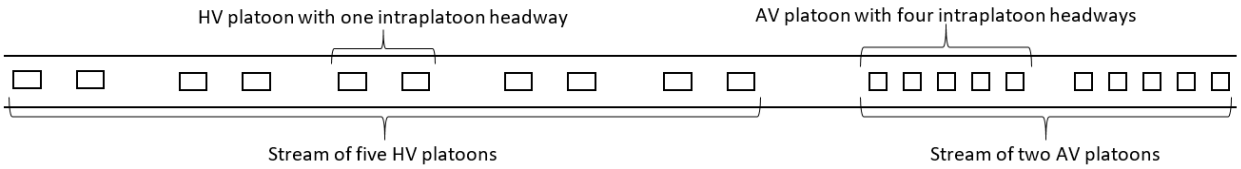


Figure A.12: An illustration of homogeneous platoons. In this example, $n = 20$, $p = 0.5$, $\delta^A = 0.2$, and $\delta^H = 0.5$.

W^D by at most 0.79 minutes (28.72%) compared to the original W^D. In terms of θ, the maximum increase in θ is $3,300$ vehicles per hour (43.68%) under the I policy and $1,870$ vehicles per hour (39.98%) under the D policy, both compared to their corresponding original values. As a result, the interval of p in which the D policy outperforms the I policy, in terms of θ, slightly shrinks from $p \in [0.25, 0.55]$ in the base case to $p \in [0.32, 0.55]$.

Case 2. (Homogeneous platoons) In order to investigate the robustness of our results with respect to platoon mixture (i.e., the degree to which a vehicle, HV or AV, forms homogeneous platoons), we perform a sensitivity analysis on this feature as follows. Whereas, platoons can consist of both AVs and HVs under our original I policy, we consider an extreme case, in which platoons are entirely homogeneous so that each platoon consists of only AVs or only HVs (see Figure A.12). Note that this change affects only the I policy. We first explain our approach and then discuss the numerical result from the calibrated model.

We update the mean intraplatoon headway in (10), as follows:

$$1/\xi_n^I = [\frac{np\delta^A(1/\delta^A - 1)}{\xi_n^{IAA}} + \frac{n(1-p)\delta^H(1/\delta^H - 1)}{\xi_n^{IHH}}]/n = 0.495p + 0.367(1-p).$$

In this expression, the term $np\delta^A(1/\delta^A - 1)$ consists of $np\delta^A$, which is the number of AV platoons, and $(1/\delta^A - 1)$, which is the number of intraplatoon headways that exist within an AV platoon. Hence, $np\delta^A(1/\delta^A - 1)$ represents the total number of intraplatoon headways within all AV platoons, and $\frac{np\delta^A(1/\delta^A-1)}{n\xi_n^{IAA}}$ indicates the mean intraplatoon headway of AVs. Similarly, $\frac{n(1-p)\delta^H(1/\delta^H-1)}{n\xi_n^{IHH}}$ represents the mean intraplatoon headway of HVs. We also update the mean interplatoon headway in (11), as follows:

$$1/\psi_n^I = \{\frac{(np\delta^A-1)^+}{\psi_n^{IAA}} + \frac{[n(1-p)\delta^H-1]^+}{\psi_n^{IHH}} + \frac{1}{2}(\frac{1}{\psi_n^{IAH}} + \frac{1}{\psi_n^{IHA}})\}/n$$

$$= \frac{(0.1np-1)^+(10,800-2.97n) + \frac{n}{2}(10,800-7.56n)}{n(3,000+0.6n)} + \frac{\{\frac{n}{2} + [0.667n(1-p)-1]^+\}[10,800-0.55n(46.67e^{-\frac{n^2}{21,049}}+3.13)]}{n^2(46.67e^{-\frac{n^2}{21,049}}+3.13)}.$$

In this expression, the terms $np\delta^A - 1$ and $n(1-p)\delta^H - 1$ represent the number of AV interplatoon headways and the number of HV interplatoon headways, respectively. In addition, the term $\frac{1}{2}(\frac{1}{\psi_n^{IHA}} + \frac{1}{\psi_n^{IAH}})$ represents the interplatoon headway between the stream of AV platoons and the stream of HV platoons, considering the following two scenarios: there exists an AV platoon that follows an HV platoon and maintains an interplatoon headway of $\frac{1}{\psi_n^{IHA}}$ seconds, or there exists an HV platoon that follows an AV platoon and maintains an interplatoon headway of $\frac{1}{\psi_n^{IAH}}$ seconds. We assume either of these scenarios happens with probability $\frac{1}{2}$.

As discussed in the paper, the mean headway of vehicles when there are n vehicles on the segment is calculated as $h_n^I = \frac{1}{\xi_n^I} + \frac{1}{\psi_n^I}$. Substituting h_n^I in $V_n^I(p) = N/nh_n^I$, we obtain $V_n^I(p)$ for $n = 1, 2, \cdots, 555$, and $p \in [0, 1]$ as follows:

$$V_n^I(p) = \min\{74.7, \frac{3,600N}{0.495pn + 0.367n(1-p) + \frac{(0.1np-1)^+(10,800-2.97n)+\frac{n}{2}(10,800-7.56n)}{n(3,000+0.6n)} + \frac{\{\frac{n}{2}+[0.667n(1-p)-1]^+\}[10,800-0.55n(46.67e^{-\frac{n^2}{21,049}}+3.13)]}{n^2(46.67e^{-\frac{n^2}{21,049}}+3.13)}}\}.$$

For a given p, although the overall number of platoons is equal to that of the original I policy, i.e., $n[p\delta^A + (1-p)\delta^H] = n\delta^I$, the chance of having an HV at the head of a platoon increases from $(1-p)$ in the original model to $\frac{\delta^H}{\delta^I}(1-p)$. Since $\delta^H < \delta^I$, this leads to a higher mean interplatoon headway than that under the original I policy. As a result, the performance of the I policy is slightly inferior to that of the original I policy.

As Figure A.14 illustrates, this change in the platoon mixture can increase W by at most 1.28 minutes (15.65%), and decrease θ by 793.27 vehicles per hour (13.49%). As Figure A.14(a) depicts, the I policy leads to a lower W than the D policy for all values of p, except when p is between $p^{(DI,W)} = 0.93$ and 0.95. Figure A.14(b) shows that, when $0.21 \le p \le 0.64$, the D policy outperforms the new I policy in terms of θ. This interval is wider than what we have for the original I policy, i.e., $p \in [0.25, 0.55]$, presented in Figure 6. This is still aligned with our policy recommendations in §6.

Case 3. (AV mean platoon size increases from 10 **to** 20 **vehicles)** As the AV technology advances, these vehicles might be able to form larger platoons. Hence, we perform a sensitivity analysis on mean platoon size of AVs. For this case, we assume the mean platoon size of AVs, $\frac{1}{\delta^{DA}}$, is equal to 20 vehicles, and we update the mean platoon size under the I policy as $\frac{1}{\delta^{I}} = \frac{6}{4-3.7p}$. Substituting these values in $V_n = \frac{N}{nh_n}$, we get $V_n^{DA} = \min\{74.7, \frac{3{,}600+1.08n}{0.7025n}\}$ for $n = 1, 2, \cdots, 185$, and we have $V_n^I(p)$ for $n = 1, 2, \cdots, 555$ and $p \in [0, 1]$ as follows:

$$V_n^I(p) = \min\{74.7, \frac{21,600N/[n(2+3.7p)]}{\frac{4-3.7p}{2+3.7p}\left[\frac{(10{,}800-7.56n)p+4.59np^2}{3{,}000+0.6n} + \frac{10{,}800(1-p)}{n\left(46.67e^{-\frac{n^2}{21{,}049}}+3.13\right)} - 0.55(1-p)\right] + [0.55p^2 + 1.4p(1-p) + 1.1(1-p)]}\}.$$

Figure A.15 illustrates that this change improves the performance of both policies, because having larger platoons leads to smaller mean headways. However, our recommended policies remain mostly the same; in terms of W, the I policy outperforms the D policy for all values of p, and in terms of θ, the I policy is superior to the D policy unless $p \in [0.25, 0.60]$ (compared to $p \in [0.25, 0.55]$ depicted in Figure 6).

Case 4. (AV-AV and HV-AV mean intraplatoon headways decrease from 0.55 **and** 1.4 **seconds, respectively, to** 0.1 **seconds)** As the AV technology advances, these vehicles might be able to maintain much smaller intraplatoon headways. Hence, we perform a sensitivity analysis on mean intraplatoon headway of AVs. For this case, we reduce the mean intraplatoon headways of AVs to 0.1 seconds, i.e., $\frac{1}{\xi^{DA}} = \frac{1}{\xi^{IAA}} = \frac{1}{\xi^{IHA}} = 0.1$ sec. By substituting these values in $V_n = \frac{N}{nh_n}$, we get $V_n^{DA} = \min\{74.7, \frac{3{,}600+2.16n}{0.45n}\}$ for $n = 1, 2, \cdots, 185$, and we have $V_n^I(p)$ for $n = 1, 2, \cdots, 555$ and $p \in [0, 1]$ as follows:

$$V_n^I(p) = \min\{74.7, \frac{10,800N/[n(1+1.7p)]}{\frac{2-1.7p}{1+1.7p}\left[\frac{(10{,}800-7.56n)p+4.59np^2}{3{,}000+0.6n} + \frac{10{,}800(1-p)}{n\left(46.67e^{-\frac{n^2}{21{,}049}}+3.13\right)} - 0.55(1-p)\right] + (1.1-p)}\}.$$

As Figure A.16 depicts, this change improves the performance of both policies. The effect of this change on the performance of the I policy is more significant when p is high, because as p increases, the mean headway decreases. However, since the D policy designates a lane to AVs, this policy starts to benefit from this change at lower values of p. In addition, the jam throughput of

the D policy increases by 78%, from $5,419$ vehicles per hour in the base case to $9,627$ vehicles per hour. The improvement in the throughput of the I policy is not as substantial as that of the D policy, because the mean headways of HVs remain unchanged. Taken together, this leads to the I policy dominating with respect to W for all values of p, and a wider region of p, i.e., $p \in [0.26, 0.72]$, compared to $p \in [0.25, 0.55]$ of the base case, in which the D policy outperforms the I policy in terms of θ.

Case 5. (AV mean interplatoon headway decreases to half of the safe stopping time) As the AV technology advances, these vehicles might be able to maintain smaller interplatoon headways. Hence, we perform a sensitivity analysis on mean interplatoon headway of AVs. For this case, we update the AV mean interplatoon equations as follows: $\frac{1}{\psi_n^{DA}} = 0.5(3.6 - \frac{21.6}{V_n^D A})$, $\frac{1}{\psi_n^{IAA}} = \frac{10{,}800-2.97n}{6{,}000+0.6n}$, and $\frac{1}{\psi_n^{IHA}} = \frac{10{,}800-7.56n}{6{,}000+0.6n}$. By substituting these values in $V_n = \frac{N}{nh_n}$, we get $V_n^{DA} = \min\{74.7, \frac{3{,}600+1.08n}{0.675n}\}$ for $n = 1, 2, \cdots, 185$, and we have $V_n^I(p)$ for $n = 1, 2, \cdots, 555$ and $p \in [0, 1]$ as follows:

$$V_n^I(p) = \min\{74.7, \frac{10,800N/[n(1+1.7p)]}{\frac{2-1.7p}{1+1.7p}\left[\frac{(10{,}800-7.56n)p+4.59np^2}{6{,}000+0.6n} + \frac{10{,}800(1-p)}{n\left(46.67e^{-\frac{n^2}{21{,}049}}+3.13\right)} - 0.55(1-p)\right] + [0.55p^2, 1.4p(1-p) + 1.1(1-p)]}\}.$$

As Figure A.17 illustrates, although this change improves the performance of both policies over the base case, it does not affect our recommended policies in §6: The I policy is recommended for all values of p when considering W, and for all values of p, except for $p \in [0.27, 0.59]$, when considering θ.

Case 6. (Simultaneous improvement of Cases 3, 4, and 5) Technological improvement of AVs might affect the AV mean headway in one (or more) of the following ways: (1) by increasing mean platoon size (Case 3), (2) by decreasing mean intraplatoon headway (Case 4), and (3) by decreasing mean interplatoon headway (Case 5). In Case 6, we apply all three changes simultaneously to ensure the robustness of our policy recommendations. In this case, we have $V_n^{DA} = \min\{74.7, \frac{3{,}600+0.54n}{0.185n}\}$ for $n = 1, 2, \cdots, 185$, and we have $V_n^I(p)$ for $n = 1, 2, \cdots, 555$ and

$p \in [0,1]$ as follows:

$$V_n^I(p) = \min\{74.7, \frac{21,600N/[n(2+3.7p)]}{\frac{4-3.7p}{2+3.7p}\left[\frac{(10{,}800-7.56n)p+4.59np^2}{6{,}000+0.6n} + \frac{10{,}800(1-p)}{n\left(46.67e^{-\frac{n^2}{21{,}049}}+3.13\right)} - 0.55(1-p)\right] + (1.1-p)}\}.$$

Figure A.18 shows that technological improvement of AVs significantly boosts the performance of both policies. In terms of W, this change reduces W^I by at most 3.55 minutes (81.51%) compared to the original W^I, and reduces W^D by at most 1.55 minutes (65.80%) compared to the original W^D. In terms of θ, the maximum increase in θ is $4,092$ vehicles per hour (57.68%) under the I policy and $6,654$ vehicles per hour (142.31%) under the D policy, both compared to their corresponding original values. In this case, when $p \geq 0.93$, the D policy is no longer highly loaded, and has the maximum achievable throughput, i.e., $\theta = \lambda = 11,342$ vehicles per hour. The performance of the I policy also improves substantially. Under this policy, when $p \geq 0.75$, the highway is no longer highly loaded. Despite these performance improvements, our recommended policies stay intact. In other words, if the performance metric considered by the policy makers is W, we recommend the I policy; otherwise, we recommend the D policy in a moderate region of p, i.e., $p \in [0.32, 0.55]$, and the I policy in the other regions.

Case 7. (Mean interplatoon headway of an AV-HV pair increases from 1.1 seconds to 2.2 seconds) Due to unfamiliarity of human drivers with AVs, an HV might maintain a larger gap from an AV than it does from another HV. Hence, we examine the performance of the I policy when the intraplatoon headway of an AV-HV pair is increased to 2.2 seconds, i.e., $\frac{1}{\xi^{IAH}} = 2.2$ sec. This change affects only the I policy. By substituting this value in $V_n^I(p) = \frac{N/n}{\frac{\delta^I}{\psi_n^I} + \frac{1-\delta^I}{\xi_n^I}}$, we obtain $V_n^I(p)$ for $n = 1, 2, \cdots, 555$ and $p \in [0,1]$ as follows:

$$V_n^I(p) = \min\{74.7, \frac{10,800N/[n(1+1.7p)]}{\frac{2-1.7p}{1+1.7p}\left[\frac{(10{,}800-7.56n)p+4.59np^2}{3{,}000+0.6n} + \frac{10{,}800(1-p)}{n\left(46.67e^{-\frac{n^2}{21{,}049}}+3.13\right)} - 0.55(1-p)\right] + [0.55p^2, 1.4p(1-p) + 1.1(1-p^2)]}\}.$$

Figure A.19 shows that, as expected, if HVs maintain larger headways from AVs than HVs, the performance of the I policy is negatively affected (especially in terms of θ). However, even by doubling the AV-HV headway, our overall policy recommendations are not significantly impacted;

only the moderate region of p in which the D policy is recommended based on θ becomes wider, i.e., $p \in [0.23, 0.61]$ (compared to $p \in [0.25, 0.55]$ in the base case, depicted in Figure 6).

Case 8. (Highway length increases from one mile to two miles) To check the robustness of our results with respect to the choice of highway length L, we consider a highway segment of two miles. As our analysis shows (see (A.6) and (A.7) in Appendix A.3), for a highly loaded highway, mean travel time W should be proportional to the length of the highway, and throughput θ, which represents the number of vehicles that exit the highway per hour, should be independent of the choice of L. To demonstrate this, we substitute $L = 2$ miles into $V_n = \frac{NL}{nh_n}$. As a result, V_n^{DA} is equal to $\min\{74.7, \frac{(3{,}600L+2.16}{0.855n}\}$ for $n = 1, 2, \cdots, 370$, and we have $V_n^I(p)$ for $n = 1, 2, \cdots, 1110$ and $p \in [0, 1]$ as follows:

$$V_n^I(p) = \min\{74.7, \frac{10,800NL/[n(1+1.7p)]}{\frac{2-1.7p}{1+1.7p}\left[\frac{(10,800*L-7.56n)p+4.59np^2}{3,000*L+0.6n} + \frac{10,800*L(1-p)}{n\left(46.67e^{-\frac{n^2}{21,049}}+3.13\right)} - 0.55(1-p)\right] + [0.55p^2 + 1.4p(1-p) + 1.1(1-p)]}\}.$$

Figure A.20(a) depicts a comparison between the D policy and the I policy in terms of W. We observe that the results are identical to those of the base case illustrated in Figure 6(a), except that the graph is vertically scaled up by a factor of two. Similarly, in terms of θ, Figure A.20(b) shows that our results for a two-mile highway segment are identical to those depicted in Figure 6(b).

Case 9. (Mean intraplatoon headway of an HV-AV pair reduces from 1.4 seconds to 0.55 seconds) To assess the robustness of our analysis to the intraplatoon headway of an HV-AV pair, $\frac{1}{\xi_n^{IHA}}$, we reduce this parameter from 1.4 seconds in the base case to the value of the intraplatoon headway of an AV-AV pair, i.e., 0.55 seconds. By substituting this value in (10), we have $1/\xi_n^I = 1.1 - 0.55p$, and after some simplifications, we have obtain $V_n^I(p)$ for $n = 1, 2, \cdots, 555$ and $p \in [0, 1]$ as follows:

$$V_n^I(p) = \min\{74.7, \frac{10,800N/[n(1+1.7p)]}{\frac{2-1.7p}{1+1.7p}\left[\frac{(10,800-7.56n)p+4.59np^2}{3,000+0.6n} + \frac{10,800(1-p)}{n\left(46.67e^{-\frac{n^2}{21,049}}+3.13\right)} - 0.55(1-p)\right] + (1.1-0.55p)}\}.$$

Figure A.21 depicts that this change slightly improves the performance of the I policy, and has no effect on the D policy. Hence, our recommended policies stay unchanged, and our main results

are robust to the value of this parameter. Compared to the results of the base case, illustrated in Figure 6, the interval of p in which the D policy outperforms the I policy in terms of θ slightly shrinks from $p \in [0.25, 0.55]$ to $p \in [0.26, 0.52]$.

Case 10. (AV mean intraplatoon headway increases from 0.55 seconds to 1.1 seconds) One of the main advantages of AVs over HVs is their ability to maintain a low intraplatoon headway. To examine the importance of this feature, we increase the mean intraplatoon headway of AVs from 0.55 seconds to 1.1 seconds, which is the mean intraplatoon headway that HVs maintain; i.e., we let $\frac{1}{\xi^{DA}} = \frac{1}{\xi^{IAA}} = 1.1$ sec. By substituting these values in $V_n = \frac{N}{nh_n}$, we get $V_n^{DA} = \min\{74.7, \frac{3{,}600+2.16n}{1.35n}\}$ for $n = 1, 2, \cdots, 185$, and we have $V_n^I(p)$ for $n = 1, 2, \cdots, 555$ and $p \in [0, 1]$ as follows:

$$V_n^I(p) = \min\{74.7, \frac{10,800N/[n(1+1.7p)]}{\frac{2-1.7p}{1+1.7p}\left[\frac{(10{,}800-7.56n)p+4.59np^2}{6{,}000+0.6n} + \frac{10{,}800(1-p)}{n\left(46.67e^{-\frac{n^2}{21{,}049}}+3.13\right)} - 0.55(1-p)\right] + [1.1+0.3p(1-p)]}\}.$$

As figure A.22 illustrates, although the I policy outperforms the benchmark model on both metrics, the highway remains highly loaded even when p is high. The performance of the D policy deteriorates even more compared to the base case. The jam throughput of this model reduces from $5,419$ vehicles per hour in the original model to $3,330$ vehicles per hour. Now the D policy outperforms the I policy only when $p \in [0.92, 0.97]$ in terms of W, and when $p \in [0.26, 0.29]$ in terms of θ. Except for these values of p, we recommend the I policy over the D policy based on both metrics.

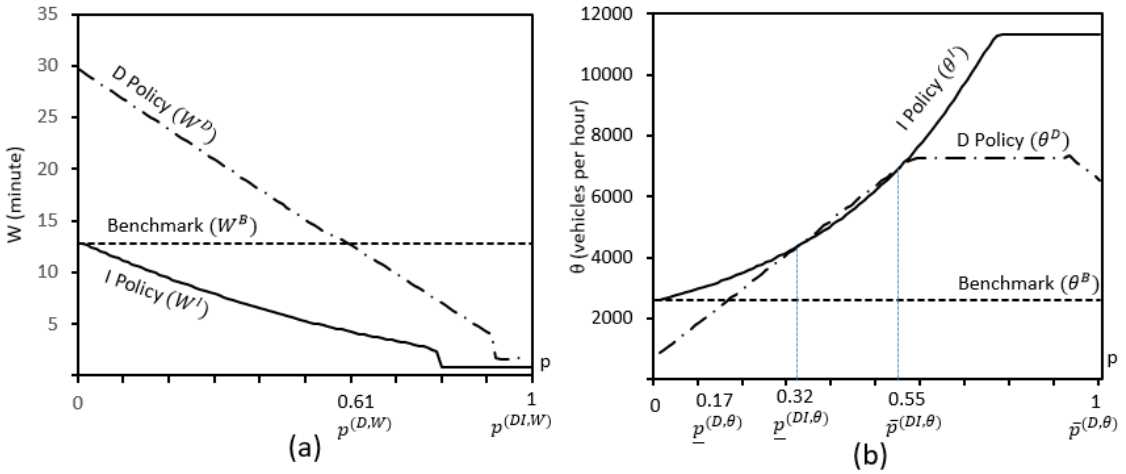


Figure A.13: A comparison between the D policy and the I policy for Case 1 when $\frac{1}{\psi_n^{DA}} = \frac{1}{\psi_n^{IAA}} = \frac{1}{\psi_n^{IHA}} = \frac{1}{\xi_n^{IHA}} = 0.55$ seconds and $\lambda = 11,342$ vehicles per hour: (a) mean travel time, and (b) throughput.

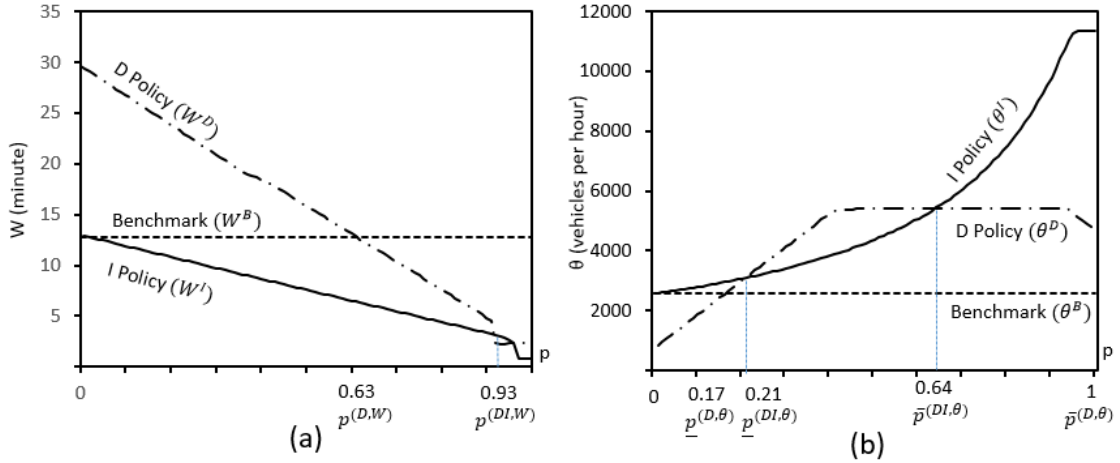


Figure A.14: A comparison between the D policy and the I policy for Case 2 of homogeneous platoons, when $\lambda = 11,342$ vehicles per hour: (a) mean travel time, and (b) throughput.

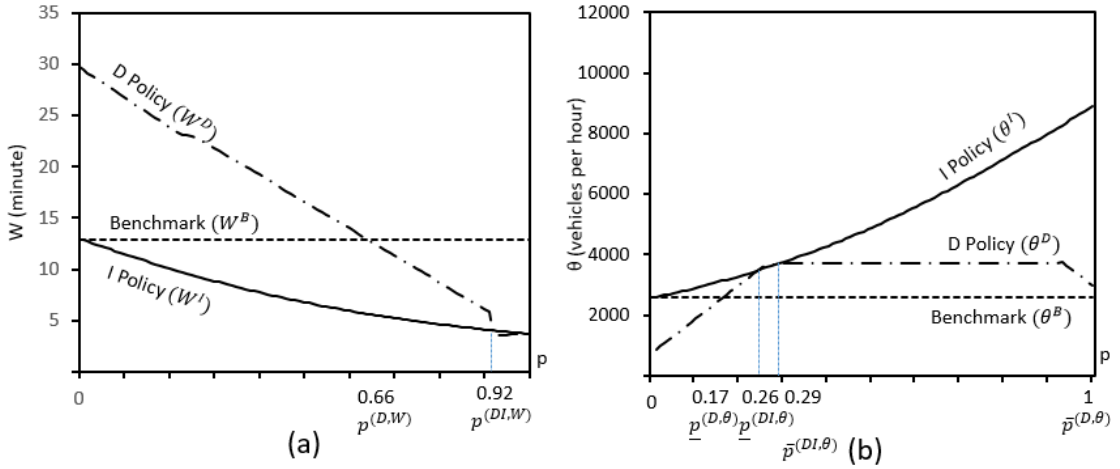


Figure A.22: A comparison between the D policy and the I policy for Case 10 when $\frac{1}{\xi^{DA}} = 1.1$ seconds and $\lambda = 11,342$ vehicles per hour: (a) mean travel time, and (b) throughput.

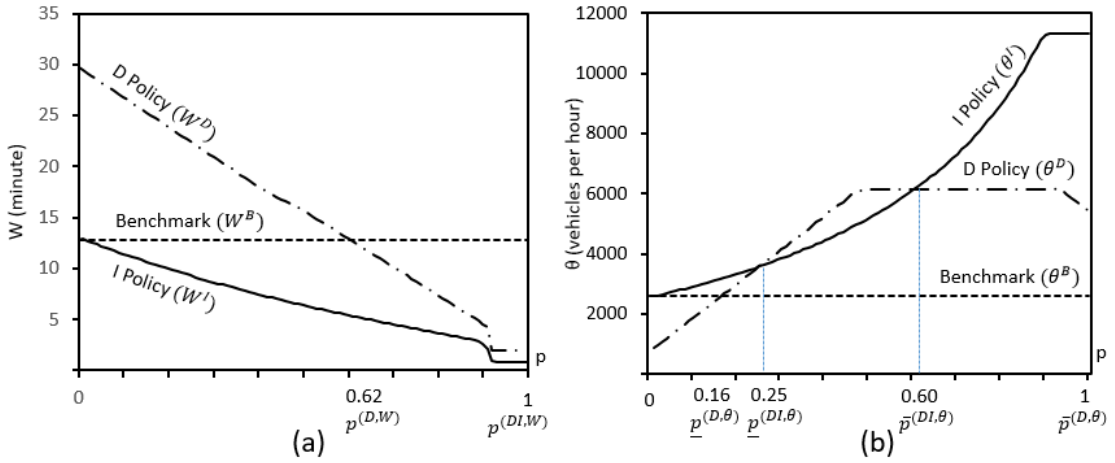


Figure A.15: A comparison between the D policy and the I policy for Case 3 when $\frac{1}{\delta^{DA}} = 20$ and $\lambda = 11,342$ vehicles per hour: (a) mean travel time, and (b) throughput.

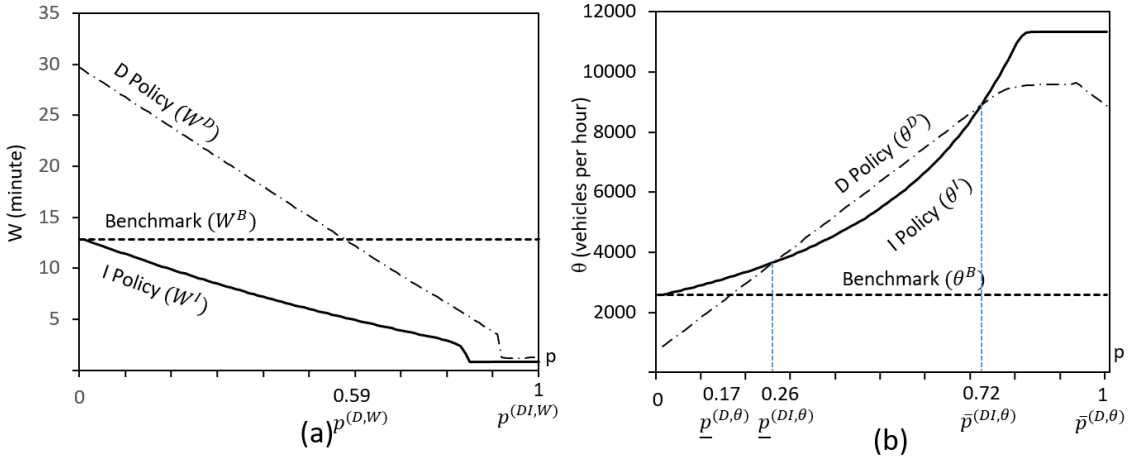


Figure A.16: A comparison between the D policy and the I policy for Case 4 when $\frac{1}{\xi^{DA}} = \frac{1}{\xi^{IAA}} = \frac{1}{\xi_n^{IHA}} = 0.1$ seconds and $\lambda = 11,342$ vehicles per hour: (a) mean travel time, and (b) throughput.

A.7 Simulation

We create a discrete-event agent-based simulation for a multi-lane highway. The goal of this simulation is to verify the validity of our queueing model and analysis.

Description of Simulation Algorithm

We develop a discrete-event cellular automata simulation (DECAS). This approach combines two methods, Discrete-event simulation (DES) and cellular automata (CA), used in prior literature. DES is commonly used to simulate queueing models (e.g., Law et al. (2000) and Ross (2006)). CA models are capable of explicitly representing individual vehicle interactions and relating these interactions to macroscopic traffic flow metrics, such as mean travel time and throughput (e.g.,

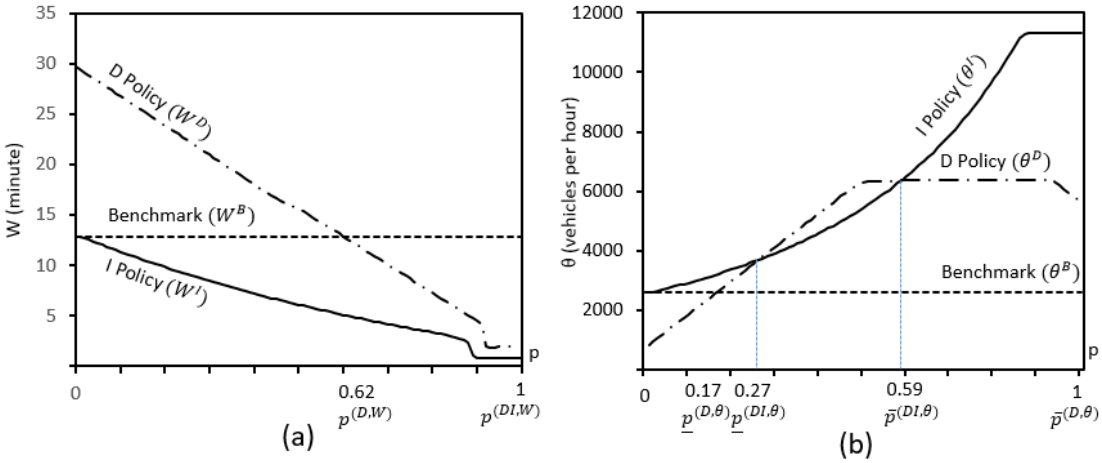


Figure A.17: A comparison between the D policy and the I policy for Case 5 when $\frac{1}{\psi_n^{DA}}$ and $\frac{1}{\psi_n^{IAA}}$ are equal to half of the safe stopping time and $\lambda = 11,342$ vehicles per hour: (a) mean travel time, and (b) throughput.

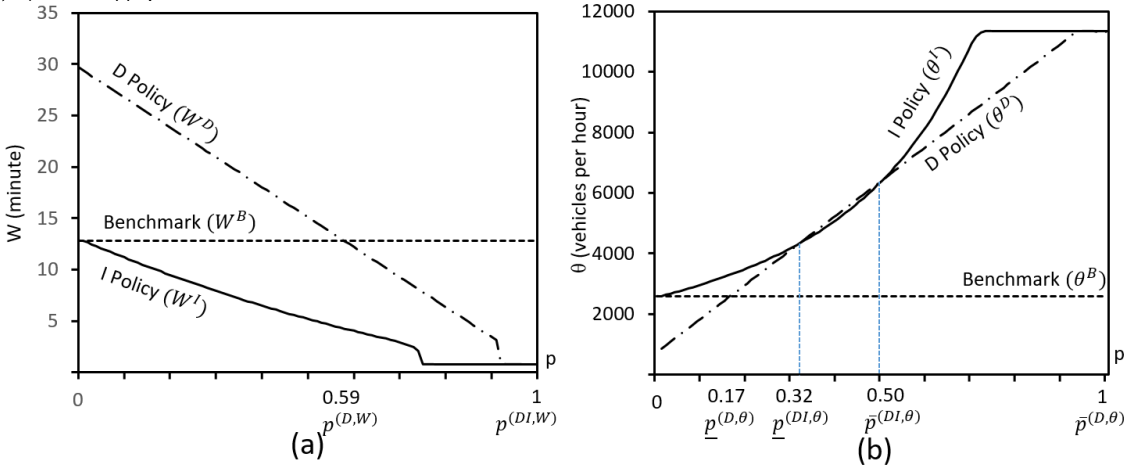


Figure A.18: A comparison between the D policy and the I policy for Case 6 when $\lambda = 11,342$ vehicles per hour: (a) mean travel time, and (b) throughput.

Benjaafar et al. 1997). Thus, DECAS is appropriate for simulating a queueing model of traffic flow in our setting.

In our DECAS model, a highway is modeled as a grid. Each cell of the grid can be occupied by at most one vehicle. In a typical CA model, the state of the system (i.e., speed and location of vehicles that are present on the grid) evolves according to a predefined set of rules at every time step (usually one second). Instead of such a discrete-time simulation, we employ a discrete-event simulation by updating the state of the system when one of the following "events" happens: (i) arrival of a new vehicle to the highway segment ("arrival" in short) and (ii) departure of an existing vehicle from the highway segment ("departure" in short). This approach significantly improves the

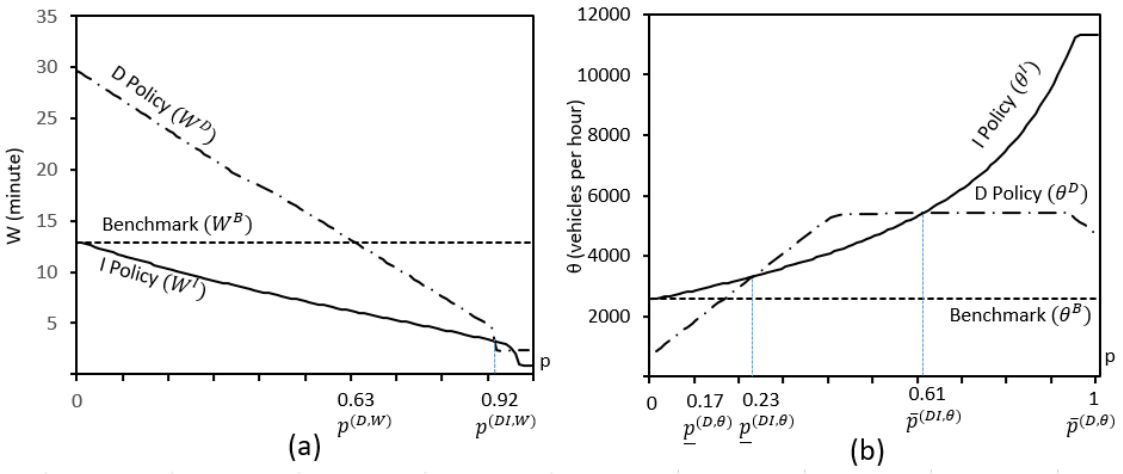


Figure A.19: A comparison between the D policy and the I policy for Case 7 when $\frac{1}{\xi_n^{IAH}} = 2.2$ seconds and $\lambda = 11,342$ vehicles per hour: (a) mean travel time, and (b) throughput.

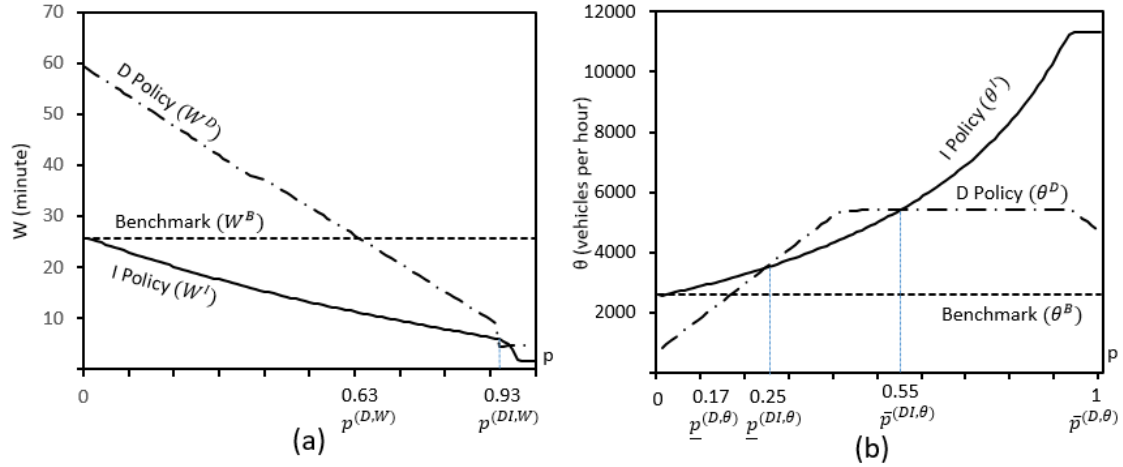


Figure A.20: A comparison between the D policy and the I policy for Case 8 when $L = 2$ miles and $\lambda = 11,342$ vehicles per hour: (a) mean travel time, and (b) throughput.

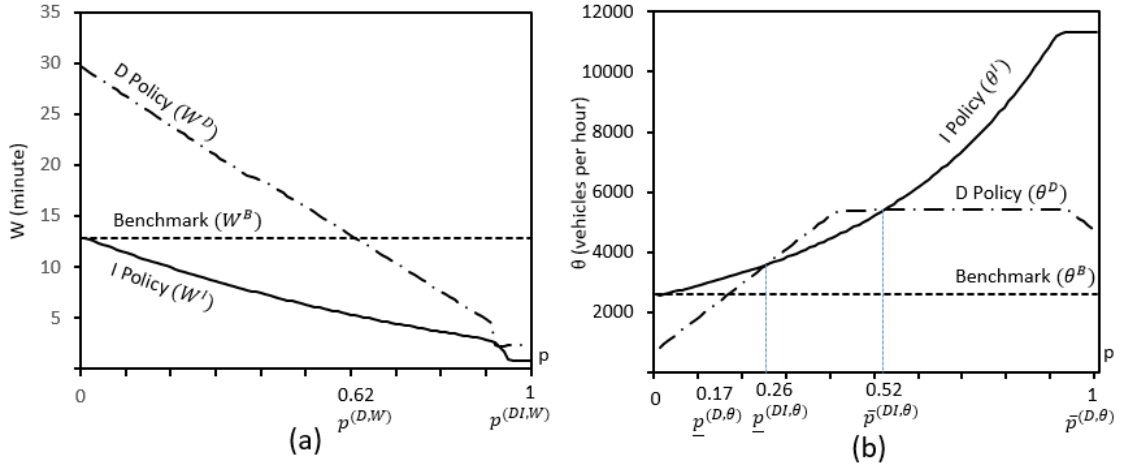


Figure A.21: A comparison between the D policy and the I policy for Case 9 when $\frac{1}{\xi_n^{IHA}} = 0.55$ seconds and $\lambda = 11,342$ vehicles per hour: (a) mean travel time, and (b) throughput.

speed of large-scale simulation in our setting.

As compared to the calibrated analytical model presented in §4, our simulation model incorporates the following general features:

- Length of the highway segment $L > 1$ (whereas we use the normalized length $L = 1$ in the numerical analysis of the main body).

- Lane changing: If the speed of a vehicle is higher than the vehicle immediately in front, the vehicle is allowed to leave its current platoon (of which the size can be one or larger) and to create a new platoon in an adjacent lane or merge into the existing platoon in an adjacent lane, if the following conditions hold. First, there is enough space (e.g., at least one empty cell) between this vehicle and the vehicle immediately in front in the adjacent lane. Second, the gap between this vehicle and the vehicle immediately behind in the adjacent lane is so high that, if the vehicle behind travels at the maximum speed of the highway, the advancement of this vehicle is smaller than the gap.

- Platoon formation process: As a vehicle arrives to the highway, it decides whether or not to join the existing platoon immediately in front. A vehicle can also leave its current platoon and create a new platoon (or merge into another platoon), as it changes its lane.

- Transient behavior of vehicles: The simulation model allows speed-up (i.e., the ability of a vehicle to increase its speed if there is a large gap between this vehicle and the vehicle immediately in front) and speed-down (i.e., the ability of a vehicle to reduce its speed to reduce chances of colliding with the vehicle in front).

Our simulation implicitly includes two more features of traffic flows: the negative effect of lane-changing on speed, and mergers of platoons. First, right after a vehicle changes its lane, if its speed is lower than that of the vehicle immediately behind, the vehicle immediately behind either reduces its speed to match the speed of the vehicle in front or changes its lane. Second, when the speed of a vehicle is higher than that of the vehicle immediately in front, and the vehicle is not able to change its lane, this vehicle is forced to reduce its speed to match the speed of vehicle immediately in front; then the follower vehicle merges into the platoon immediately in front.

Simulation Settings

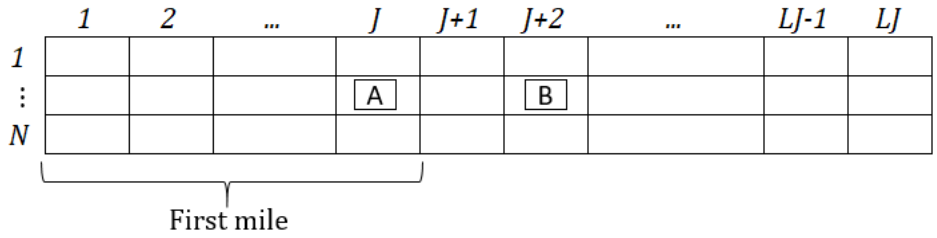


Figure A.23: Highway segment grid

As Figure A.23 depicts, we model an N-lane highway segment of length L with jam density J as a grid of NLJ cells. A vehicle occupies one cell of the grid, and moves from one cell to another, as it moves from left to right. Here we describe the simulation mechanism for the benchmark model. Similar mechanisms are used for the D and I policies.

As we mentioned before, there are two events in our simulation: arrival and departure. In the event when a new vehicle arrives to the highway, one of the following two scenarios occurs. If all cells in column 1 through column NJ (i.e., cells on the first mile of the segment) are occupied, the vehicle is blocked.[5] Otherwise, it occupies the first empty cell in column 1 through column NJ. With probability $1-\delta^B$ (where $1/\delta^B$ denotes the mean platoon size of HVs as in the base model), this vehicle joins the last platoon ahead, and with probability δ^B, it starts a new platoon. If it joins the last platoon, its speed is set equal to the speed of this platoon.[6] Otherwise, this vehicle starts a new platoon, and the speed of this vehicle is determined by generating a random number from a truncated normal distribution with mean V_n^B and standard deviation σ miles per hour.[7] In the event when a vehicle departs from the highway segment, we remove this vehicle from the highway grid. A vehicle is eligible for departing from the highway when it is at the head of a lane, i.e., it is the rightmost vehicle in a lane, and it can travel the remaining cells before other vehicles. If there is a tie among the rightmost vehicles in the lanes, we allow all of them to depart the highway

[5]The first mile of the segment is used as warm-up state to place vehicles on the segment and form platoons. Suppose the first unoccupied cell exists in lane l, column k. If k is higher than 1, we move all the vehicles in lane l before column k forward by one cell, and place the new vehicle in lane l, column 1. Note that the blocking rule used in the simulation is more stringent than that in the queueing model: whereas a vehicle is blocked in the queuing model when there are NJL vehicles on the highway segment, a vehicle is blocked in the simulation when there are NJ vehicles on the first mile of the highway segment, even though there may exist empty cells between the $(J+1)^{th}$ and the LJ^{th} columns. Using the same blocking rule as the queueing model, requires assigning an arriving vehicle to an empty cell that is possibly located at the end of the segment. Thus, we use a more stringent blocking rule to ensure that all vehicle start service at the beginning of the segment.

[6]This assumption guarantees the same speed for all vehicles in one platoon.

[7]To determine σ, we calculate 95% confidence intervals of the estimated V_n^B for $n = 1, 2, \cdots, NJL$ in §4. Next, we set σ equal to half of the maximum width of these confidence intervals.

Notation	Definition	Notation	Definition
$V(i,l)$	Speed of the i^{th} vehicle in the l^{th} lane	V_{max}	Free-flow speed of the highway
$x(i,l)$	Column number of the i^{th} vehicle in the l^{th} lane	$g(i,l)$	Gap between the i^{th} and the $(i+1)^{th}$ vehicle (i.e., the vehicle immediately in front) in the l^{th} lane, which is defined as $x(i+1,l) - x(i,l) - 1$, $g(i,l) = 0$ if the i^{th} and the $(i+1)^{th}$ vehicles are in two consecutive cells.
$d_{front}(i,l)$	Distance between the i^{th} vehicle in the l^{th} lane and the closest vehicle in front of this vehicle in the adjacent lane, which is set to $J-i$ if there is no vehicle in front.	$d_{back}(i,l)$	Distance between the i^{th} vehicle in the l^{th} lane and the closest vehicle behind of the i^{th} vehicle in the adjacent lane, which is set to $i-1$ if there is no vehicle behind.
d_{up}	Minimum gap between two vehicles that allows the follower vehicle to increase its speed.	a	Acceleration in meters per second squared
V_{front}	Speed of the front vehicle in the adjacent lane	$V_{right}(l)$	Speed of the rightmost vehicle in the l^{th} lane
$x_{right}(l)$	Location of the rightmost vehicle in the l^{th} lane	$i_{l,right}$	Column number of the rightmost vehicle in lane l
$i_{l,left}$	Column number of the leftmost vehicle in lane l	Ł$_{depart}$	The set of lane(s) from which the next departure(s) will take place, i.e., $_l\{\frac{J-x_{right}(l)}{JV_{right}(l)}\}$.
n_{depart}	Number of simultaneous departures, i.e., len(Ł$_{depart}$)	t_A	Time until next arrival
t_D	Time until next departure	t_E	Time until next event (arrival or departure), which is equal to $\min\{t_A, t_D\}$

Table A4: Table of additional notation used in the simulation algorithm.

simultaneously, i.e., there can be up to N simultaneous departures. When one of these two events occurs, it triggers the movement of the existing vehicles on the highway segment. These vehicles move according to one of the following scenarios:

- Suppose the gap between a vehicle 'A' and the vehicle 'B' immediately in front allows vehicle A to maintain its speed without a collision with vehicle B. In this scenario, vehicle A advances according to its speed and the time since the last event.

- Suppose the gap between a vehicle 'A' and the vehicle 'B' immediately in front is so large that vehicle A is able to speed up. In this scenario, if vehicle A and vehicle B belong to the same platoon (resp., two different platoons), vehicle A increases its speed – at a specified acceleration – so that the headway between vehicle A and vehicle B is at least equal to the mean intraplatoon headway (resp., mean interplatoon headway).

- Suppose the gap between a vehicle 'A' and the vehicle 'B' immediately in front is so small

that, if the vehicle maintains its speed, a collision happens. In this case, if the speed of the vehicle immediately in front in an adjacent lane is higher than the speed of vehicle A, vehicle A changes its lane and moves to the adjacent lane, contingent on fulfillment of lane-changing conditions, discussed previously.[8] Otherwise, vehicle A moves to the cell behind vehicle B, while reducing its speed to the speed of vehicle B.

Detailed Simulation Algorithm

We define the additional notation that we use in the simulation in Table A4, and then present the algorithm used in our simulation of the benchmark model with no AVs. Similar algorithms are used in simulating the HV and AV queues of the D policy as well as the I policy.[9]

Algorithm

Initialization

Initiate t_A by generating a random number from an exponential distribution with rate λ.

Set t_D equal to ∞.

Determining the event type

if $t_A \leq t_D$ (i.e., the event is an arrival) **then**

  if Column 1 through column NJ of the segment are occupied **then**

    Block the new vehicle.

[8] If there are two adjacent lanes on both sides of vehicle A, this vehicle tries moving to its right lane only when moving to its left lane is not feasible.

[9] In the simulation, if there is a tie among lanes, we choose the lane with the lower number.

else

Increase n by one, and assign this vehicle to the first empty cell in column 1 through NJ of lane 1 through N.

Store the assigned location of this vehicle as $x(1, l)$.

Generate a uniform random number r.

if $r \geq \delta$ and there exists at least one platoon in lane l **then**

Add this vehicle to the last platoon in lane l by setting $V(i, l)$ equal to $V(i + 1, l)$.

else

Determine $V(i, l)$ by generating a random number from a truncated normal distribution with mean V_n^B and standard deviation σ.

else

Decrease n by $n_{depart} = len(L_{depart})$, and remove the rightmost vehicle(s) from $l \in L_{depart}$.

for lane $l \in \{1, 2, \cdots, N\}$ **do**

for vehicle $i \in \{i_{l,right}, \cdots, i_{l,left}\}$ **do**

Updating speed and location

if $JV(i, l)t_E \leq g(i, l)$ (i.e., the advancement of vehicle i is less than the gap between this vehicle and the vehicle in front) **then**

if $g(i, l) - JV(i, l)t_E \geq d_{up}$ **then**

if the i^{th} and $(i + 1)^{th}$ belong to the same platoon **then**

Set $V(i, l)$ equal to $min\{[g(i, l)/J - V_n^B \xi^B]/t_E, at_E + V(i, l), V_{max}\}$, and $x(i, l)$ equal to $x(i, l) + round(JV(i, l)t_E)$. Update $g(i, l)$ and $g(i - 1, l)$ accordingly.

else

Set $V(i, l)$ equal to $min\{[g(i, l)/J - V_n^B \psi_n^B]/t_E, at_E + V(i, l), V_{max}\}$, and $x(i, l)$ equal to $x(i, l) + round(JV(i, l)t_E)$. Update $g(i, l)$ and $g(i - 1, l)$ accordingly.

else

Advance the vehicle by $round(JV(i, l)t_E)$ cells. Update $g(i, l)$ and $g(i-1, l)$ accordingly.

else

Lane changing

Look ahead:

if $d_{front} \geq 1$ and $V_{front} > V(i,l)$ **then**

Look backward:

if $d_{back} \geq Jt_E V_{max}$ **then**

Let l' be the new lane that vehicle i moved to, and i' be the number of this vehicle in lane l'. Set $V(i',l')$ and $x(i',l')$ equal to V_{front} and $x(i,l)+1$, respectively.

Remove $V(i,l)$ and $x(i,l)$.

else

Set $V(i,l)$ equal to $V(i+1,l)$, and set $x(i,l)$ equal to $x(i+1,l)-1$. Let $g(i,l)=0$, and update $g(i-1,l)$ accordingly.

else

Set $V(i,l)$ equal to $V(i+1,l)$, and set $x(i,l)$ equal to $x(i+1,l)-1$. Let $g(i,l)=0$, and update $g(i-1,l)$ accordingly.

Parameter update

Update t_D by $\min_l\{\frac{J-x_{right}(l)}{JV_{right}(l)}\}$, and set l_{depart} and n_{depart} equal to $_l\{\frac{J-x_{right}(l)}{JV_{right}(l)}\}$ and $len(L_{depart})$, respectively.

Update t_A by generating a random number from an exponential distribution with rate λ.

Simulation Results

As in our calibrated model in §4, we run the simulation for a highway segment with three lanes, and a jam density of 185 vehicles per mile per lane. We consider a segment of four miles: we use the first mile as warm-up state to place vehicles on the segment and form platoons, and use the results from the remaining three miles. The arrival rate to this highway segment is $11,342$ vehicles per hour. We set the minimum gap between two vehicles required for a following vehicle to increase its speed, d_{up}, equal to 20 cells, which is equivalent to 0.1 miles. The acceleration a is set equal to 2 meters per second squared, which is the acceleration value used by Liu et al. (2018).[10]

To investigate the performance of the D and I policies, for each value of $p \in \{10, 20, \cdots, 100\}$ we run the simulation 30 times for a time horizon of four hours. Since we assume the highway is

[10]Performing a sensitivity analysis on the values of d_{up} and a, we observe that all of our main results are robust.

empty at the beginning of the simulation horizon, we use the results of the last hour to ensure that the simulation is in steady state. Figures A.24(a) and (b) represent a comparison between the mean travel time W of the D policy and that of the I policy in the simulation and the queueing model, respectively. As these figures illustrate, W in the simulation is close to, but slightly worse than, that in the queueing model (e.g., $W^B = 13.8$ minutes per mile in the simulation versus $W^B = 12.8$ minutes per mile in the queueing model). This is likely because any fractional advancements of vehicles are rounded down as we update the location of vehicles in the simulation; for example, if a vehicle is able to move 2.5 cells according to its speed, it moves only two cells to ensure it will not collide with a vehicle in front of it (that, for example, may only have been able to move two cells). In addition, Figures A.24(a) and (b) show that, similar to our numerical results in §5, for any given value of p, the I policy outperforms the benchmark model in terms of mean response time W. When the AV proportion p is at least 64% (i.e., $p^{(D,W)} = 0.64$), the D policy is also capable of reducing W over that of the benchmark model. This value of $p^{(D,W)}$ is very close to that in the base case, i.e., $p^{(D,W)} = 0.63$. As in the queueing model, the I policy leads to a lower W than the D policy for all values of p.

Figures A.24(c) and (d) represent a comparison between throughput θ of the D policy and that of the I policy in the simulation and the queueing model, respectively. As these figures depict, θ in the simulation is lower than that in the queueing model, especially under the D and I policies when the AV proportion p is not high. This discrepancy likely arises due to the more stringent blocking rule in the simulation than that in the queueing model: whereas a vehicle is blocked in the queueing model when the entire highway segment is full, a vehicle is blocked in the simulation when the first mile of the segment is full. This results in a lower effective arrival rate (i.e., the arrival rate of vehicles that are not blocked) in the simulation than that in the queueing model. Under the D policy, $\theta^D(p)$ in the simulation shows a similar pattern to that of $\theta^D(p)$ in the queueing model. Although $\theta^D(0) = 716$ vehicles per hour in the simulation is close to $\theta^D(0) = 740$ vehicles per hour in the base model, $\theta^D(p)$ in the simulation plateus at a lower value of p than it does in the queueing model; again, likely due to the more stringent blocking rule used in the simulation. Similarly, under the I policy we observe that the simulated throughput $\theta^I(p)$ is increasing in p. In this case, when all vehicles are AVs (i.e, $p = 1$), no vehicle is blocked and $\theta^I(1)$ becomes equal to the arrival rate λ to the highway segment. Furthermore, Figures A.24(c) and (d) illustrate that the

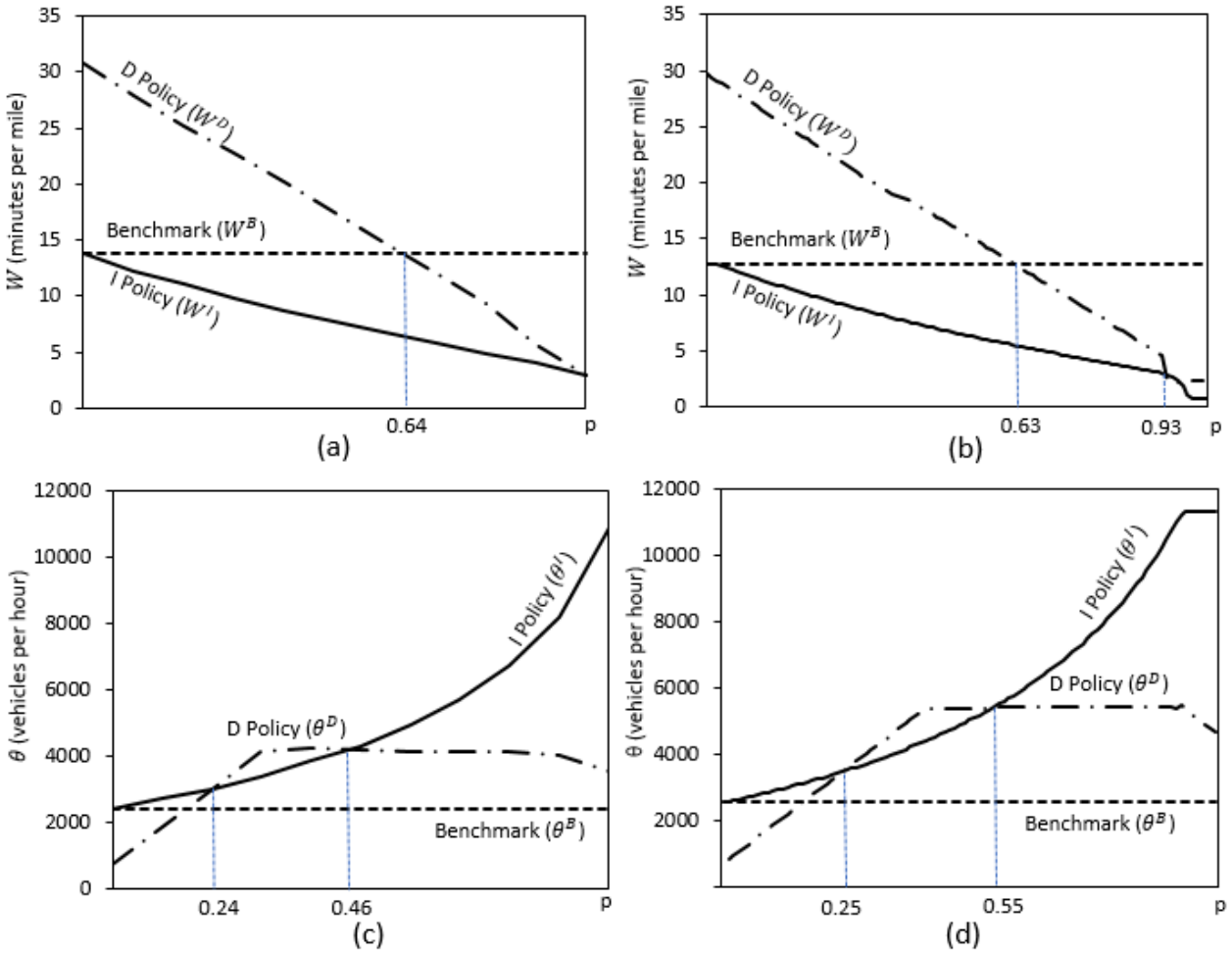


Figure A.24: A comparison between the D policy and the I policy when $\lambda = 11,342$ vehicles per hour: (a) mean travel time of the simulation, (b) mean travel time of the queueing model, (c) throughput of the simulation, and (d) throughput of the queueing model.

interval of p in which $\theta^D(p)$ is higher than $\theta^I(p)$ is $[0.24, 0.46]$, which is similar to, but smaller than that in the queueing model, $[0.25, 0.55]$. This shows that the stringent blocking rule used in the simulation affects the D policy more than the I policy, because the I policy blocks fewer vehicles that the D policy by using the capacity of all of the lanes, i.e., the "pooling effect."

Appendix B

Additional Material for Chapter 2

B.1 Notation Summary

Table B1: Summary of notation

Symbol	Definition
j	Index for parking area ($j \in \{1, 2\}$)
x	Departure time from H
x_{max}	Latest departure time from H
x_j	Earliest departure time of commuters who choose Area $j \in \{1, 2\}$
$x_{j,max}$	Latest departure time of commuters who choose Area $j \in \{1, 2\}$
x_D	Earliest departure time from H for commuters who choose Area 2 and experience downtown congestion
y	Drop-off time at W for commuters who leave H at time x
y_{max}	Latest Drop-off time at W
y_j	Earliest Drop-off time at W for commuters who choose Area $j \in \{1, 2\}$
$y_{j,max}$	Latest Drop-off time at W for commuters who choose Area $j \in \{1, 2\}$
y_D	Earliest Drop-off time at W of commuters who choose Area 2 and experience downtown congestion
R_I	Inbound bottleneck capacity
R_W	Drop-off rate at W
R_W^*	Optimal drop-off rate at W
$R_D(y)$	Travel rate in D for a commuter who leaves H at time x
$R_{D,j}(y)$	Travel rate in D for a commuter who leaves H at time x and chooses Area $j \in \{1, 2\}$
M	Maximum travel rate in D, i.e., $R_D(0)$
θ	Weight of number of AVs present in D in the equation for $R_D(y)$
a	The difference between the drop-off capacity and the maximum travel rate in D
N	Total number of commuters
T	Official start time at work for all commuters
$p_j(x)$	Parking fee in Area $j \in \{1, 2\}$ for commuters who leave H at time x
K_j	Capacity of parking Area $j \in \{1, 2\}$
K_j^*	Optimal capacity of parking Area $j \in \{1, 2\}$
$\lambda_H(x)$	Departure rate from H at time x
$\lambda_j(y)$	Departure rate from W to Area $j \in \{1, 2\}$ for AVs that leave H at time x
$\tau_I(x)$	Inbound delay time for a commuter who leaves H at time x
$\tau_W(x)$	Drop-off congestion time for a commuter who leaves H at time x
$\tau_{D,j}(x)$	Congestion time in D for a commuter who leaves H at time x and chooses Area $j \in \{1, 2\}$
$t_{H,I}$	Free-flow travel time from H to I
$t_{I,D}$	Free-flow travel time from I to D
$t_{W,1}$	Free-flow travel time from W to 1
$t_{D,2}$	Free-flow travel time from D to 2
α	Unit cost of driving time for commuters
α'	Unit cost of driving time for AVs
β	Unit cost for early schedule delay
γ	Unit cost for late schedule delay
$C_j(x)$	Total cost of a commuter who leaves H at time x and chooses Area $j \in \{1, 2\}$
n	The parameter that counts the number of times that the social planner switches from routing AVs to Area 2 to routing them to Area 1.
$\epsilon_1(x)$	Non-negative value added to the parking fee of Area 1
$\epsilon_2(x)$	Non-negative value added to the downtown congestion toll

B.2 Model Calibration

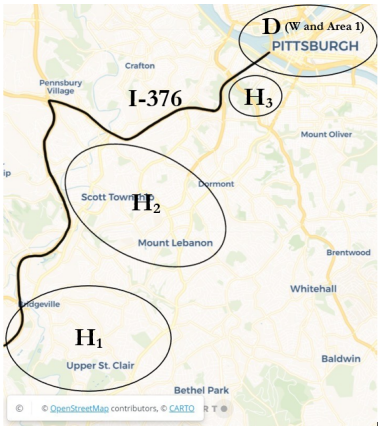
We calibrate our model to data from the Pittsburgh Metropolitan Area. Table 2.1 summarizes our calibrated parameter values, and Figure B.1 illustrates this area. Our parameter values are based on the data from Pi et al. (2019) and our own estimation. The following parameter values are from Pi et al. (2019):

- There are $N = 20,000$ morning commuters from H_1 who use their personal vehicles to travel on highway I-376. This route is known as the southwest commuting corridor to Downtown Pittsburgh (D).
- I-376 has two lanes entering D, each of them has a capacity of $2,300$ vehicles per hour. Thus, the inbound bottleneck capacity R_I is $4,600$ vehicles per hour.
- The free-flow travel time from W to Area 1, $t_{W,1}$, is equal to 2 minutes. Both Area 1 and W are located inside D.
- We consider two possible locations of Area 2. In our base case, Area 2 is located in H_2 with the free-flow travel time from D to Area 2, $t_{D,2} = 15$ minutes and zero parking fee, i.e., $p_2(x) = 0$ for $x \in [0, x_{max}]$. We also consider another case where Area 2 is located in H_3 with $t_{D,2} = 5$ minutes and $p_2(x) = p_1(x)/2$ for $x \in [0, x_{max}]$.
- The daily parking price in Downtown Pittsburgh varies from \$1.80 to \$16. So we let $p_1(0) =$ \$1.80 and $p_1(x_{max}) =$ \$16.
- The capacity K_1 of Area 1 is equal to $10,000$ vehicles.
- The early arrival cost β is estimated at \$3.90 per hour.

We estimate the remaining model parameters as follows:

- We assume the passengered travel cost α is \$4.50 per hour. This is lower than the value of α (= \$6.4 per hour) for HVs estimated by Pi et al. (2019), because AV owners do not need to drive and are able to spend their commute time on other activities (e.g., sleeping, reading, etc.) that give positive utility. In addition, we assume the passengerless travel cost α' is half

Figure B.1: (Color in PDF File) An Illustration of Downtown and Southern Region of Pittsburgh.



Notes. This figure is a modified version of Figure 8(b) in Pi et al. (2019).

of the passengered travel cost, i.e., $\alpha' = \alpha/2 = \$2.25$ per hour. These parameter values are rough estimates, so we perform sensitivity analyses on the values of α and α' in Appendix B.4.1.

- We assume Area 1's parking fee is equal to $p_1(x) = 14.20x/x_{max} + 1.80$, so the minimum and maximum parking fees match the corresponding values in Pi et al. (2019), i.e., $p_1(0) = \$1.80$ and $p_1(x_{max}) = \$16$. This assures that commuters who leave H early take cheaper parking spots, and commuters who leave H late are left with more expensive parking spots.

- To determine the curbside drop-off capacity, we measure the total curbside length in Downtown Pittsburgh, which is about 0.6 miles for the southwest corridor commuters. Assuming there can be up to 50 drop-offs per mile and each drop-off takes 30 seconds, the drop-off capacity R_W is equal to $0.6 \times 50 \times 3,600/30 = 3,600$ drop-offs per hour (see Barth (2019) for more details).

- To characterize the travel rate in D, $R_{D,2}(y)$, we assume $\theta = 1$ and $M = 59.8$. This results in the following expression for $R_{D,2}(y)$: $\lambda_2(y) - a\theta e^{\theta y} = \lambda_2(y) - 0.2e^y$, where $a = (R_W - M)/\theta$ (See Lemma B.1 for further details). This is a decreasing function of y, which means that as D gets more congested, the exit rate from D decreases.

B.3 Proofs

Lemma B.1. *If $\tau_W(x) \neq 0$ for all $x \in [0, x_{max}]$, then $\tau_{D,2}(x) = \frac{y\int_0^y a\theta e^{\theta z}dz}{\int_0^y [\lambda_2(z) - a\theta e^{\theta z}]dz}$.*

Proof. If $\tau_W(x) \neq 0$ for all $x \in [0, x_{max}]$, then $\min\{\lambda_H(x), R_W\} = R_W$. By substituting this in the $R_D(y)$ equation from Assumption A3, we have:

$$\begin{aligned} R_D(y) &= M - \theta \int_0^y [R_W - R_D(z)]dz \\ &= M - \theta R_W y + \theta \int_0^y R_D(z)dz. \end{aligned} \tag{B.1}$$

Equation (B.1) is a first order differential equation in the form of $\theta f(y) - f'(y) - \theta R_W y + M = 0$, where $f(y) = \int_0^y R_D(z)dz$, $f'(y) = R_D(y)$, and $f(0) = 0$. By solving this first order differential equation, we have the following:

$$R_D(y) = R_W - a\theta e^{\theta y},$$

where $a = (R_W - M)/\theta (> 0)$. By Assumption A2, AVs that choose Area 1 do not experience any delay, so the travel rate of AVs that head to Area 1 at time y is equal to the number of AVs that head to Area 1 at time y, $R_{D,1}(y) = \lambda_1(y)$. Therefore, the travel rate $R_{D,2}(y)$ of AVs that go to Area 2 is equal to $R_D(y) - R_{D,1}(y) = R_W - a\theta e^{\theta y} - \lambda_1(y) = \lambda_2(y) - a\theta e^{\theta y}$.

Finally, by Assumption A5, we attain the congestion time in D for AVs that leave H at time x and choose Area 2, as follows:

$$\tau_{D,2}(x) = \frac{\int_0^y [\lambda_2(z) - R_{D,2}(z)]^+ dz}{\frac{\int_0^y R_{D,2}(z)dz}{y}} = \frac{y\int_0^y \{\lambda_2(z) - [\lambda_2(z) - a\theta e^{\theta z}]\}^+ dz}{\int_0^y [\lambda_2(z) - a\theta e^{\theta z}]dz} = \frac{y\int_0^y a\theta e^{\theta z}dz}{\int_0^y [\lambda_2(z) - a\theta e^{\theta z}]dz}.$$

□

Proof of Proposition 2.1. (a) Case (i): We first assume that both $\tau_I(x)$ and $\tau_W(x)$ are positive for $x \in (0, x_{max}]$, and then we show that our results hold even when $\tau_I(x)$ and $\tau_W(x)$ are not positive. Since in this case $\alpha' t_{W,1} + p_1(0) \geq \alpha' t_{D,2}$, by (2.2) and (2.3), $C_2(x)$ is lower than $C_1(x)$ at time $x = 0$. In this case, by Condition 1, all commuters choose Area 2, i.e., $\lambda_2(x) = R_W$ and $\lambda_1(x) = 0$. Since $\tau_{D,2}(x)$ is increasing in x, there may exist $x_1 \in [0, x_{max}]$ such that $\alpha' t_{W,1} + p_1(x_1) = \alpha'[t_{D,2} + \tau_{D,2}(x_1)]$, which means $C_1(x_1) = C_2(x_1)$. Thus, by Condition 1, commuters can choose to

park in Area 1. This means that when $x \geq x_1$, the cost associated with parking in Area 1 must be equal to that in Area 2, i.e., $\alpha' t_{W,1} + p_1(x) = \alpha'[t_{D,2} + \tau_{D,2}(x)]$. After rearranging this equation, we have $\tau_{D,2}(x) = \frac{y\int_0^y a\theta e^{\theta z}dz}{\int_0^y \lambda_2(z) - a\theta e^{\theta z}dz} = t_{W,1} + \frac{p_1(x)}{\alpha'} - t_{D,2}$, so $\lambda_2(y)$ is equal to $ae^{\theta y} + \frac{yae^{\theta y} + \int_0^y a\theta e^{\theta z}dz}{t_{W,1} - t_{D,2} + p_1(x)/\alpha'} - \frac{[yp_1'(x)/\alpha']\int_0^y a\theta e^{\theta z}dz}{[t_{W,1} - t_{D,2} + p_1(x)/\alpha']^2}$.

To derive $\lambda_H(x)$ for $x < x_1$, we substitute, $\tau_I(x)$, $\tau_W(x)$, and $\tau_{D,2}(x)$ into (2.3). By Assumption A5, since $\tau_I(x)$, $\tau_W(x) > 0$, we have $\tau_I(x) = \frac{\int_0^x [\lambda(u) - R_I]du}{R_I}$ and $\tau_W(x) = \frac{\int_0^{x+\tau_I(x)} [R_I - R_W]du}{R_W}$. In addition, since $\lambda_2(y) = R_W$, by Lemma B.1 $\tau_{D,2}(x) = \frac{y\int_0^y ae^{\theta z}dz}{\int_0^y \lambda_2(z) - ae^{\theta z}dz} = \frac{y\int_0^y ae^{\theta z}dz}{\int_0^y R_W - ae^{\theta z}dz}$. Hence, $C_2(x) = (\alpha - \beta)\frac{\int_0^y \lambda_H(z)dz}{R_W} - \alpha x + \beta T + \alpha'\{t_{D,2} + \frac{y\int_0^y ae^{\theta z}dz}{\int_0^y R_W - ae^{\theta z}dz}\}$, where $y = x + \tau_I(x) + \tau_W(x) = \frac{\int_0^x \lambda_H(u)du}{R_W}$. By Condition 2, regardless of departure time x, the total travel cost of all commuters must be equal, i.e., $\frac{\partial C_2}{\partial x} = (\alpha - \beta)\lambda_H(x)/R_W - \alpha + \alpha'\frac{\partial \tau_{D,2}(x)}{\partial x} = 0$. After rearranging this equation, we have $\lambda_H(x) = \frac{\alpha - \alpha'\frac{\partial \tau_{D,2}(x)}{\partial x}}{\alpha - \beta} R_W$, where $\frac{\partial \tau_{D,2}(x)}{\partial x} = \frac{\partial \tau_{D,2}(x)}{\partial y} \times \frac{\partial y}{\partial x} = \frac{R_W y^2 a\theta e^{\theta y} - (\int_0^y a\theta e^{\theta z}dz)^2}{(\int_0^y R_W - a\theta e^{\theta z}dz)^2} \times \frac{\lambda_H(x)}{R_W}$. Thus, $\lambda_H(x) = \frac{\alpha(\int_0^y R_W - a\theta e^{\theta z}dz)^2}{(\alpha - \beta)(\int_0^y R_W - a\theta e^{\theta z}dz)^2 + \alpha' R_W y^2 a\theta e^{\theta y} - \alpha'(\int_0^y a\theta e^{\theta z}dz)^2}$. For $x \geq x_1$, we derive $\lambda_H(x)$ using equation (2.2). By Condition 2, regardless of departure time x, the total travel cost of all commuters must be equal, i.e., $\frac{\partial C_1}{\partial x} = (\alpha - \beta)\frac{\lambda_H(x)}{R_W} - \alpha + p'(x) = 0$. After rearranging this equation, for $x \in [x_1, x_{max}]$, we have $\lambda_H(x) = \frac{\alpha - p_1'(x)}{\alpha - \beta} R_W$.

Next, we show that our results hold even when $\tau_I(x)$ and $\tau_W(x)$ are not positive. First, if $\tau_I(x) = 0$ for some $x \in (0, x_{max}]$ and $\tau_W(x) > 0$ for all $x \geq 0$, all of our results hold. This happens because $y = x + \tau_W(x)$ is still equal to $\frac{\int_0^x \lambda_H(u)du}{R_W}$, and $C_2(x)$ is equal to $C_2(x) = (\alpha - \beta)\frac{\int_0^x \lambda_H(u)du}{R_W} - \alpha x + \beta T + \alpha'[t_{D,2} + \tau_{D,2}(x)]$, and the rest of the proof remains unchanged. Next, let $\hat{x}$ be the smallest $x \in [0, x_{max}]$ such that $\tau_W(x) = 0$. Since $R_W \leq R_I$, $\tau_I(\hat{x})$ is also equal to zero. In this case, $C_2(\hat{x}) = \beta(T - \hat{x}) + \alpha'[t_{D,2} + \tau_{D,2}(\hat{x})]$. By setting the derivative of $C_2(\hat{x})$ equal to zero, i.e., $\frac{\partial C_2}{\partial \hat{x}} = \beta + \alpha'\frac{\partial \tau_{D,2}(\hat{x})}{\partial \hat{x}} = 0$, we find that $\frac{\partial \tau_{D,2}(\hat{x})}{\partial \hat{x}}$ must be equal to β/α'. We show that this is consistent with our result for $\lambda_H(x)$ by contradiction. First, we assume that $\frac{\partial \tau_{D,2}(\hat{x})}{\partial \hat{x}} < \beta/\alpha'$. In this case, $\frac{\alpha - \alpha'\frac{\partial \tau_{D,2}(\hat{x})}{\partial \hat{x}}}{\alpha - \beta} > 1$, and hence, $\lambda_H(x) > R_W$, which is a paradox; since $\tau_W(x) = 0$, $\lambda_H(x)$ cannot exceed the drop-off capacity R_W. Next, we assume that $\frac{\partial \tau_{D,2}(\hat{x})}{\partial \hat{x}} > \beta/\alpha'$, which leads to $\lambda_H(x) < R_W$. In this case, compared to a commuter who leaves H at time $\hat{x}$, a commuter who leaves at time $\hat{x} + 1$ incurs $\$\beta$ more in work schedule penalty, but saves $\$\alpha'\frac{\partial \tau_{D,2}(\hat{x})}{\partial \hat{x}}$ in downtown congestion cost. Since we assume $\frac{\partial \tau_{D,2}(\hat{x})}{\partial \hat{x}} > \beta/\alpha'$, this commuter is better off departing H at time $\hat{x}$ instead of time $\hat{x} + 1$, so she can unilaterally change her decision without increasing any other commuter's cost. This means that there is no UE in this case, which is a paradox. Thus,

$\frac{\partial \tau_{D,2}(\hat{x})}{\partial \hat{x}} = \beta/\alpha'$ and our results still hold.

Lastly, for case (i) of UE1 to hold, we need to assume that Area 1 can accommodate all commuters who choose this area, i.e., $K_1 \geq \int_{y_1}^{y_{max}} \lambda_1(u)du = N - \int_0^{y_{max}} \lambda_2(u)du = N - \int_0^{y_1} \lambda_2(u)du - \int_{y_1}^{y_{max}} \lambda_2(u)du = N - R_W y_1 - \int_{y_1}^{y_{max}} C du = N - R_W y_1 - a(e^{\theta(N/R_W - y_1)} - 1)[1 + \frac{N/R_W}{t_{W,1} - t_{D,2} + p_1(x_{max})/\alpha'}]$. If this condition does not hold, i.e., case (ii) of UE1 when $K_1 < N - R_W y_1 - a(e^{\theta(N/R_W - y_1)} - 1)[1 + \frac{N/R_W}{t_{W,1} - t_{D,2} + p_1(x_{max})/\alpha'}]$, after Area 1 becomes full at time $y_{1,max}$, commuters have to choose Area 2. The proof of this part is very similar to that of case (i), and hence it is omitted.

(b) Case (i): We first assume that both $\tau_I(x)$ and $\tau_W(x)$ are positive for $x \in (0, x_{max}]$, and then we show that our results hold even when $\tau_I(x)$ and $\tau_W(x)$ are not positive. When $\alpha' t_{W,1} + p_1(0) \leq \alpha' t_{D,2}$, then by (2.2) and (2.3) $C_2(x)$ is higher than $C_1(x)$ at time 0. In this case, by Condition 1, all commuters choose Area 1, i.e., $\lambda_1(x) = R_W$ and $\lambda_2(x) = 0$. To derive $\lambda_H(x)$, we substitute $\tau_I(x) = \frac{\int_0^x [\lambda(u) - R_I]du}{R_I}$, and $\tau_W(x) = \frac{\int_0^{x+\tau_I(x)} [R_I - R_W]du}{R_W}$ from Assumption A5 into (2.2) as follows:

$$C_1(x) = (\alpha - \beta)\frac{\int_0^x [\lambda_H(u) - R_I]du}{R_I} + (\alpha - \beta)\frac{(R_I - R_W)(x + \tau_I(x) - 0)}{R_W} + \beta(T - x) + \alpha' t_{W,1} + p_1(x). \tag{B.2}$$

By substituting $x + \tau_I(x)$ and $x + \tau_I(x) + \tau_W(x)$ with $\frac{\int_0^x \lambda_H(u)du}{R_I}$ and $\frac{\int_0^x \lambda_H(u)du}{R_W}$, respectively, we can simplify (B.2) as follows: $C_1(x) = (\alpha - \beta)\frac{\int_0^x \lambda_H(u)du}{R_W} - \alpha x + \beta T + \alpha' t_{W,1} + p_1(x)$. By Condition 2, regardless of departure time x, the total travel cost of all commuters must be equal, i.e., $\frac{\partial C_1}{\partial x} = (\alpha - \beta)\frac{\lambda_H(x)}{R_W} - \alpha + p_1'(x) = 0$. After rearranging this equation, we have $\lambda_H(x) = \frac{\alpha - p_1'(x)}{\alpha - \beta} R_W$.

Since $p_1(x)$ is increasing in x, there may exist $x_2 \in [0, x_{max}]$ such that $\alpha' t_{W,1} + p_1(x_2) = \alpha' t_{D,2}$, or $x_2 = p_1^{-1}(\alpha'(t_{D,2} - t_{W,1}))$. Thus, $C_1(x_2) = C_2(x_2)$, and by Condition 1 commuters can choose to park in Area 2, i.e., $\lambda_2(y) > 0$ for $y \in [y_2, y_{max}]$. In this case, $C_1(x) = C_2(x)$, and we have $\alpha' t_{W,1} + p_1(x) = \alpha'\{t_{D,2} + \tau_{D,2}(x)\}$, where by Lemma B.1 $\tau_{D,2}(x)$ is equal to $\frac{(y - y_2)\int_0^{y - y_2} a\theta e^{\theta z} dz}{\int_0^{y - y_2} \lambda_2(z) - a\theta e^{\theta z} dz}$. By simplifying this equation, we get $\int_0^{y - y_2} \lambda_2(z)dz = (1 + \frac{y - y_2}{t_{W,1} + p_1(x)/\alpha' - t_{D,2}}) \int_0^{y - y_2} a\theta e^{\theta z} dz$. We take the derivative of this equation with respect to y to characterize $\lambda_2(y)$ as follows: $ae^{\theta(y - y_2)} + \frac{(y - y_2)ae^{\theta(y - y_2)} + \int_0^{y - y_2} a\theta e^{\theta z} dz}{t_{W,1} - t_{D,2} + p_1(x)/\alpha'} - \frac{p_1'(x)[(y - y_2)/\alpha'] \int_0^{y - y_2} a\theta e^{\theta z} dz}{[t_{W,1} - t_{D,2} + p_1(x)/\alpha']^2}$. In addition, the number of AVs that head to Area 1, $\lambda_1(y)$, is equal to the total number of AVs that leave W minus the number of AVs that head to Area 2, i.e., $R_W - \lambda_2(y)$.

Next, we derive T $(= y_{max})$. The last commuter does not experience any work schedule penalty, so the official start time at work is when the last commuter is dropped off at W, i.e., $T = y_{max} =$

$x_{max} + \tau_I(x_{max}) + \tau_W(x_{max}) = \frac{\int_0^{x_{max}} \lambda_H(u)du}{R_W}$. Since $\int_0^{x_{max}} \lambda_H(u)du$ is the total number of all commuters, N, we have $T = \frac{\int_0^{x_{max}} \lambda_H(u)du}{R_W} = \frac{N}{R_W}$.

Lastly, we show that our results hold even when $\tau_I(x)$ and $\tau_W(x)$ are not positive. First, if $\tau_I(x) = 0$ for some $x \in (0, x_{max}]$ and $\tau_W(x) > 0$, all of our results hold. This happens because $x + \tau_W(x)$ is still equal to $\frac{\int_0^x \lambda_H(u)du}{R_W}$, and $C_1(x)$ is equal to $C_1(x) = (\alpha - \beta)\frac{\int_0^x \lambda_H(u)du}{R_W} - \alpha x + \beta T + \alpha' t_{W,1} + p_1(x)$, and the rest of the proof remains unchanged. Next, let $\hat{x}$ be the smallest $x \in [0, x_{max}]$ such that $\tau_W(x) = 0$. Since $R_W \leq R_I$, $\tau_I(\hat{x})$ is also equal to 0. In this case, $C_1(\hat{x}) = \beta(T - \hat{x}) + \alpha' t_{W,1} + p_1(\hat{x})$. By setting the derivative of $C_1(\hat{x})$ equal to zero, we find that $p_1'(\hat{x})$ must be equal to β. We show that this is consistent with our result for $\lambda_H(x)$ by contradiction. First, we assume that $p_1'(\hat{x}) < \beta$. In this case, $\frac{\alpha - p_1'(\hat{x})}{\alpha - \beta} > 1$, and hence, $\lambda_H(x) > R_W$, which is a paradox; since $\tau_W(x) = 0$, $\lambda_H(x)$ cannot exceed the drop-off capacity R_W. Next, we assume that $p_1'(\hat{x}) > \beta$, which leads to $\lambda_H(x) < R_W$. In this case, compared to a commuter who leaves H at time $\hat{x}$, a commuter who leaves at time $\hat{x} + 1$ incurs $\$\beta$ more in work schedule penalty, but saves $\$p_1'(\hat{x})$ in downtown congestion cost. Since we assume $p_1'(\hat{x}) > \beta$, this commuter is better off departing H at time $\hat{x}$ instead of time $\hat{x} + 1$, so she can unilaterally change her decision without increasing any other commuter's cost. This means that there is no UE in this case, which is a paradox. Thus, $p_1'(\hat{x}) = \beta$ and our results still hold.

Case (ii): The proof of this case is similar to that of case (i), except that we need to characterize $x_{1,max}$ and x_2. The earliest time that Area 1 becomes full, $x_{1,max}$, satisfies the following equation: $\int_0^{x_{1,max}} \lambda_H(u)du = \int_0^{x_{1,max}} \frac{\alpha - p_1'(x)}{\alpha - \beta} R_W dx = K_1$, so $[\alpha x_{1,max} - p_1(x_{1,max}) + p_1(0)]\frac{R_W}{\alpha - \beta} = K_1$. If $p_1(x)$ is linear in x, then $x_{1,max} = \frac{K_1(\alpha - \beta)}{R_W(\alpha - p_1')}$. In addition, the drop-off time at W for commuters who leave H at time x_1 is equal to $y_1 = K_1/R_W$. To characterize the earliest departure time from H for commuters who choose Area 2, we find the value of x_2 that satisfies Conditions 1 and 2. Commuters who leave at time $x_{1,max}$ incur the cost $C_1(x_{1,max}) = (\alpha - \beta)y_{1,max} - \alpha x_{1,max} + \beta T + \alpha' t_{W,1} + p_1(x_{1,max})$. If commuters who choose Area 2 leave right at time $x_{1,max}$, their travel cost is equal to $C_2(x_{1,max}) = (\alpha - \beta)y_{1,max} - \alpha x_{1,max} + \beta T + \alpha' t_{D,2}$. If $C_2(x_{1,max})$ is higher than $C_1(x_{1,max})$, by Conditions 1 and 2, commuters who choose Area 1 must shift back their departure time from H by $\Delta = [C_2(x_{1,max}) - C_1(x_{1,max})]/\beta$ minutes to increase their work schedule penalty. So for $\Delta = [C_2(x_{1,max}) - C_1(x_{1,max})]/\beta = [\alpha'(t_{D,2} - t_{W,1}) - p_1(x_{1,max})]/\beta$ minutes no commuter leaves H.

Hence, we shift the timeline for Δ minutes to guarantee that the departure time of first commuters is time 0. In this case, $x_2 = x_{1,max} + \Delta$ and $T = N/R_W + \Delta$. Lastly, the largest K_1 that results in case (ii) corresponds to the case when the earliest time that Area 2 is chosen, x_2, is equal to the time $x_{1,max}$ that Area 1 is full, i.e., $\alpha' t_{D,2} = \alpha' t_{W,1} + p_1(x_{1,max})$. In this case, $x_{1,max} = p_1^{-1}(\alpha'(t_{D,2} - t_{W,1}))$, and $K_1 = [\alpha x_{1,max} - p_1(x_{1,max}) + p_1(0)]\frac{R_W}{\alpha-\beta} = [\alpha p_1^{-1}(\alpha'(t_{D,2} - t_{W,1})) - \alpha'(t_{D,2} - t_{W,1}) + p_1(0)]\frac{R_W}{\alpha-\beta}$. Thus, case (ii) holds for all $K_1 \leq [\alpha p_1^{-1}(\alpha'(t_{D,2} - t_{W,1})) - \alpha'(t_{D,2} - t_{W,1}) + p_1(0)]\frac{R_W}{\alpha-\beta}$.

Case(iii): This case can be shown similar to case (i), except that, at time $x_{1,max}$, Area 1 becomes full, and the remaining AVs have to go to Area 2. □

Proof of Proposition 2.2. (a) Case (i): We first show that, if $t_{2,D} \geq t_{W,1}$, then $\lambda_1(y) = R_W$ and $\lambda_2(y) = 0$. Next, we prove that $\lambda_H(x) = R_W$. Since $\lambda_1(y) + \lambda_2(y) = R_W$, we can restate the optimization problem stated in (2.5) as follows:

$$\min_{\lambda_H(x),\lambda_2(y)} \int_0^{x_{max}} \lambda_H(x)\{\alpha[\tau_I(x) + \tau_W(x)] + \beta(T - y)\}dx + \int_0^{y_{max}} R_W\alpha' t_{W,1} + \lambda_2(y)\alpha'[t_{2,D} - t_{W,1} + \tau_{2,D}(x)]dy$$

subject to:

$$y = x + \tau_I(x) + \tau_W(x)$$
$$0 \leq \lambda_H(x) \leq N$$
$$0 \leq \lambda_2(y) \leq R_W.$$

We use a two-stage approach to find the optimal solution of this control theory problem. In the first stage we find the value of $\lambda_2(y)$ that minimizes the second component of the objective function, i.e., $\int_0^{y_{max}} R_W\alpha' t_{W,1} + \lambda_2(y)\alpha'[t_{2,D} - t_{W,1} + \tau_{2,D}(x)]dy$. We are allowed to do this because the first component does not depend on $\lambda_2(y)$. In the second stage, we substitute $\lambda_2(y)$ with its optimal value from stage one, and find the value of $\lambda_H(x)$ that minimizes the entire objective function.

Stage one:

$$\min_{\lambda_2(y)} \int_0^{y_{max}} R_W\alpha' t_{W,1} + \lambda_2(y)\alpha'[t_{2,D} - t_{W,1} + \tau_{2,D}(x)]dy$$

subject to:

$$0 \leq \lambda_2(y) \leq R_W.$$

The objective function is a linear function of $\lambda_2(y)$ for $y \in [0, y_{max}]$, so the optimal value of $\lambda_2(y)$

lies on one of the boundaries of its feasible region, $[0, R_W]$. Since $t_{D,2} \geq t_{W,1}$ and $\tau_{D,2}(x) \geq 0$, $t_{D,2} - t_{W,1} + \tau_{D,2}(x) \geq 0$. Thus, the objective function is a linear increasing function of $\lambda_2(y)$ for $y \in [0, R_W]$, and $\lambda_2(y) = 0$ minimizes this function.

Stage two:

$$\min_{\lambda_H(x)} \int_0^{x_{max}} \lambda_H(x)\{\alpha[\tau_I(x) + \tau_W(x)] + \beta(T - y)\}dx + R_W \alpha' t_{W,1} y_{max} \tag{B.3}$$

subject to:

$$y = x + \tau_I(x) + \tau_W(x)$$

$$0 \leq \lambda_H(x) \leq N.$$

We can state y as a function of $\lambda_H(x)$, by substituting the values of $\tau_I(x)$ and $\tau_W(x)$ as follows: $y = x + \tau_I(x) + \tau_W(x) = x + \tau_I(x) + \frac{\int_0^{x+\tau_I(x)}[R_I - R_W]du}{R_W} = x + \tau_I(x) + \frac{R_I - R_W}{R_W}[x + \tau_I(x)] = \frac{R_I}{R_W}[x + \tau_I(x)] = \frac{R_I}{R_W}[x + \frac{\int_0^x \lambda_H(u) - R_I du}{R_I}] = \frac{\int_0^x \lambda_H(u)du}{R_W}$. By substituting this in (B.3), the objective function becomes minimizing $\int_0^{x_{max}} \lambda_H(x)[(\alpha - \beta)\frac{\int_0^x \lambda_H(u)du}{R_W} - \alpha x + \beta T]dx + R_W \alpha' t_{W,1} y_{max}$. After removing the constant term $R_W \alpha' t_{W,1} y_{max}$ from the objective function, we have the following:

$$\min_{0 \leq \lambda_H(x) \leq N} \int_0^{x_{max}} \lambda_H(x)[(\alpha - \beta)\frac{\int_0^x \lambda_H(u)du}{R_W} - \alpha x + \beta T]dx.$$

This is a control theory problem, where $\lambda_H(x)$ is the control variable (i.e., the variable that we can change to control the value of the objective function) and $\int_0^x \lambda_H(u)du$ is the state variable (i.e., the variable that changes as we change the control variable). We use the maximum principle approach to find the optimal value of the control variable $\lambda_H(x)$.[1] Let $f(x)$ represent the derivative of the state variable, i.e., $f(x) = \frac{\partial[\int_0^x \lambda_H(u)du]}{\partial x} = \lambda_H(x)$, and $g(x)$ denote the negative of the objective function, i.e., $g(x) = -\lambda_H(x)[(\alpha - \beta)\frac{\int_0^x \lambda_H(u)du}{R_W} - \alpha x + \beta T]$. The Hamiltonian equation for this problem is defined as follows: $H = \rho(x)f(x) + g(x) = \rho(x)\ \lambda_H(x) - \lambda_H(x)[(\alpha - \beta)\frac{\int_0^x \lambda_H(u)du}{R_W} - \alpha x + \beta T]$, where $\rho(x)$ is the marginal return vector. The variable $\rho(x)$ also satisfies the following two equations: $\frac{\partial \rho(x)}{\partial x} = -\frac{\partial H}{\partial[\int_0^x \lambda_H(u)du]} = (\alpha - \beta)\frac{\lambda_H(x)}{R_W}$, and $\rho(x_{max}) = \frac{\partial g}{\partial \lambda_H(x)}(x_{max}) = (\alpha - \beta)\frac{N}{R_W} - \alpha x_{max} + \beta T$. From these two equations, we get $\rho(x) = (\alpha - \beta)\frac{\int_0^x \lambda_H(u)du}{R_W} - \alpha x_{max} + \beta T$ for $x \in [0, x_{max}]$. We find the optimal value of $\lambda_H(x)$ that maximizes H. Since H is a linear function of $\lambda_H(x)$, if

[1] For further information on the maximum principle approach see Sethi & Thompson (2000).

$\rho(x) - (\alpha - \beta)\frac{\int_0^x \lambda_H(u)du}{R_W} + \alpha x - \beta T$ is positive, then $\lambda_H(x) = N$ maximizes H, otherwise $\lambda_H(x) = R_W$ is the optimal solution. Substituting $\rho(x)$ with $(\alpha - \beta)\frac{\int_0^x \lambda_H(u)du}{R_W} - \alpha x_{max} + \beta T$, we have $\rho(x) - (\alpha - \beta)\frac{\int_0^x \lambda_H(u)du}{R_W} + \alpha x - \beta T = -\alpha(x_{max} - x)$. Since $x \leq x_{max}$, $-\alpha(x_{max} - x)$ is negative, and the optimal value of $\lambda_H(x)$ for $x \in [0, x_{max}]$ is R_W.

Case (ii): The proof of this case is similar to the proof of case (i), except that at time $y_2 = x_2$ Area 1 becomes full, i.e., $K_1 = y_2 R_W$, and the remaining AVs must be routed to Area 2, i.e., $\lambda_2(y) = \lambda_H(x)$. In this case, $\lambda_H(x)$ can take two values R_W (which results in $\tau_{D,2}(x) \neq 0$) and $R_W - a$ (which results in $\tau_{D,2}(x) = 0$, since $\max\{R_{D,2}(x)\} = R_W - a$). Let time x_D ($\geq x_2$) be the time when the social planner sets $\lambda_H(x)$ for commuters who are routed to Area 2 equal to R_W, i.e., $\lambda_H(x) = R_W - a$ for $x \in [x_2, x_D)$ and $\lambda_H(x) = R_W$ for $x \in [x_D, x_{max}]$. The ideal start time at work, T, in this case is equal to $\frac{N - K_1 - (R_W - a)(x_D - x_2)}{R_W} + x_D = \frac{N - K_1 - (R_W - a)x_D + (R_W - a)x_2 + R_W x_D}{R_W} = \frac{N - K_1 + a x_D + (R_W - a)(K_1/R_W)}{R_W} = \frac{N + a(x_D - x_2)}{R_W}$. Our goal is to characterize x_D. In this case, the objective function becomes $\int_0^{x_2} \beta R_W(T - x)dx + \int_{x_2}^{x_D} \beta(R_W - a)(T - x)dx + \int_{x_D}^{T} \beta R_W(T - x)dx + \alpha' K_1 t_{W,1} + \alpha'(N - K_1)t_{D,2} + \alpha' R_W \int_0^{T - x_D} \tau_{D,2}(x)dx$. This objective function is convex in x_D, because its second derivative, which is equal to $\beta a(R_W - a)/R_W + \alpha' \frac{(R_W - a)^2}{R_W} \tau'_{D,2}(T - x_D)$, is positive. Therefore, the value of x_D that leads to the SO cost satisfies the first degree condition, i.e., $\beta a x_D(R_W - a)/R_W + \beta a^2 K_1/R_W^2 - \alpha'(R_W - a)\tau_{D,2}(T - x_D) = 0$. Note that if the value of x_D that satisfies the first degree condition is not in the interval $[x_2, T]$, then the optimum is the boundary point $x_2 = K_1/R_W$. The boundary point T cannot be the optimum, as the first derivative is positive at T.

(b) The proof of this case is similar to that of part (a). In particular, $\lambda_H(x)$ is equal to either $R_W - a$ or R_W. At time 0, since $t_{D,2} < t_{W,1}$ and congestion time in D, $\tau_{D,2}(x)$, is zero, all AVs are routed to Area 2. For $x < x_D$, $\lambda_H(x) = \max\{R_{D,2}(x)\} = R_W - a$, so $\tau_{D,2}(x)$ remains equal to zero and $\lambda_2(x) = R_W - a$. At time x_D, $\lambda_H(x)$ increases to R_W, and as long as $\tau_{D,2}(x)$ is low enough that $t_{D,2} + \tau_{D,2}(x) \leq t_{W,1}$, all AVs are routed to Area 2, i.e., $\lambda_2(y) = R_W$ and $\lambda_1(y) = 0$. Let time $y_D + y_1$ be the earliest drop-off time at W such that $t_{D,2} + \tau_{D,2}(x) = t_{W,1}$. All AVs that leave W at time $y_D + y_1$ are routed to Area 1, because $\tau_{D,2}(x)$ is an increasing function of y, but $t_{W,1}$ is a constant value. After $y_D + y_1$, all AVs are routed to Area 1, i.e., $\lambda_1(y) = R_W$ and $\lambda_2(y) = 0$, until $\tau_{D,2}(x)$ becomes zero. Let $y_D + y_2$ be the earliest drop-off time at W such that $\tau_{D,2}(x)$ becomes equal to zero again. At time $y_D + y_2$, AVs are again routed to Area 2, and this cycle continues until all N

AVs are routed to a parking area. Next, we characterize $x_D \in [0, T]$, which is the time until when the social planner sets $\lambda_H(x) = R_W - a$ and routes AVs to Area 2, i.e., $\lambda_2(y) = R_W - a$ for $y < y_D$. After time $x_D = y_D$, the social planner alternately routes AVs to Area 2 and Area 1. Our goal is to find the optimal value of x_D that minimizes the total system cost, i.e., $\int_0^{x_D}(R_W - a)\beta(T - x)dx + \int_{x_D}^{T} R_W\beta(T - x)dx + \int_0^{T-x_D} \alpha' R_W \min\{\tau_{D,2}(x), t_{W,1} - t_{D,2}\}dx + N\alpha' t_{D,2}$. The ideal start time at work, T, in this case is equal to $\frac{N+ax_D}{R_W}$. This objective function is convex in x_D, because its second derivative, which is equal to $\beta a(R_W - a)/R_W + \alpha'(R_W - a)^2\tau'_{D,2}(\min\{y_1, T - x_D - y_2\lfloor\frac{T-x_D}{y_2}\rfloor\})$, is positive. Therefore, the value of x_D that leads to the SO cost satisfies the first degree condition, i.e., $\beta a x_D(R_W - a)/R_W - \alpha'(R_W - a)\tau_{D,2}(\min\{y_1, T - x_D - y_2\lfloor\frac{T-x_D}{y_2}\rfloor\}) = 0$. Note that if the value of x_D that satisfies the first degree condition is not in the interval $[0, T]$, then the optimum is the boundary point $x = 0$. The boundary point T cannot be the optimum, as the first derivative is positive at T.

Case (ii): The proof of this part is analogous to the proof of case (i), except that at time $y_{1,max}$ Area 1 becomes full, i.e., $K_1 = \int_0^{y_{1,max}} \lambda_1(y)dy$, and all AVs that leave W after time $y_{1,max}$ must be routed to Area 2. □

Lemma B.2. *Under SO, if $\lambda_2(x) = R_W$ for $x \in [0, \hat{x}]$, then $\tau_{D,2}(x)$ is a convex increasing function of $x(\in [0, \hat{x}])$.*

Proof. If $\lambda_2(x) = R_W$ for $x \in [0, \hat{x}]$, by Lemma B.1 the congestion time in D is equal to $\tau_{D,2}(x) = \frac{xa(e^{\theta x}-1)}{R_W x - a(e^{\theta x}-1)}$. Taking the derivative of $\tau_{D,2}(x)$ with respect to x, we have $\tau'_{D,2}(x) = \frac{R_W x^2 a\theta e^{\theta x} - (ae^{\theta x} - a)^2}{[R_W x - a(e^{\theta x}-1)]^2}$. Since $R_D(x) = R_W - a\theta e^{\theta x} > 0$, then $\int_0^x R_D(u)du = \int_0^x R_W - a\theta e^{\theta u}du = R_W x - (ae^{\theta x} - a) > 0$. In addition, it can be shown that $xa\theta e^{\theta x} \geq ae^{\theta x} - a$, so $R_W x^2 a\theta e^{\theta x} - (ae^{\theta x} - a)^2 \geq R_W x(ae^{\theta x} - a) - (ae^{\theta x} - a)^2 = (ae^{\theta x} - a)[R_W x - (ae^{\theta x} - a)]$. Thus, both the numerator and the denominator of $\tau'_{D,2}(x)$ are non-negative, and $\tau_{D,2}(x)$ is increasing in x.

Next, to prove the convexity of $\tau_{D,2}(x)$, we show that its second derivative is positive. The second derivative of $\tau_{D,2}(x)$ is equal to $\tau''_{D,2}(x) = \frac{aR_W[2R_W(x^2e^{\theta x}/2 + e^{\theta x} - xe^{\theta x} - 1) + ae^{\theta x}(x + xe^{\theta x} + 2 - 2e^{\theta x})]}{[R_W x - a(e^{\theta x}-1)]^3} + 2\frac{R_W a\theta[e^{\theta x}(x-1)+1]}{[R_W x - a(e^{\theta x}-1)]^2}$. The second term and the denominator of the first term of $\tau''_{D,2}(x)$ are positive. The numerator of this term is also positive, because $[2R_W(x^2e^{\theta x}/2 + e^{\theta x} - xe^{\theta x} - 1) + ae^{\theta x}(x + xe^{\theta x} + 2 - 2e^{\theta x})]$ is greater than $[2ae^{\theta x}(x^2e^{\theta x}/2 + e^{\theta x} - xe^{\theta x} - 1) + ae^{\theta x}(x + xe^{\theta x} + 2 - 2e^{\theta x})] > 0$. Therefore, $\tau_{D,2}(x)$ is convex. □

Proof of Proposition 2.3. (a) The SO1 solution becomes a UE solution, if Conditions 1 and 2 are satisfied. In other words, for SO1 to be a UE, all commuters must incur the same travel cost, regardless of their departure time from H or their parking location. Our approach is to equalize all commuters' costs, by charging commuters a parking fee, $p_1(x)$, or a congestion toll, $\pi_2(x)$. Here we derive the functions $p_1(x)$ and $\pi_2(x)$ for case (ii). The functions $p_1(x)$ and $\pi_2(x)$ can be derived similarly for case (i).

Under SO1, the total travel cost for commuters who go to Area 1 is $C_1(x) = \beta(T-x)+\alpha' t_{W,1} + p_1(x)$ for $0 \le x \le x_2$ and that for commuters who go to Area 2 is $C_2(x) = \beta(T-x)+\alpha' t_{D,2} + \pi_2(x)$ for $x_2 \le x \le x_D$ and $C_2(x) = \beta(T-x) + \alpha'[t_{D,2} + \tau_{D,2}(x)] + \pi_2(x)$ for $x_D \le x \le x_{max}$. Note that since there is no passengered congestion time, the congestion in I, $\tau_I(x)$, and drop-off congestion, $\tau_W(x)$, are not included in the cost. The commuters who incur the maximum cost are the ones who depart home at one of the boundary points $x = 0, x_2$ or $x_{\max}$. This is because $C_1(x)$ is decreasing in x, and $C_2(x)$ is convex (because $\beta(T-x)$ is linearly decreasing in x, and, by Lemma B.2, $\tau_{D,2}(x)$ is convex and increasing in x.) This means that the maximum commuter cost is equal to the cost at one of these three boundary points. In other words, the maximum commuter cost is equal to $\max\{\beta T + \alpha' t_{W,1}, \beta(T-x_2) + \alpha' t_{D,2}, \alpha' t_{D,2} + \alpha' \tau_{D,2}(x_{max})\}$. As such, the parking fee $p_1(x)$ (resp., the congestion toll $\pi_2(x)$) is the difference between $\max\{\beta T + \alpha' t_{W,1}, \beta(T-x_2) + \alpha' t_{D,2}, \alpha' t_{D,2} + \alpha' \tau_{D,2}(x_{max})\}$ and $C_1(x)$ (resp., $C_2(x)$). So, we have the following:

$$
\begin{aligned}
p_1(x) &= \max\{\beta T + \alpha' t_{W,1}, \beta(T-x_2) + \alpha' t_{D,2}, \alpha' t_{D,2} + \alpha' \tau_{D,2}(x_{max})\} - \beta(T-x) - \alpha' t_{W,1} \\
&= \beta x + \max\{0, -\beta x_2 + \alpha'(t_{D,2} - t_{W,1}), -\beta T + \alpha'(t_{D,2} - t_{W,1}) + \alpha' \tau_{D,2}(x_{max})\} \text{ for } 0 \le x \le x_2, \text{ and} \\
\pi_2(x) &= \max\{\beta T + \alpha' t_{W,1}, \beta(T-x_2) + \alpha' t_{D,2}, \alpha' t_{D,2} + \alpha' \tau_{D,2}(x_{max})\} - \beta(T-x) - \alpha' t_{D,2} \\
&= \beta x + \max\{-\alpha'(t_{D,2} - t_{W,1}), -\beta x_2, -\beta T + \alpha' \tau_{D,2}(x_{max})\} \text{ for } x_2 \le x \le x_D, \text{ and} \\
\pi_2(x) &= \max\{\beta T + \alpha' t_{W,1}, \beta(T-x_2) + \alpha' t_{D,2}, \alpha' t_{D,2} + \alpha' \tau_{D,2}(x_{max})\} - \beta(T-x) - \alpha'[t_{D,2} + \tau_{D,2}(x)] \\
&= \beta x - \alpha' \tau_{D,2}(x) + \max\{-\alpha'(t_{D,2} - t_{W,1}), -\beta x_2, -\beta T + \alpha' \tau_{D,2}(x_{max})\} \text{ for } x \ge x_D
\end{aligned}
$$

Lastly, we add a positive constant $\epsilon_1(x)$ to $p_1(x)$ when $x \in [x_2, x_{max}]$ to make sure that commuters who leave H after time x_2 do not choose Area 1. Similarly, we add a positive constant $\epsilon_2(x)$ to $\pi_2(x)$ when $x \in [0, x_2]$ to make sure that commuters who leave H before time x_2 do not choose

Area 2.

(b) Similar to part (a), we derive $p_1(x)$ and $\pi_2(x)$ for case (i) of SO2. In this case, the total travel cost for commuters who go to Area 1 is $C_1(x) = \beta(T-x) + \alpha' t_{W,1} + p_1(x)$ and that for commuters who go to Area 2 is $C_2(x) = \beta(T-x) + \alpha' t_{D,2} + \pi_2(x)$ for $x \leq x_D$ and $C_2(x) = \beta(T-x) + \alpha'[t_{D,2} + \tau_{D,2}(x)] + \pi_2(x)$ for $x_D \leq x$. The commuters who incur the maximum cost are the ones who depart home at one of the boundary points $x = 0$ or $x_D + x_1$. This is because $C_1(x)$ is decreasing in x, and $C_2(x)$ is convex (because $\beta(T-x)$ is linearly decreasing in x, and, by Lemma B.2, $\tau_{D,2}(x)$ is convex and increasing in x.) In other words, the maximum commuter cost is equal to $\max\{\beta T + \alpha' t_{D,2}, \beta(T - x_D - x_1) + \alpha' t_{W,1}\}$. As such, the parking fee $p_1(x)$ (resp., the congestion toll $\pi_2(x)$) is the difference between $\max\{\beta T + \alpha' t_{D,2}, \beta(T - x_D - x_1) + \alpha' t_{W,1}\}$ and $C_1(x)$ (resp., $C_2(x)$). So, we have the following:

$$
\begin{aligned}
p_1(x) &= \max\{\beta T + \alpha' t_{D,2}, \beta(T - x_D - x_1) + \alpha' t_{W,1}\} - \beta(T-x) - \alpha' t_{W,1} \\
&= \beta x + \max\{-\beta(x_D + x_1), \alpha'(t_{D,2} - t_{W,1})\}, \text{ and} \\
\pi_2(x) &= \max\{\beta T + \alpha' t_{D,2}, \beta(T - x_D - x_1) + \alpha' t_{W,1}\} - \beta(T-x) - \alpha' t_{D,2} \\
&= \beta x + \max\{-\beta(x_D + x_1) - \alpha'(t_{D,2} - t_{W,1}), 0\} \text{ for } x_2 \leq x \leq x_D, \text{ and} \\
\pi_2(x) &= \max\{\beta T + \alpha' t_{D,2}, \beta(T - x_D - x_1) + \alpha' t_{W,1}\} - \beta(T-x) - \alpha'[t_{D,2} + \tau_{D,2}(x)] \\
&= \beta x - \alpha' \tau_{D,2}(x) + \max\{-\beta(x_D + x_1) - \alpha'(t_{D,2} - t_{W,1}), 0\} \text{ for } x \geq x_D
\end{aligned}
$$

Lastly, we add a positive constant $\epsilon_1(x)$ to $p_1(x)$ when $nx_2 \leq x \leq \min\{x_D + nx_2 + x_1, x_{max}\}$ for $n = 0, 1, \cdots, [N - (R_W - a)x_D]/(R_W y_2)$ to make sure that commuters do not choose Area 1. Similarly, we add a positive constant $\epsilon_2(x)$ to $\pi_2(x)$ when $x \notin [nx_2, \min\{x_D + nx_2 + x_1, x_{max}\}]$ for $n = 0, 1, \cdots, [N - (R_W - a)x_D]/(R_W y_2)$ to make sure that commuters who leave during these time intervals do not choose Area 2. The functions $p_1(x)$ and $\pi_2(x)$ for case (ii) can be derived similarly. **Remarks** In §2.6.1, for simplicity, we assume that, for the intervals of x when $\epsilon_1(x)$ and $\epsilon_2(x)$ are non-zero, they are equal to $|\alpha'(t_{D,2} - t_{W,1})|$. However, as discussed in the proof of Proposition 2.3, any positive values of $\epsilon_1(x)$ and $\epsilon_2(x)$ for those intervals yield the same parking location decisions. □

Proof of Corollary 2.1. Under SO1, let C_{SO1} be the optimal value of the objective function

in (2.5), where $\lambda_H(x)$, $\lambda_1(y)$ and $\lambda_2(y)$ are substituted with those in Table 2.4, i.e., $C_{SO1} = \frac{\beta T^2}{2} - a\beta(x_D - x_2)(T - \frac{x_D+x_2}{2}) + \alpha' K_1(t_{W,1} - t_{D,2}) + \alpha' t_{D,2} N + \alpha' R_W \int_{x_D}^{T} \tau_{D,2}(x)dx$. The total system cost C_{SO1} is continuous and differentiable with respect to $R_W \in (0, K_1 + R_W]$. Since the number of downtown parking spots, K_1, and the drop-off capacity, R_W, are both limited, the total available space downtown is limited, i.e., $K_1 + R_W < \infty$. In addition, since the drop-off capacity does not exceed the inbound bottleneck capacity R_I, there exist R_W^* that minimizes C_{SO1} on the closed interval $[0, \min\{K_1 + R_W, R_I\}]$. In this case, the optimal downtown parking space capacity is equal to $K_1^* = K_1 + R_W - R_W^*$. Similarly, we can show that under SO2, there exist R_W^* and K_1^* that minimize the total system cost. □

B.4 Additional Analysis

B.4.1 Robustness of Case (i) of UE1

Our calibrated model parameters satisfy the condition for UE1: $\alpha' t_{D,2} = 2.25 \times 15/60 \leq 2.25 \times 2/60 + 1.80 = \alpha' t_{W,1} + p_1(0)$. In addition, since Area 1 is not chosen by any of the commuters (i.e., $y_1 > y_{max} = 333$), the condition for case (i) of UE1 is satisfied: $K_1 = 10,000 > N - R_W y_{max} - a(e^{\theta(N/R_W - N/R_W)} - 1)[1 + \frac{N/R_W}{t_{W,1} - t_{D,2} + p_1(x_{max})/\alpha'}] = 20,000 - 60 \times 333 - 0 = 0 > N - R_W y_1 - a(e^{\theta(N/R_W - y_1)} - 1)[1 + \frac{N/R_W}{t_{W,1} - t_{D,2} + p_1(x_{max})/\alpha'}]$.

We demonstrate that the observation of case (i) of UE1 is robust over a wide range of the passengerless travel cost α', number of commuters N, and drop-off rate R_W. In addition, we show that even when Area 2 is located in H_3 (see Figure B.1), case (i) of UE1 is observed.

For our calibrated model, the observation of case (i) of UE1 is robust over a wide range of model parameters. In particular, we show that the two conditions ($\alpha' t_{D,2} \leq \alpha' t_{W,1} + p_1(0)$ and $K_1 \geq N - R_W y_1 - a(e^{\theta(N/R_W - y_1)} - 1)[1 + \frac{N/R_W}{t_{W,1} - t_{D,2} + p_1(x_{max})/\alpha'}]$) required for case (i) of UE1 are satisfied for reasonably estimated parameters values.[2] First, the condition $\alpha' t_{D,2} \leq \alpha' t_{W,1} + p_1(0)$ depends on the distance of Area 2 from D ($t_{D,2} = 15$ minutes), the travel time from W to Area 1 ($t_{W,1} = 2$ minutes), the parking fee in Area 1 for the first group of commuters ($p_1(0) = \$1.80$), and the passengerless travel cost ($\alpha' = \$2.25$ per hour). Given that the travel time from W to Area 1,

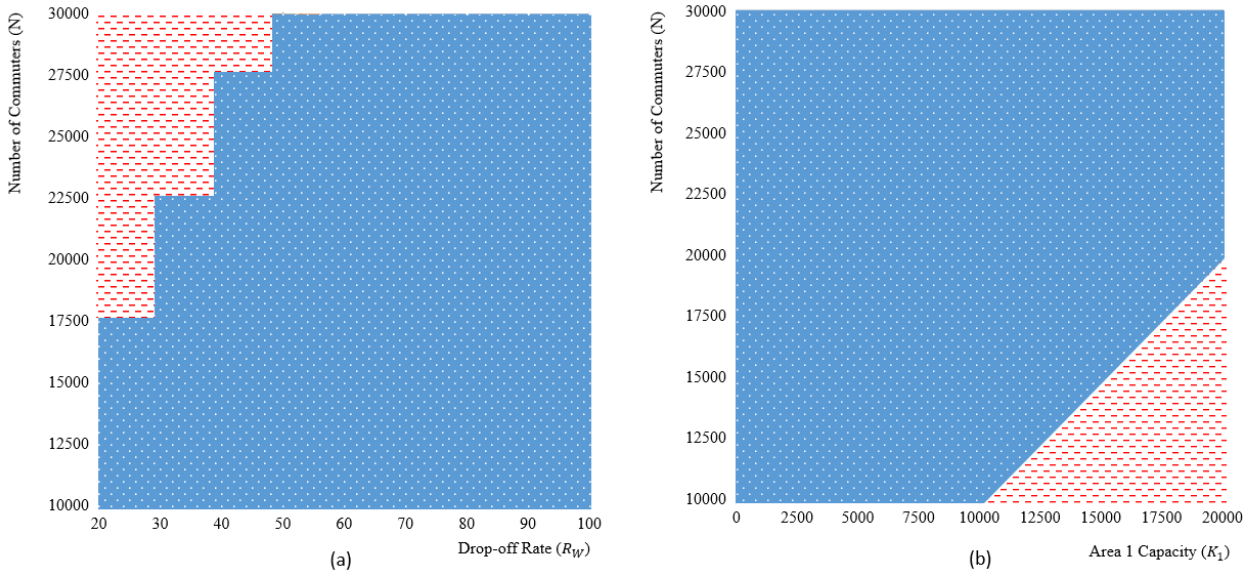
[2]Since neither of the conditions for case (i) of UE1 depends on the value of α, our results hold for any positive value of $\alpha \geq \alpha'$.

$t_{W,1}$, also includes the time it takes to park, our estimate of 2 minutes for $t_{W,1}$ is a conservative one. Hence, any higher estimate of $t_{W,1}$ still satisfies the condition for UE1. To show the robustness of observing UE1 for the three remaining parameters, we consider two alternative locations for Area 2: H_1 and H_3 (see Figure B.1). First, suppose Area 2 is located in H_1, which is the farthest feasible option from D. Then the free-flow travel time $t_{D,2}$ from D to the Area 2 is 30 minutes. As such, for all values of α' less than $p_1(0)/(t_{D,2} - t_{W,1}) = \frac{p_1(0)}{(30-2)/60} = 2.14p_1(0)$, we observe UE1. In fact, for our current estimate of \$1.80 for Area 1's parking fee for the first group of commuters, $p_1(0)$, UE1 is observed when the passengerless travel cost α' is less than \$3.86 per hour. However, $p_1(0) = \$1.80$ is itself a conservative estimate for the daily parking fee in Downtown Pittsburgh because, as Parkopedia (2020) shows, the daily parking rate in Downtown Pittsburgh is usually at least \$4. This means that the necessary condition for observing UE1 is satisfied for all $\alpha' \leq \$8.57$ per hour, which is an even higher margin than the value of passengered travel cost $\alpha = \$6.4$ per hour estimated in Pi et al. (2019).

Next, we analyze the case where Area 2 is located in H_3 (see Figure B.1). As mentioned in Appendix B.2, in this case the travel time from D to Area 2 , $t_{D,2}$, is 5 minutes, and the parking fee in Area 2 is half of that in Area 1, i.e., $p_2(x) = p_1(x)/2$. In this case, we still observe UE1, because the cost associated with parking in Area 2 at time zero is lower than that in Area 1, i.e., $\alpha' t_{D,2} + p_2(0) = 2.25 \times 5/60 + 0.90 \leq 2.25 \times 2/60 + 1.80 = \alpha' t_{W,1} + p_1(0)$. In fact, for all values of α' that are less than or equal to $\frac{p_1(0)-p_2(0)}{t_{D,2}-t_{W,1}} = \frac{1.8-0.90}{(5-2)/60} = \$18$ per hour, we observe UE1. In addition, even if α' is as high as the passengered travel cost $\alpha = \$6.4$ per hour estimated by Pi et al. (2019), as long as the Area 2's parking fee is at least \$0.96 cheaper than Area 1's parking fee at time zero, i.e., $p_1(0) - p_2(0) \geq 0.96$, we observe UE1. It is worth noting that, similar to the case when Area 2 is located in H_2, when Area 2 is located in H_3 all commuters choose Area 2. In addition, the total travel cost that a commuter incurs increases from \$22.23 in the former case to \$22.75 in this case. Given that H_2 results in a marginally lower cost than H_3, the latter location could be a valid option for Area 2. We further discuss this option in Appendix B.4.4.

The second condition for observing case (i) of UE1 is $K_1 \geq N - R_W y_1 - a(e^{\theta(N/R_W - y_1)} - 1)[1 + \frac{N/R_W}{t_{W,1} - t_{D,2} + p_1(x_{max})/\alpha'}]$. In our calibrated model, all commuters choose Area 2 and Area 1 is never used, i.e., $y_1 > y_{max}$. This case of UE1 is observed when the cost associated with parking in Area 2 is always lower than that in Area 1, i.e., $\alpha'[t_{D,2} + \tau_{D,2}(x)] \leq \alpha' t_{W,1} + p_1(x)$ for $x \in [0, x_{max}]$. The cost

Figure B.2: Robustness region of: (a) case (i) of UE1, and (b) case (ii) of SO1.



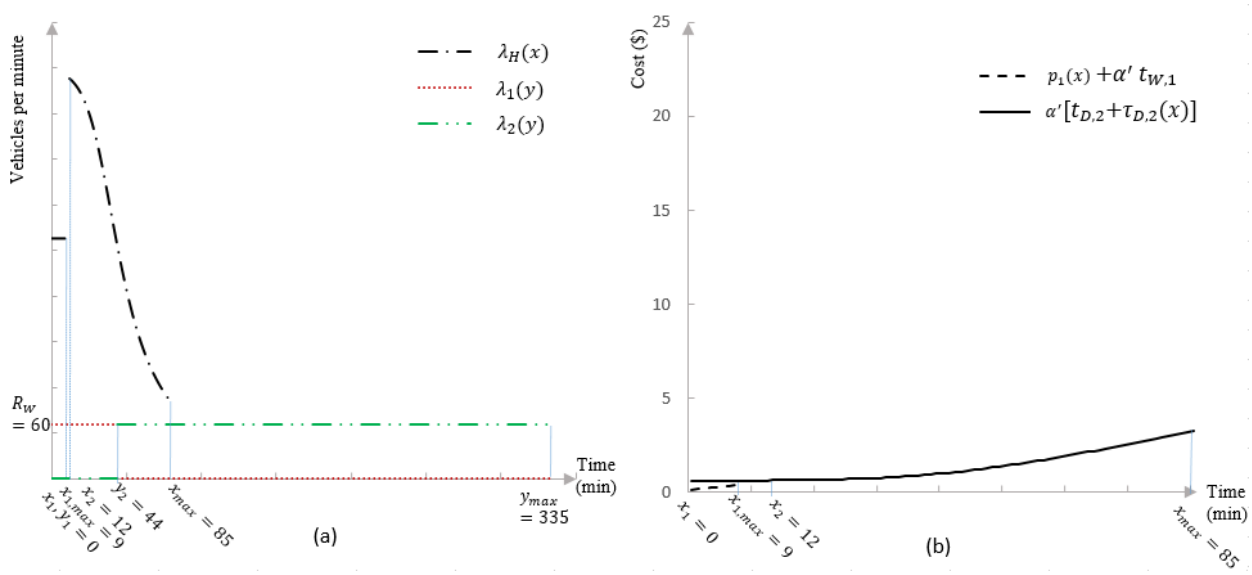
Notes. In (a), the (blue) shaded region, which includes the point (60, 20000), illustrates the parameter values for which we observe case (i) of UE1. In (b), the (blue) shaded region, which includes the point (10000, 20000), illustrates the parameter values for which we observe case (ii) of SO1.

associated with parking in Area 2 is a convex increasing function of x (see Lemma B.2 in Appendix B.3), and that in Area 1 is linearly increasing in x, so to observe this case, it suffices to show that Area 2 is cheaper than Area 1 at x_{max}, i.e., $\alpha'[t_{D,2} + \tau_{D,2}(x_{max})] \leq \alpha' t_{W,1} + p_1(x_{max})$. Considering the most conservative estimates for our parameter values, i.e., $t_{D,2} = 30$ minutes, $t_{W,1} = 2$ minutes, and $p_1(x_{max}) = \$16$, Area 2 is cheaper than Area 1 at time x_{max} when $\tau_{D,2}(x_{max}) \leq 122$ minutes. Figure B.2(a) illustrates the range of number of commuters (N) and that of the drop-off congestion rate (R_W) that satisfy the condition $\tau_{D,2}(x_{max}) \leq 122$ minutes. As shown in this figure, only when R_W is significantly lower and N is significantly higher than our estimated values, case (i) is not observed. Therefore, our observation of case (i) of UE1 in Pittsburgh is robust.

B.4.2 A Numerical Analysis of UE2

As mentioned in §2.4.1, UE2 is observed in cities with low parking fees, so for the numerical example depicted in Figure 2.4, we let $p_1(x) = 8x/x_{max}$, which is lower than $p_1(x) = 1.8 + 14.2x/x_{max}$ in our calibrated model for $x \geq 0$. In addition, we choose H_3 (see Figure B.1) with $t_{D,2} = 5$ minutes as the location of Area 2, since the distance from downtown to suburban neighborhoods is shorter in small

Figure B.3: An illustration of case (ii) of UE2: (a) departure rates, and (b) costs associated with parking in Areas 1 and 2.



Note$_1$. In (a), $\lambda_H(x)$ denotes the departure rate from H at time x, and $\lambda_1(y)$ and $\lambda_2(y)$ denote the departure rates from W at time y to Areas 1 and 2, respectively. In (b), $p_1(x) + \alpha' t_{W,1}$ represents the cost associated with parking in Area 1, and $\alpha'[t_{D,2} + \tau_{D,2}(x)]$ is that in Area 2.

cities. In addition, we let the capacity K_1 of Area 1 equal to $20,000$. Under these assumption, the model parameters in this example satisfy the condition for UE2: $\alpha' t_{D,2} = 2.25 \times 5/60 > 2.25 \times 2/60 + 0 = \alpha' t_{W,1} + p_1(0)$, and the condition for case (i) of UE2: $K_1 = 20,000 \geq 17,594.8 = N - [\frac{N/R_W}{t_{W,1} - t_{D,2} + p_1(x_{max})/\alpha'} + 1]a(e^{\theta(N/R_W - y_2)} - 1)$.

Under UE2, Area 1 is the primary parking area chosen by commuters, and Area 2 is an auxiliary parking area. Depending on the magnitude of Area 1's capacity K_1, three different variations of UE2 can occur: case (i) when K_1 is high, case (ii) when K_1 is low, and case (iii) when K_1 is moderate. An in-depth discussion and numerical analysis of case (i) is provided in §2.4.2. The case of moderate K_1, i.e., case (iii), is similar to the first case, except that after time $y_{1,max}$ Area 1 becomes full and all the remaining commuters go to Area 2. To avoid repetition, we do not present a numerical example of this case. Figure B.3(a) portrays a numerical example of case (ii), where we assume that the capacity of Area 1 , K_1, is equal to $2,638$ parking spots, so the condition on K_1 for case (ii) is satisfied, i.e., $K_1 = 2,638 \leq 4,143.8 = \frac{R_W(\alpha - p_1'(x))p_1^{-1}(\alpha'(t_{D,2} - t_{W,1}))}{\alpha - \beta}$. The remaining parameters are the same as those in the numerical example discussed in §2.4.2. As this figure shows, all commuters who leave H early (i.e., before time $x_{1,max} = 9$ minutes) park in Area 1, i.e.,

$\lambda_1(y) = R_W = 60$ AVs per minute and $\lambda_2(y) = 0$ for $0 \leq y \leq y_{1,max} = 44$ minutes. As Figure B.3(b) shows, although Area 1 is the cheaper parking option, commuters who choose this Area incur a higher work schedule penalty as they arrive at work earlier than commuters who choose Area 2. This leads to a period of time from $x_{1,max} = 9$ minutes to $x_2 = 12$ minutes, during which no commuter leaves H, as Area 1 is full and the cost of traveling to Area 2 is so high that commuters rather leave later to decrease their total travel cost by reducing their work schedule penalty. After $x_2 = 12$ minutes, all commuters park in Area 2, i.e., $\lambda_2(y) = R_W$ for $y_2 = 44 \leq y \leq y_{max} = 335$ minutes.

B.4.3 Robustness of Case (ii) of SO1

For our calibrated model, we show that the observation of case (ii) of SO1 is robust over a wide range of model parameters, including $t_{D,2}$, $t_{W,1}$, N and K_1. In particular, we show that the two conditions ($t_{D,2} \geq t_{W,1}$ and $K_1 < N$) required for case (ii) of SO1 are satisfied for reasonably estimated parameters values. The first condition, $t_{D,2} \geq t_{W,1}$, depends on the travel time from D to Area 2, $t_{D,2}$, and the travel time from W to Area 1, $t_{W,1}$. For Pittsburgh, even if we assume Area 2 is located in H_3, which is the closest feasible option to D (see Figure B.1), the free-flow travel time $t_{D,2}$ from D to the Area 2 (which is 5 minutes) is still higher than the travel time $t_{W,1}$ from W to Area 1 (which is 2 minutes). This means as long as $t_{W,1}$ is not higher than 2.5 times of our estimated value, we observe SO1. In addition, to observe case (ii) we need the condition $K_1 < N$. Figure B.2(b) illustrates the range of the number of commuters (N) and that of Area 1 capacity (K_1) that satisfy this condition. We observe from this figure that for a wide range of parameter values for N and K_1 the condition for case (ii) of SO1 is satisfied.

B.4.4 The Location of Area 2

There are two potential locations for Area 2 outside Downtown Pittsburgh: H_2 and H_3 (see Figure B.1). In the analysis of our calibrated model, we assume Area 2 is located in H_2 rather than H_3. This is because there is already an abundance of parking spots in H_2, while the City needs to invest on building a new parking area in H_3, should Area 2 be located there. Yet, one can argue that building new parking spots in H_3 is a beneficial investment in the long run, as it reduces the total distance traveled by AVs. To this end, we calculate the annual savings in the total system cost under

SO due to relocating Area 2 from H_2 to H_3. On the one hand, if Area 2 is moved to H_3, the travel time of each commuter decreases by 10 minutes (from $t_{D,2} = 15$ minutes for H_2 to $t_{D,2} = 5$ minutes for H_3). This leads to $(N - K_1) \times \alpha' \times 10/60 = (20,000 - 10,000) \times 2.25 \times 10/60 = \$3,750$ in daily savings (where $N - K_1$ represent the number of commuters that go to Area 2 under SO), or $\$945,000$ in annual savings (considering 252 work days per year). On the other hand, according to Strong Towns (2018), the cost of creating one parking spot can vary from $\$5,000$ to $\$50,000$. Assuming building one parking spot outside Downtown Pittsburgh costs $\$5,000$, the social planners need to invest 50 million dollars to build 10,000 ($= N - K_1$) parking spots in H_3. Even when doubling the commute cost savings due to two-way trips to the parking area, this investment requires more than 25 years to break even. Thus, H_3 does not seem to be a feasible option for Area 2.

B.4.5 The HV Case

To compare the effect of AVs on the morning commute case, we analyze a morning commute model for HVs. In this model, after going through the inbound bottleneck, commuters who use HVs drive to Area 1 to park their vehicles, and then walk to W. These commuters do not experience any drop-off congestion at W (as HVs do not have the ability to drop off their commuters), neither do they have the option to park outside downtown (as the walking time to W is prohibitively high). However, they might experience some congestion as they drive around downtown to find empty parking spots; we call this the parking cruising congestion time, $\tau_1(x)$. Following Qian & Rajagopal (2014) and Qian & Rajagopal (2015), we let $\tau_1(x)$ be equal to $\frac{t_{W,1}}{K_1(x)/K_1}$, where $K_1(x)$ is the number of empty spots in Area 1 available to a commuter who leaves H at x. Then, the travel cost for a commuter consists of the following elements: the inbound congestion cost (i.e., $\alpha\tau_I(x)$), parking cruising congestion cost (i.e., $\alpha\tau_1(x)$), cost of walking from Area 1 to W (i.e., $\alpha'' t_{1,W}$, where α'' and $t_{1,W}$ represent the monetary value of walk time and the walk time from Area 1 to W, respectively), work schedule penalty (i.e., $\beta[T - x - \tau_I(x) - t_{W,1} - t_{1,W}]$, and parking fee of Area 1 (i.e., $p_1(x)$). Taken together, the travel cost of a commuter under the HV case is equal to $C_1(x) = \alpha\tau_I(x) + \alpha\tau_1(x) + \alpha'' t_{1,W} + \beta[T - x - \tau_I(x) - \tau_1(x) - t_{1,W}] + p_1(x)$. To estimate this travel cost, we use the model parameters discussed in Table 2.1, except that we use $\alpha = \$9$ for HVs, which is equal to twice the value of α for AVs in the calibrated model. This is because the HV commute time is more valuable than the AV commute time, since an AV commuter can

dedicate her commute time to an activity other than driving. In addition, in the prior literature (e.g., Zhang et al. (2008), Qian et al. (2011), Qian et al. (2012), Qian & Rajagopal (2014), Qian & Rajagopal (2015), and Zhang et al. (2019)), the estimates for α range from \$9.91 to \$36 per hour, so our estimate of $\alpha = \$9$ per hour for HVs is a conservative one. In addition, we estimate two model parameters specific to the HV case that are not available in Pi et al. (2019): α'' and $t_{1,W}$. First, we assume α'' is equal to $\alpha = \$9$ per hour in our model. This is a lower bound on α', as the cost of walking is higher than passengered travel cost. Second, based on the size of Downtown Pittsburgh, we estimate that $t_{1,W}$ is approximately 15 minutes for a commuter.

Using these model parameters, we calculate the departure rate from H. To this end, similar to the proof of Proposition 2.1, we set the derivative of the commuters' travel cost $C_1(x)$ equal to zero, i.e., $(\alpha-\beta)\frac{\lambda_H(x)-R_I}{R_I}+(\alpha-\beta)\frac{K_1 t_{W,1}\lambda_H(x)}{[K_1-\int_0^x \lambda_H(u)du]^2}-\beta+p_1'(x)=0$. This is an ODE in the form of $\frac{\alpha-\beta}{R_I}f(x)f'(x)+K_1 t_{W,1}f'(x)+1\frac{\alpha-p_1'(x)}{\alpha-\beta}f(x)\sqrt{f(x)}=0$, where $f(x)=[K_1-\int_0^x \lambda_H(u)du]^2$, $f'(x)=-2\lambda(x)[K_1-\int_0^x \lambda_H(u)du]$, and $f(0)=K_1^2$. The solution to this ODE is equal to $f(x)=R_I^2[\frac{8K_1 t_{W,1}}{R_I}-4c\frac{\alpha-p_1'(x)}{\alpha-\beta}x+c^2+4(\frac{\alpha-p_1'(x)}{\alpha-\beta})^2x^2-(2\frac{\alpha-p_1'(x)}{\alpha-\beta}x-c)\sqrt{\frac{8K_1 t_{W,1}}{R_I}-4c\frac{\alpha-p_1'(x)}{\alpha-\beta}x+c^2+4(\frac{\alpha-p_1'(x)}{\alpha-\beta})^2x^2}]/8$. Using our model parameters and $c = 569.9$ (which is the constant factor that guarantees $f(0) = K_1^2$), $\lambda_H(x)$ ranges from 84.42 vehicles per minute at time zero to 49.77 vehicles per minute at $x_{max}=\frac{N(\alpha-\beta)[1/R_I+t_{W,1}/(K_1-N)]+\max\{p_1(x)\}-p(0)}{\alpha}=253$. This results in \$22.61 for the travel cost of a commuter under UE, and \$283,243 for the total system cost (which does not include the downtown parking fee).

To have a fair comparison between the HV case and the AV case, similar to the HV case, in our model we allow commuters to drive to Area 1, park their AVs and walk to W. However, for our calibrated model the cost associated with this option, i.e., $(\alpha-\beta)\tau_I(x)+\beta(T-x)+p_1(x)+(\alpha''-\beta)t_{1,W}=(\alpha-\beta)\tau_I(x)+\beta(T-x)+p_1(x)+(9-3.9)\times 15/60=(\alpha-\beta)\tau_I(x)+\beta(T-x)+p_1(x)+1.275$, is higher than the cost of dropping off a commuter at W and self-driving to Area 1, i.e., $(\alpha-\beta)\tau_I(x)+\beta(T-x)+p_1(x)+(\alpha-\beta)\tau_W(x)+\alpha' t_{W,1}\le(\alpha-\beta)\tau_I(x)+\beta(T-x)+p_1(x)+(\alpha-\beta)\max\{\tau_W(x)\}+\alpha' t_{W,1}=(\alpha-\beta)\tau_I(x)+\beta(T-x)+p_1(x)+(4.5-3.9)\times 1.21+2.25\times 2/60=(\alpha-\beta)\tau_I(x)+\beta(T-x)+p_1(x)+0.80$, which, as discussed in §2.4.1, is itself higher than the cost of choosing Area 2. Hence, even though the AV commuters have the option to park their AVs in Area 1 and walk to W, they choose not to do it, and continue to park in Area 2. As such, the travel cost of a commuter in the AV case is lower than that in the HV case. However, as

discussed in §2.4.1, the total system cost in the HV case is lower than that in the calibrated model. AVs also increase the total VHT compared to HVs. For the HV case, the total VHT is equal to $\int_0^{253} \lambda_H(x)[\tau_I(x) + \tau_1(x)]dx = 334,232.37$ minutes. For the AV case, the total free-flow travel time for all commuters is equal to $N \times t_{D,2} = 20,000 \times 15 = 300,000$ minutes and total congestion is equal to $\int_0^{75} \lambda_H(x)[\tau_I(x) + \tau_W(x) + \tau_{D,2}(x)]dx = 2,992,515.9$ minutes, so the total VHT is equal to $300,000 + 2,992,515.9 = 3,292,515.90$ minutes.

We also calculate SO for the HV case. Similar to the AV case, we can show that the departure rate from H under SO is equal to the capacity of the most downstream bottleneck, which is the inbound bottleneck in this case, i.e., $\lambda_H(x) = R_I = 4,600/60 = 76.67$ vehicles per minute. So the SO cost is $\int_0^T R_I[(\alpha-\beta)\frac{K_1 t_{W,1}}{K_1 - \int_0^x R_I du} + \beta(T-x) + \alpha t_{W,1}]dx$, where $T = N/R_I + \frac{K_1 t_{W,1}}{K_1 - N} + t_{1,W} = 298$ minutes, yielding $\$252,905.16$.

Next, we do a similar analysis for the HV case under UE2. Similar to the previous case, the departure rate from H, $\lambda_H(x)$, can be derived by solving an ODE. For the model parameters used in the numerical example in §2.4.2, $\lambda_H(x)$ ranges from 91.61 vehicles per minute at time zero to 15.13 vehicles per minute at $x_{max} = \frac{N(\alpha-\beta)[1/R_I + t_{W,1}/(K_1-N)] + \max\{p_1(x)\} - p(0)}{\alpha} = 328$ minutes. The total system cost is equal to $\int_0^{x_{max}} \lambda_H(x)\{(\alpha-\beta)[\tau_I(x) + \tau_1(x) + t_{1,W}] + \beta(T-x)\}dx = \$518,351$. Under SO, since the optimal departure rate from H is equal to the inbound bottleneck capacity, $R_I = 4,600/60 = 76.67$ commuters per minute, the SO cost becomes equal to $\int_0^{x_{max}} R_I[(\alpha - \beta)\frac{K_1 t_{W,1}}{K_1 - \int_0^x R_I du} + \beta(T-x) + (\alpha-\beta)t_{W,1}]dx = \$311,748.76$.

B.4.6 Pricing and Tolling Schemes Benchmarks for the Calibrated Model

Benchmark 1 (Dynamic parking pricing for Area 1 without congestion tolling): In our calibrated model, since Area 1 is closer than Area 2, i.e., $t_{W,1} \leq t_{D,2}$, the social planner wants to fill Area 1 before commuters go to Area 2. To this end, the social planner needs to determine Area 1's parking fee in such a way that the cost of choosing Area 1, $\alpha' t_{W,1} + p_1(x)$, is lower than the cost of choosing Area 2, $\alpha' t_{D,2}$, when Area 1 is not full, i.e., for $x \in [0, x_{1,max}]$. This means that Area 1's parking fee $p_1(x)$ can be at most equal to $\alpha'(t_{D,2} - t_{W,1}) = (2.25/60) \times (15-2) = \$0.49$ for $x \in [0, x_{1,max}]$. By Assumption A1, $p_1(x)$ is increasing in x, hence we assume that $p_1(x) = 0.0169x$ for $x \in [0, x_{1,max} = 29]$, which is a linear increasing function of x with $max\{p_1(x)\} = 0.49$. In

this case, $\alpha' t_{D,2} > \alpha' t_{W,1} + p_1(0)$, so by Proposition 1, we observe UE2, where the travel cost of commuters under UE is equal to \$21.74. In addition, the total system cost, which excludes the parking fees, is equal to $\int_0^{x_{1,max}} \lambda_H(x)[\alpha\tau_I(x)+\alpha\tau_W(x)+\beta(T-y)+\alpha' t_{W,1}]dx+\int_{x_{1,max}}^{x_{max}} \lambda_H(x)[\alpha\tau_I(x)+\alpha\tau_W(x)+\beta(T-y)+\alpha' t_{D,2}+\alpha'\tau_{D,2}(x)]dx = \$432,654$.

Benchmark 2 (A congestion tolling scheme): Similar to Benchmark 1, since Area 1 is closer than Area 2, the social planner wants commuters to first fill Area 1 and follow UE2. To this end the social planner imposes a flat toll π_2 on commuters who want to go to Area 2 before Area 1 becomes full at time $x_{1,max}$. This toll ensures that the cost of choosing Area 1, $\alpha' t_{W,1} + p_1(x)$, must be lower than the cost of choosing Area 2, $\alpha' t_{D,2} + \pi_2$, when Area 1 is not full, i.e., for $x \in [0, x_{max}]$. Area 2's flat toll π_2 must be at least equal to $\alpha'(t_{W,1} - t_{D,2}) + \max\{p_1(x)\} = \alpha'(t_{W,1} - t_{D,2}) + p_1(x_{1,max})$, where $x_{1,max} = 51$ minutes and $p_1(x_{1,max}) = 1.8 + 0.0426x_{1,max} = 3.99$, so the flat toll $\pi_2(x)$ is equal to $(2.25/60) \times (2 - 15) + 3.99 = \$3.50$ for $x \in [0, x_{max} = 74]$. Using this flat toll, $\alpha' t_{D,2} + \pi_2 > \alpha' t_{W,1} + p_1(0)$, so by Proposition 1, we observe UE2, where the travel cost of commuters under UE is equal to \$23.54. In addition, the total system cost, which excludes both Area 1's parking fees and the flat tolls, is equal to $\int_0^{x_{1,max}} \lambda_H(x)[\alpha\tau_I(x) + \alpha\tau_W(x) + \beta(T-y) + \alpha' t_{W,1}]dx + \int_{x_{1,max}}^{x_{max}} \lambda_H(x)[\alpha\tau_I(x) + \alpha\tau_W(x) + \beta(T-y) + \alpha' t_{D,2} + \alpha'\tau_{D,2}(x)]dx = \$416,076$. Note that the congestion toll π_2 does not necessarily need to be a flat toll. In fact, our analysis holds for any $\pi_2(x)$ function that leads to a higher travel cost to Area 2 than that to Area 1.

B.5 Late Arrivals

We discuss the case when commuters can choose to arrive at W after the official work start time T, and show that our main results are robust. As stated in Assumption A7, so far we assume that no commuter would intentionally plan to arrive after T, so the latest group of commuters who leave H at x_{max} arrive at W exactly at T. However, when relaxing Assumption A7, the departure time for commuters who arrive at W at T, which we denote it by x^*, is no longer equal to the latest departure time x_{max}. In fact, in this case $x^* < x_{max}$, and by definition of arrival time at W, $y^* = T < y_{max}$. Proposition B.1 presents various forms of UEs in this case when the penalty of arriving late at W, γ, is not prohibitively high (although it is still higher than the penalty of arriving early, i.e., $\beta \leq \gamma$).

Proposition B.1. *(a) [UE1L] Suppose $\alpha' t_{D,2} \leq \alpha' t_{W,1} + p_1(0)$. There exists a UE which is presented in Table B2, where $E = \frac{\alpha(\int_0^y R_W - a\theta e^{\theta z} dz)^2}{(\alpha+\gamma)(\int_0^y R_W - a\theta e^{\theta z} dz)^2 - \alpha'(\int_0^y a\theta e^{\theta z} dz)^2 + \alpha' a\theta e^{\theta y} R_W y^2} R_W$, $F = \frac{\alpha - p'(x)}{\alpha+\gamma} R_W$, $y_{max} = x_{max} = n/R_W$, $y^* = x^* + \tau_I(x^*) + \tau_W(x^*)$, x^* satisfies $(\alpha - \beta)T - \alpha x^* + \alpha' \tau_{D,2}(x^*) = 0$, and x_1 satisfies $\alpha' t_{W,1} + p_1(x_1) = \alpha'[t_{D,2} + \tau_{D,2}(x_1)]$. In addition, $\lambda_1(y)$ and $\lambda_2(y)$ for cases (i-1)-(i-2) and cases (ii-1)-(ii-2) follow those from case (i) and case (ii) in Table 2.2, respectively.*

Table B2: A characterization of departure rate from H under UE1L.

Condition	$\lambda_H(x)$
(i-1) $K_1 \geq N - R_W y_1$ and $\alpha'[\tau_2(x^*) + t_{D,2}] \leq \alpha' t_{W,1} + p_1(x^*)$	A for $0 \leq x < x^*$ E for $x^* \leq x \leq \min\{x_1, x_{max}\}$ F for $\min\{x_1, x_{max}\} \leq x \leq x_{max}$
(i-2) $K_1 \geq N - R_W y_1$ and $\alpha'[\tau_2(x^*) + t_{D,2}] \leq \alpha' t_{W,1} + p_1(x^*)$	A for $0 \leq x < x_1$ B for $x_1 \leq x \leq x^*$ F for $x^* \leq x \leq x_{max}$
(ii-1) $K_1 < N - R_W y_1$ and $\alpha'[\tau_2(x^*) + t_{D,2}] \leq \alpha' t_{W,1} + p_1(x^*)$	A for $0 \leq x < x^*$ E for $x^* \leq x < x_1$ or $x_2 \leq x \leq x_{max}$ F for $x_1 \leq x < x_2$
(ii-2) $K_1 < N - R_W y_1$ and $\alpha'[\tau_2(x^*) + t_{D,2}] \leq \alpha' t_{W,1} + p_1(x^*)$	A for $0 \leq x < x_1$ B for $x_1 \leq x < x^*$ F for $x^* < x < x_2 + \{x^* - x_2\}^+$ E for $\max\{x^*, x_2\} \leq x \leq x_{max}$

(b) [UE2L] Suppose $\alpha' t_{D,2} \geq \alpha' t_{W,1} + p_1(0)$. There exists a UE which is presented in Table B3, where x_2 satisfies $\alpha' t_{W,1} + p_1(x_2) = \alpha' t_{D,2}$, $y_{max} = x_{max} = N/R_W$, $y^* = T = \frac{\gamma N + R_W(p_1(x_{max}) - p_1(0))}{R_W(\beta+\gamma)}$, and x^* satisfies $\int_0^{x^*} \lambda_H(u)du = R_W y^*$. In addition, $\lambda_1(y)$ and $\lambda_2(y)$ for case (i), cases (ii-1)-(ii-2), and cases (iii-1)-(iii-2) follow those from case (i), case (ii), and case (iii) in Table 2.3, respectively.

Proof of Proposition B.1. (a) The proof of this part is similar to the proof of Proposition 2.1(a), except that late arrivals change the value of $\lambda_H(x)$. We discuss cases (i-1) and (i-2) in Table B2; the proof of the remaining two cases follow a similar procedure.

The work schedule penalty for commuters who arrive at work late, i.e., $y \geq T$, is equal to $\gamma(y - T)$. So for the commuters who choose Area 1. the travel cost is modified as follows: $C_1(x) = (\alpha + \gamma)\frac{\int_0^x \lambda_H(u)du}{R_W} - \alpha x - \gamma T + \alpha' t_{W,1} + p_1(x)$. For the commuters who choose Area 2, the travel cost is modified as follows: $C_2(x) = (\alpha + \gamma)\frac{\int_0^x \lambda_H(u)du}{R_W} - \alpha x - \gamma T + \alpha'[\tau_2(x) + t_{D,2}]$. Thus, the departure rate from H is equal to $E = \frac{\alpha(\int_0^y R_W - a\theta e^{\theta z} dz)^2}{(\alpha+\gamma)(\int_0^y R_W - a\theta e^{\theta z} dz)^2 - \alpha'(\int_0^y a\theta e^{\theta z} dz)^2 + \alpha' a\theta e^{\theta y} R_W y^2} R_W$ when commuters choose Area 2, and $F = \frac{\alpha - p'(x)}{\alpha+\gamma} R_W$ otherwise. Lastly, depending on the level of congestion for commuters who leave at time x^*, there are two cases associated with case (i) of UE1. In particular, if $\tau_2(x^*) \leq t_{W,1} - t_{D,2} + p_1(x^*)/\alpha'$, then the departure time of commuters who arrive at W at T is

Table B3: A characterization of departure rate from H under UE2L.

Condition	$\lambda_H(x)$
(i) $K_1 \geq N - [\frac{N/R_W}{t_{W,1}-t_{D,2}+p_1(x_{max})/\alpha'} + 1]a(e^{\frac{N}{R_W}y_2} - 1)$	B for $0 \leq x \leq x^*$ F for $x^* \leq x \leq x_{max}$
(ii-1) $K_1 \leq \frac{R_W(\alpha - p_1'(x))p_1^{-1}(\alpha'(t_{D,2}-t_{W,1}))}{\alpha-\beta}$ and $\frac{\gamma N + R_W[p_1(x_{max})-p_1(0)]}{(\alpha-\beta)R_W} \leq \frac{K_1}{R_W}$	B for $0 \leq x < x^*$ F for $x^* \leq x < x_{1,max}$ 0 for $x_{1,max} \leq x < x_2$ E for $x_2 \leq x \leq x_{max}$
(ii-2) $K_1 \leq \frac{R_W(\alpha - p_1'(x))p_1^{-1}(\alpha'(t_{D,2}-t_{W,1}))}{\alpha-\beta}$ and $\frac{\gamma N + R_W[p_1(x_{max})-p_1(0)]}{(\alpha-\beta)R_W} > \frac{K_1}{R_W}$	B for $0 \leq x < x_{1,max}$ 0 for $x_{1,max} \leq x < x_2$ A for $x_2 \leq x \leq x^*$ E for $x^* \leq x \leq x_{max}$
(iii-1) $\frac{R_W(\alpha - p_1'(x))p_1^{-1}(\alpha'(t_{D,2}-t_{W,1}))}{\alpha-\beta}$ $< K_1 < N - [\frac{N/R_W}{t_{W,1}-t_{D,2}+p_1(x_{max})/\alpha'} + 1]a(e^{\frac{N}{R_W}y_2} - 1)$ and $\frac{\gamma N + R_W[p_1(x_{max})-p_1(0)]}{(\alpha-\beta)R_W} \leq \frac{K_1}{R_W}$	B for $0 \leq x < x^*$ F for $x^* \leq x < x_2$ E for $x_2 \leq x \leq x_{max}$
(iii-2) $\frac{R_W(\alpha - p_1'(x))p_1^{-1}(\alpha'(t_{D,2}-t_{W,1}))}{\alpha-\beta}$ $< K_1 < N - [\frac{N/R_W}{t_{W,1}-t_{D,2}+p_1(x_{max})/\alpha'} + 1]a(e^{\frac{N}{R_W}y_2} - 1)$ and $\frac{\gamma N + R_W[p_1(x_{max})-p_1(0)]}{(\alpha-\beta)R_W} > \frac{K_1}{R_W}$	B for $0 \leq x < x_2$ A for $x_2 \leq x \leq x^*$ E for $x^* \leq x \leq x_{max}$

sooner than the earliest time some commuters choose Area 1, i.e., $x^* \leq x_1$, and we have case (i-1). Otherwise, we have case (i-2).

(b) The proof of this part is similar to the proof of Proposition 2.1(b). For all cases of UE1L, we derive $y^* = T$ by letting $C_1(0) = T + \alpha' t_{W,1} + p_1(0)$ equal to $C_1(x_{max}) = \gamma N/R_W - \gamma T + \alpha' t_{W,1} + p_1(x_{max})$. Hence, $y^* = T = \frac{\gamma N + R_W[p_1(x_{max})-p_1(0)]}{(\alpha-\beta)R_W}$. In addition, by (2.1), x^* satisfies $y^* = \frac{\int_0^{x^*} \lambda_H(u)du}{R_W}$. Similar to the proof of part (a), the departure rate from H, $\lambda_H(x)$, changes for late arrivals. For case (i), since all commuters choose Area 1, $\lambda_H(x)$ follows B before time x^*, and F after time x^*. However, for each case (ii) and (iii) under UE2, there exist two cases: when $K_1/R_W \geq \frac{\gamma N + R_W[p_1(x_{max})-p_1(0)]}{(\alpha-\beta)R_W}$, $x^* \leq x_2$ and we observe cases (ii-1) and (iii-1), and otherwise we observe cases (ii-2) and (iii-2), respectively. □

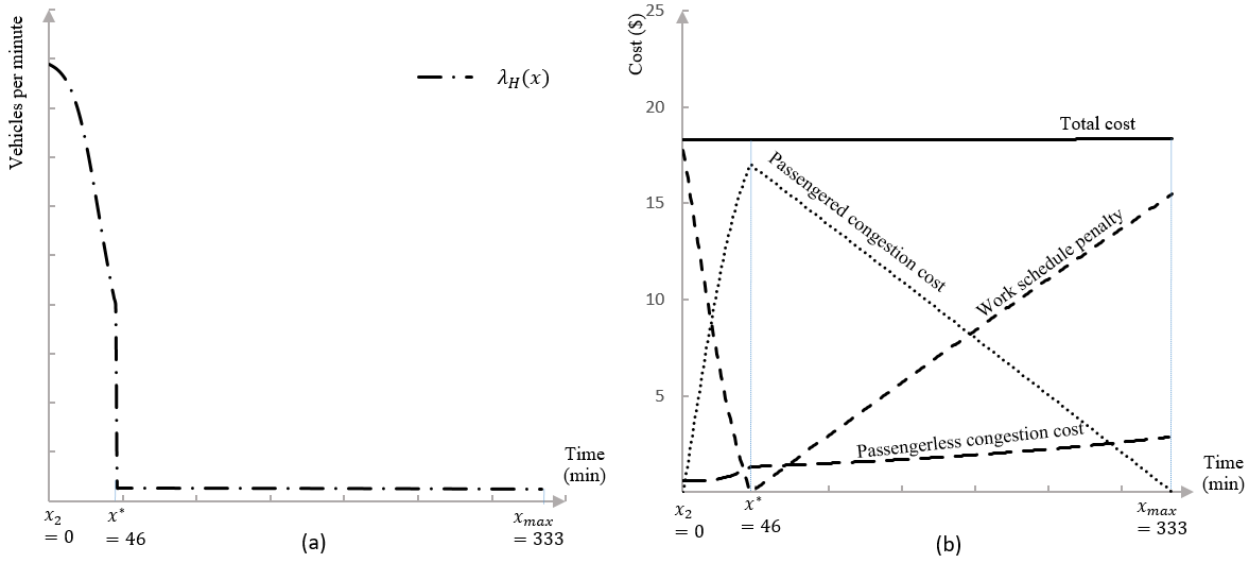
Similar to Proposition 2.1, Proposition B.1 indicates that there exist two forms of UE: UE1L and UE2L. In fact, under UE1L and UE2L the departure rate to the parking areas, i.e., $\lambda_1(y)$ and $\lambda_1(y)$, are the same as those under UE1 and UE2, respectively. However, under UE1L and UE2L, the departure rate from H, $\lambda_H(x)$, for the commuters who arrive at W after time T differs from that in Proposition 2.1. Hence, depending on the model parameters, each case of UE1 or UE2, except case (i) of UE2, is divided to two cases under UE1L or UE2L. Here we focus on UE1L and its difference with UE1. The reason why each case of UE1 is divided into two cases under

UE1L is as follows: In both of these cases, commuters choose Area 2 until time x_1 when some commuters go to Area 2, so depending on the level of downtown congestion and parking fees at time x^*, the earliest departure time of commuters who choose Area 1, x_1, can be before or after time x^*, creating two separate cases. In particular, in case (i-1) of UE1L, the cost of traveling to Area 2, $\alpha'[\tau_{D,2}(x^*) + t_{D,2}]$, is still lower than that to Area 1, $\alpha' t_{W,1} + p_1(x^*)$, so $x_1 > x^*$; otherwise, $x_1 \leq x^*$ which is presented in case (i-2). Note that the condition $K_1 \geq N - R_W y_1$ is exactly the condition we have for case (i) of UE1 in Proposition 2.1. Similarly, case (ii-1) and (ii-2) correspond to when $x_1 > x^*$ and $x_1 \leq x^*$, respectively.

We illustrate this proposition for our calibrated model in Figure B.4. Following Pi et al. (2019), we assume that the penalty of arriving late at W, γ, is \$15.2 per hour. Using this value of γ and the model parameters discussed in Table 2.1, we observe case (i-1) of UE1L.[3] As Figure B.4(a) shows, the departure rate from H for the commuters who leave before $x^* = 46$ minutes is identical to that in Figure 2.2(a), but for the commuters who leave after $x^* = 46$ minutes, there is a significant decrease in $\lambda_H(x)$. This is because commuters who depart H after x^* and arrive at W late incur a higher work schedule penalty per hour than those who arrive early (remember that $\beta = \$3.9 \leq \gamma = \$15.2$ per hour). Hence, the rate of the commuters who leave after x^* is significantly lower than that for the commuters who leave before x^*. In fact $\lambda_H(x)$ decreases from 199.02 commuters per minute to 13.17 commuters per minute immediately after $x^* = 46$ minutes. The low departure rate of late commuters expands the departure window from H from 75 minutes in Figure 2.2(a) to 333 minutes in this case. However, since the departure rates to the parking areas stay the same as those in Figure 2.2(a), the entire duration of the morning commute remains the same, i.e., $y_{max} = 333$ in both cases. Figure B.4(b) depicts the different cost components that commuters incur in our calibrated model. In this case, the total travel cost incurred by each commuter decreases from \$22.23 in Figure 2.3(b) to \$18.33 for two major reasons. First, the reduced departure rate from H after x^* decreases the congestion costs. Second, since some commuters arrive after time T, the work schedule penalty that commuters experience is more balanced, and the maximum work schedule delay is reduced. This is expected because commuters are now given an option to arrive late.

[3]The calibrated model parameters satisfy the condition for case (i-1) of UE1L: $\alpha' t_{D,2} = 2.25 \times 15/60 \leq 2.25 \times 2/60 + 1.80 = \alpha' tW, 1 + p_1(0)$. In addition, since Area 1 is not chosen by any of the commuters (i.e., $y_1 > y_{max} = 333$), the conditions for case (i-1) of UE1 is satisfied: $K_1 = 10,000 > N - R_W y_{max} = 20,000 - 60 \times 333 > N - R_W y_1$ and $\alpha'[\tau_{D,2}(x^*) + t_{D,2}] = 2.25 \times (20.07 + 15)/60 \leq 2.25 \times 2/60 + 3.76 = \alpha' t_{W,1} + p_1(x^*)$.

Figure B.4: An illustration of UE1L in the calibrated model: (a) departure rate from H, and (b) different cost components commuters incur.



Notes. In (b), the total cost $C_2(x)$ has three components: the work schedule penalty, which is equal to $\beta[T-y]$ for commuters who arrive at W before time T and $\gamma[y-T]$ for commuters who arrive at W after time T, the passengered congestion cost, which is equal to $\alpha[\tau_I(x)+\tau_W(x)]$, and the passengerless congestion cost, which is equal to $\alpha'[t_{D,2}+\tau_{D,2}(x)]$.

We observe from Figure B.4 that the main insights from §2.4.1 continue to hold in this extension. In particular, most commuters prefer to leave early (81% of commuters leave before x^*), creating a spike in the passengered congestion cost, and all commuters choose to park in Area 2. This verifies the results shown in §2.4.1 that under UE, AVs eliminate the demand for downtown parking, expand the duration of the morning commute, and create more congestion than HVs. Since the latter two effects are not desirable social outcomes, in Proposition B.2, we investigate the effect of late arrivals on the SO case.

Proposition B.2. *(a) [SO1L] Suppose $t_{D,2} \geq t_{W,1}$. There exists an SO, presented in Table 2.4. Table B4 presents the expression for $x^* = T$. In this table, cases (ii-1)-(ii-3) correspond to case (ii) in Table 2.4. In addition, $y_2 = x_2 = K_1/R_W$, $x_{max} = y_{max} = \frac{N}{R_W}$ for case (i), and $x_{max} = y_{max} = \frac{N+a(x_D-x_2)}{R_W}$ for cases (ii-1)-(ii-3).*
(b) [SO2L] Suppose $t_{D,2} < t_{W,1}$. There exists an SO, presented in Table 2.5. Table B5 presents the expressions for $x_{max} = y_{max} = \frac{N+ax_D}{R_W}$ and $x^* = T$. In this table, cases (i-1)-(i-2) correspond to case (i) and (ii-1)-(ii-2) correspond to case (ii) of Table 2.4.

Table B4: A characterization of departure time from H for commuters who arrive at W at T under SO1L.

Condition	x^*
(i) $K_1 \geq N$	$\frac{N\gamma}{R_W(\beta+\gamma)}$
(ii-1) $K_1 < N \leq K_1(\beta+\gamma)/\gamma$	$\frac{R_W\gamma x_{max} - a\gamma(x_D - x_2)}{R_W(\beta+\gamma)}$
(ii-2) $K_1(\beta+\gamma)/\gamma < N \leq [(R_W - a)x_2 - ax_D](\beta+\gamma)/\gamma$	$\frac{R_W\gamma x_{max} - a\beta x_2 - \gamma x_D}{(R_W - a)(\beta+\gamma)}$
(ii-3) $N > [(R_W - a)x_2 - ax_D](\beta+\gamma)/\gamma$	$\frac{R_W\gamma x_{max} + a\beta(x_D - x_2)}{R_W(\beta+\gamma)}$

Table B5: A characterization of departure time from H for commuters who arrive at W at T under SO2L.

Condition	x^*
(i-1) $K_1 \geq [\frac{N-(R_W-a)x_D}{R_W y_2} + 1]y_1 R_W$ and $N \leq (R_W - a)(\beta+\gamma)x_D/\gamma$ or (ii-1) $K_1 < [\frac{N-(R_W-a)x_D}{R_W y_2} + 1]y_1 R_W$ and $N \leq (R_W - a)(\beta+\gamma)x_D/\gamma$	$\frac{R_W\gamma x_{max} + a\beta x_D}{R_W(\beta+\gamma)}$
(i-2) $K_1 \geq [\frac{N-(R_W-a)x_D}{R_W y_2} + 1]y_1 R_W$ and $N > (R_W - a)(\beta+\gamma)x_D/\gamma$ or (ii-2) $K_1 < [\frac{N-(R_W-a)x_D}{R_W y_2} + 1]y_1 R_W$ and $N > (R_W - a)(\beta+\gamma)x_D/\gamma$	$\frac{R_W\gamma x_{max} - a\gamma x_D}{(R_W - a)(\beta+\gamma)}$

Proof of Proposition B.2. (a) The proof of this part is identical to that of Proposition 2.2(a), except that depending on the value of K_1 and N, the value of x^* changes. For case (i) of SO1, since the total number of commuters, N, is so low that Area 1 can accommodate all commuters, there exists only one corresponding case under SO1L, which is case (i) in Table B4. However, for case (ii) of SO1, depending on the values of N and K_1, we have three cases (ii-1), (ii-2), and (ii-3), which correspond to $x^* \leq x_2$, $x_2 < x^* \leq x_D$, and $x^* > x_D$, respectively. We next derive the expression of x^* for case (i) under SO1L; the expression of x^* for the remaining cases can be derived similarly. For case (i), we find the value of x^* that minimizes the aggregate cost, i.e., $\min_{x^*} \int_0^{x^*} R_W\beta(T-x)dx + \int_{x^*}^{x_{max}} R_W\gamma(x-T)dx + \int_0^{x_{max}} R_W\alpha' t_{W,1}dx = \min_{x^*}[R_W\beta(x^*)^2]/2 + [R_W\gamma(x_{max}-x^*)^2]/2 + N\alpha' t_{W,1}$. Taking the first derivative, we have $(\beta+\gamma)x^* = \gamma N/R_W$, yielding $x^* = \frac{\gamma N}{(\beta+\gamma)R_W}$.

(b) The proof of this part is similar to the proof of part (a). For each case of SO2, we have two cases under SO2L. We derive the expression of x^* for cases (i-1) and (ii-1); the expression of x^* for cases (i-2) and (ii-2) can be calculated similarly. Suppose $N \leq (R_W - a)(\beta+\gamma)x_D/\gamma$. Then we find x^* that minimizes the aggregate cost, i.e., $\min_{x^*} \int_0^{x^*} R_W\beta(T-x)dx + \int_{x^*}^{x_{max}} R_W\gamma(x-T)dx - \int_0^{x_D} a\beta(T-x)dx + \int_0^{x_{max}} \lambda_1(y)\alpha' t_{W,1} + \lambda_2(y)\alpha'[t_{D,2} + \tau_{D,2}(x)]dx$. We can simplify this equation as $\min_{x^*}[R_W\beta(x^*)^2]/2 + [R_W\gamma(x_{max}-x^*)^2]/2 - a\beta x_D(x^* - x_D/2)$. Taking the first derivative, we have $(\beta+\gamma)x^* = \gamma x_{max} + a\beta x_D/R_W$, yielding $x^* = \frac{\gamma x_{max}R_W + a\beta x_D}{(\beta+\gamma)R_W}$. Since $N \leq (R_W - a)(\beta+\gamma)x_D/\gamma$, $x^* \geq x_D$. □

Proposition B.2 indicates that SO in this case is identical to those in Proposition 2.2, while the departure time from H that results in arriving at W at time T, x^*, can vary depending on the model parameters. In particular, under SO1L, there exist three different possible cases for x^* (represented by cases (ii-1)-(ii-3)) that correspond to the case (ii) of SO1, and under SO2L there are two different cases associated with each case of SO2. In our calibrated model, we observe case (ii-3) of SO1L.[4] All the departure time from H and parking locations decisions are exactly as those depicted in Figure 2.5, but by allowing late arrivals the total system cost decreases from $\$218,728$ to $\$180,647$. The main insights from §2.5.1 continue to hold, and the late arrivals of some commuters do not have an effect on the social planner's decisions. In addition, since SO remains unchanged, the parking fees and congestion tolls follow the same pattern as presented in Proposition 2.3.

[4]In addition to the conditions of case (ii) of SO1, our calibrated model satisfies the condition for case (ii-3) of SO1L: $N = 20,000 > [(60 - 0.2) \times 167 - 0.2 \times 215](3.9 + 15.2)/15.2 = [(R_W - a)x_2 - ax_D](\beta + \gamma)/\gamma$.

Appendix C

Additional Material for Chapter 3

C.1 Notation Summary

Table C1: Summary of notation

Symbol	Definition
j	Index for mode of transportation ($j \in \{C, R\}$)
x	Departure time from H
x_{max}	Latest departure time from H
x_j	Earliest departure time of commuters who choose Area $j \in \{1, 2\}$
R_I	Inbound bottleneck capacity
R_C	Downtown roads capacity
R_R	Drop-off rate at W
R_R^*	Optimal drop-off rate at W
N	Total number of commuters
T	Official start time at work for all commuters
$p(x)$	Downtown parking fee for commuters who leave H at x
$u(x)$	TNC's profit margin from a commuter who leaves H at x
u_0	Minimum TNC profit margin under case (i) of equilibrium
u_0	Minimum TNC profit margin under case (ii) of equilibrium
$\pi(x)$	Drop-off congestion toll for ride-hailing commuters who leave H at x
$\lambda_H(x)$	Departure rate from H at x
$\lambda_j(x)$	Departure rate from H for commuters who choose transportation mode $j \in \{C, R\}$ and leave H at x
$\tau_I(x)$	Inbound delay time for a commuter who leaves H at x
$\tau_R(x)$	Drop-off congestion time for a ride-hailing commuter who leaves H at x
$\tau_C(x)$	Downtown congestion time for a conventional commuter who leaves H at x
$t_{H,I}$	Free-flow travel time from H to I
$t_{I,D}$	Free-flow travel time from I to D
t_C	Walking time from parking areas to W
t_C	Walking time from drop-off zone to W
α	Unit cost of driving time for commuters
α'	Unit cost of walking time for commuters
β	Unit cost for early schedule delay
δ	Negative externality factor of ride-hailing drop-offs
$C_j(x)$	Total cost of a commuter who leaves H at x and chooses transportation mode $j \in \{C, R\}$
C	The daily travel cost for an individual commuter under SO
Γ	Minimum number of ride-hailing vehicles present in the drop-off zone that creates negative externality for conventional commuters

C.2 Proofs

Lemma C.1. *If $\tau_I(x) \neq 0$ for all $x \in [0, x_{max}]$, then $\tau_I(x)+\tau_C(x) = \frac{\delta \int_0^x \lambda_R(u)-R_R du+\int_0^x \lambda_C(u)-R_C du}{R_C}$ and $\tau_I(x) + \tau_R(x) = \frac{\int_0^x \lambda_R(u)-R_R du}{R_R}$.*

Proof. We prove $\tau_I(x)+\tau_R(x) = \frac{\int_0^x \lambda_R(u)-R_R du}{R_R}$. The proof of $\tau_I(x)+\tau_C(x) = [\delta \int_0^x \lambda_R(u)-R_R du+ \int_0^x \lambda_C(u) - R_C du]/R_C$ is similar and is eliminated to avoid repetition.

By A3, $\tau_I(x) = \frac{\int_0^x \lambda_H(u)-R_I du}{R_I}$ and $\tau_R(x) = \frac{\int_0^{x+\tau_I(x)} \lambda_R(u)R_I/\lambda_H(u)-R_R du}{R_R}$, hence $\tau_I(x) + \tau_R(x) = \frac{\int_0^x \lambda_H(u)-R_I du}{R_I} + \frac{\int_0^{x+\tau_I(x)} \lambda_R(u)R_I/\lambda_H(u)-R_R du}{R_R}$. After some simplifications, we have $\tau_I(x) + \tau_R(x) = \frac{\int_0^{\int_0^x \lambda_H(u)/R_I du} \lambda_R(u)R_I/\lambda_H(u)-R_R du}{R_R}$. The numerator of this expression, i.e., $\frac{\int_0^{\int_0^x \lambda_H(u)/R_I du} \lambda_R(u)R_I/\lambda_H(u)-R_R du}{R_R}$, is the total number of ride-hailing vehicles that leave H at x or earlier and are present at the drop-off zone by $x + \tau_I(x) = \int_0^x \lambda_H(u)du/R_I$. By definition of $\lambda_R(x)$, i.e., departure rate of ride-hailing commuters who leave H at x, the numerator is equal to $\int_0^x \lambda_R(x) - R_R du$. □

Proof of Proposition 3.1. We prove this proposition in two main steps: In the first step, assuming that $u(x)$ is given, we characterize an equilibrium such that Conditions 1 and 2, stated in §3.4, are satisfied. In the second step, we derive $u(x)$ that satisfies Condition 3.

Step 1: There exist two types of equilibrium that satisfies Conditions 1 and 2 depending on the decisions made by commuters who leave H at time zero:

(i) Suppose $C_C(0) < C_R(0)$, i.e., $(\alpha' - \beta)t_C + p(0) < (\alpha' - \beta)t_R + u(0)$. In this case commuters who leave H at $x \in [0, x_R)$, choose to drive. At x_R, as downtown congestion $\tau_C(x)$ for conventional commuters increases, $C_C(x_R)$ becomes equal to $C_R(x_R)$, and some commuters choose to use ride-hailing. Next, we derive the values of $\lambda_C(x)$ and $\lambda_R(x)$. Particularly, for $x \in [0, x_R)$, since all commuters choose to drive, $\lambda_R(x) = 0$. To derive $\lambda_C(x)$, by Condition 2, we set $\frac{\partial C_C}{\partial x}$ equal to zero, and derive $\lambda_C(x) = \frac{\alpha-p'(x)}{\alpha-\beta}R_C$. Similarly, for $x \in [x_R, x_{max}]$, by Condition 2, we derive $\lambda_R(x) = \frac{\alpha-u'(x)}{\alpha-\beta}R_R$. In addition, we derive $\lambda_C(x) = \frac{\alpha-p'(x)}{\alpha-\beta}R_C$ for $x \in [x_R, \bar{x}_R)$, and $\lambda_C(x) = \frac{\alpha-p'(x)}{\alpha-\beta}R_C - \frac{\delta[\beta-u'(x)]}{\alpha-\beta}R_R$ for $x \in [\bar{x}_R, x_{max}]$.

(ii) Suppose $C_C(0) \geq C_R(0)$, i.e., $(\alpha' - \beta)t_C + p(0) \geq (\alpha' - \beta)t_R + u(0)$. In this case commuters who leave H at $x \in [0, x_C)$, choose to use ride-hailing. At x_C, as drop-off congestion $\tau_R(x)$ for ride-hailing commuters increases, $C_R(x_C)$ becomes equal to $C_C(x_C)$, and some commuters choose to drive. Next, we derive the values of $\lambda_C(x)$ and $\lambda_R(x)$. Particularly, to derive $\lambda_R(x)$, by Condition 2,

we set $\frac{\partial C_R}{\partial x}$ equal to zero, and derive $\lambda_R(x) = \frac{\alpha - u'(x)}{\alpha - \beta} R_R$. For $x \in [0, x_C)$, since all commuters choose to use ride-hailing, $\lambda_C(x) = 0$. Similar to part (i), to derive $\lambda_C(x)$, for $x \in [x_C, \max\{x_C, \bar{x}_R\})$, by Condition 2, we derive $\lambda_C(x) = \frac{\alpha - p'(x)}{\alpha - \beta} R_C$, and $\lambda_C(x) = \frac{\alpha - p'(x)}{\alpha - \beta} R_C - \frac{\delta[\beta - u'(x)]}{\alpha - \beta} R_R$ for $x \in [\max\{x_C, \bar{x}_R\}, x_{max}]$.

Step 2: We find $u(x)$ such that the profit of the TNC is maximized (i.e., Condition 3 is satisfied). We first derive $u(x)$ for case (i) and then discuss case (ii).

In case (i), by substituting the value of $\lambda_R(x)$, i.e., $\frac{\alpha - u'(x)}{\alpha - \beta} R_R$ for $x \in [x_R, x_{max}]$, and zero otherwise, into (3.4) we have: $\max \int_{x_R}^{x_{max}} \frac{\alpha - u'(x)}{\alpha - \beta} R_R \{u(x) - \frac{\alpha[\beta x - u(x) + u(x_R)]}{\alpha - \beta}\} dx$. To find $u(x)$ that maximizes this profit, we set the derivative of this expression with respect to $u(x)$ equal to zero. In particular, after some simplifications, we have:

$$-u''(x)[\frac{(2\alpha - \beta)u(x) - \alpha\beta x - \alpha u(x_R)}{\alpha - \beta}] + \frac{[\alpha - u'(x)](2\alpha - \beta)u'(x)}{\alpha - \beta} = 0. \tag{C.1}$$

To solve this second degree differential equation, let $q(x) = u'(x) = \frac{du}{dx}$. We can state $u''(x)$ as a function of $q(x)$ as follows: $u''(x) = \frac{du'(x)}{dx} = \frac{dq(x)}{dx} = \frac{dq(x)}{du} \times \frac{du(x)}{dx} = q\frac{dq(x)}{du}$. By substituting $u'(x)$ and $u''(x)$ into (C.1), we can simplify this equation as follows: $-q(x)\frac{dq(x)}{du}[\frac{(2\alpha - \beta)u(x) - \alpha\beta x - \alpha u(x_R)}{\alpha - \beta}] + \frac{[\alpha - q(x)](2\alpha - \beta)q(x)}{\alpha - \beta} = 0$. This means that either $q(x) = 0$ or $u(x)$ satisfies $-\frac{dq(x)}{du}[(2\alpha - \beta)u(x) - \alpha\beta x - \alpha u(x_R)] + [\alpha - q(x)](2\alpha - \beta) = 0$. First, if $q(x)$ (which is equal to $u'(x)$) is equal to zero, then $u(x) = u(x_R)$ for $x \in [x_R, x_{max}]$. Note that $u(x_R)$ is the value of $u(x)$ that guarantees that $C_R(x)$ becomes equal to $C_C(x)$ at x_R, i.e., $(\alpha' - \beta)t_R + u(x_R) = (\alpha' - \beta)t_C + p(x_R) + (\alpha - \beta)\tau_C(x_R)$. Simplifying this equation, we have $u(x_R) = (\alpha' - \beta)(t_C - t_R) + p(0) + \beta x_R$. Next, if $q(x) \neq 0$, then we find $u(x)$ that satisfies $-\frac{dq(x)}{du}[(2\alpha - \beta)u(x) - \alpha\beta x - \alpha u(x_R)] + [\alpha - q(x)](2\alpha - \beta) = 0$. By integrating this equation with respect to u, i.e., $\int -\frac{dq(x)}{du}[(2\alpha - \beta)u(x) - \alpha\beta x - \alpha u(x_R)] + [\alpha - q(x)](2\alpha - \beta)du = 0$, we have $\alpha(2\alpha - \beta)u(x) - (2\alpha - \beta)u(x)q(x) + [\alpha\beta x + \alpha u(x_R)]q(x) + C = 0$, where C is a constant value. After substituting $q(x)$ with $u'(x)$, we have $C = -\alpha(2\alpha - \beta)u(x) + (2\alpha - \beta)u(x)u'(x) - [\alpha\beta x + \alpha u(x_R)]u'(x) \leq 0$. In addition, the objective function in (3.4) can be stated as $\frac{R_R}{\alpha - \beta}\{-(\alpha - \beta)C - \frac{\alpha^2[\beta x + u(x_R)]}{\alpha - \beta}\}$, which is negative. Thus, the first solution of the ODE, i.e., $u'(x) = 0$, which results in a positive total profit for the TNC, is the optimal solution, and $u^*(x) = u(x_R) = (\alpha' - \beta)(t_C - t_R) + p(0) + \beta x_R$ for $x \in [x_R, x_{max}]$.

Lastly, we characterized x_{max}, $\bar{x}_R$ and x_R. To characterize x_{max}, we solve the following

equation: $\int_0^{x_R} \lambda_C(x)dx + \int_{x_R}^{x_{max}} \lambda_C(x) + \lambda_R(x)dx = N$. In other words, the total number of commuters who leave H during the morning commute must be equal to N. After substituting $\lambda_C(x)$ and $\lambda_R(x)$ with the values derived previously and simplifying the equation, we have $x_{max} = \frac{N(\alpha-\beta)+R_C[p_(x_{max})-p(0)]+\alpha R_R x_R-\delta\beta R_R\bar{x}_R}{\alpha R_C+(\alpha-\delta\beta)R_R}$. To characterize $\bar{x}_R$, we find $\bar{x}_R$ that satisfies $\int_{x_R}^{\bar{x}_R} {}_R(x) - R_R dx = \Gamma$. After substituting $\lambda_R(x)$ with $\frac{\alpha R_R}{\alpha-\beta}$, we have $\bar{x}_R = x_R + \frac{\alpha-\beta}{\beta R_R}\Gamma$. To characterize x_R, we find x_R that maximizes the total profit of the TNC is equal, stated in (3.4), i.e., $\max_{x_R} \int_{x_R}^{x_{max}} \lambda_R(x)[u(x) - \alpha\tau_R(x)]dx = (x_{max} - x_R)\frac{\alpha R_R}{\alpha-\beta}[(\alpha'-\beta)(t_C - t_R) + p(0) + \frac{\beta^2}{\alpha-\beta}x_R + \frac{\alpha\beta}{2(\alpha-\beta)}(x_{max}+x_R)]$. To find the maximum, we set the derivative of this equation with respect to x_R equal to zero, i.e., $\frac{(\alpha-\beta)R_R}{\alpha R_C+(\alpha-\beta)R_R}[(\alpha'-\beta)(t_C-t_R)+p(0)-\beta^2 x_R/(\alpha-\beta)+\alpha\beta x_{max}/(\alpha-\beta)] - [(\alpha'-\beta)(t_C-t_R)+p(0)-\beta^2 x_R/(\alpha-\beta)+\alpha\beta x_R/(\alpha-\beta)] - \frac{\beta^2}{\alpha-\beta}(x_{max}-x_R) = 0$. After some simplification, we characterize $x_R^* \geq 0$ as follows: $\Big[\{R_C[p(x_{max}) - p(0)] + N(\alpha-\beta) - \delta(\alpha-\beta)\Gamma\}\{\alpha(\alpha-\beta)R_R - \beta[\alpha R_C + (\alpha-\delta\beta)R_R] - (\alpha-\beta)\alpha R_C[(\alpha'-\beta)(t_C-t_R)+p(0)]/\beta\}/[(\alpha-2\beta)(\alpha R_C)^2]\Big]^+$.

We follow a similar approach to find $u^*(x)$ for case (ii). In case (ii), $\lambda_C(x) = 0$ for $x \in [0, x_C)$, and $\lambda_C(x) > 0$ for $x \in [x_C, x_{max}]$, so we derive $u^*(x)$ for each of these departure time intervals separately. When $x \in [x_C, x_{max}]$, since $\lambda_C(x)$ and $\lambda_R(x)$ in this case follow the same function as those in case (i) when $x \in [x_R, x_{max}]$, similar to the previous case we can show that $u'(x) = 0$ and $u(x) = u(x_C)$. To avoid repetition, we do not provide this proof. When $x \in [0, x_C)$, under equilibrium, u(x) must be so low that $C_R(x) < C_C(x)$, so commuters do not choose to drive. However, since the TNC maximizes its profit, it sets $u(x)$ as high as possible. In particular, at $x = 0$, the maximum $u(0)$ is the lowest dollar value less than $u_0 = \$[(\alpha'-\beta)(t_C-t_R)+p(0)]$(the cost difference between the two transportation modes), i.e., $u(0) = \lceil 100u_0 - 1\rceil/100$. For $x \in (0, x_C)$, the rate of increase in $C_R(x)$ (which is equal to $(\alpha-\beta)\tau_R(x) + (\alpha'-\beta)t_R + \beta(T-x) + u(x)$) can at most be equal to that in $C_C(x)$ (which is equal to $(\alpha-\beta)\tau_C(x) + (\alpha'-\beta)t_C + \beta(T-x) + p(0)$, where $\tau_C(x) = \frac{\delta R_R}{R_C}$).[1] After taking the derivatives and simplifying the equations, we have the following for the derivative of ride-hailing fare: $u'(x) = \frac{\beta(\delta R_R - R_C)}{\delta R_R}$ for $x \in [0, x_C)$. This means that $\lambda_R(x) = \frac{\alpha-u'(x)}{\alpha-\beta}R_R = R_R + \frac{\beta/\delta}{(\alpha-\beta)}R_C$ and $u(x) = u(0) + \frac{\beta(\delta R_R - R_C)}{\delta R_R}x$ for $x \in [0, x_C)$.

[1] In the expression for $C_C(x)$, since no commuter has yet chosen to drive, we use $p(0)$, which is the parking fee for the first group of commuters who choose to drive. In addition, due to the externality effect of the drop-off congestion, even the first group of commuters who head to the parking area experience some downtown congestion.

We next characterize x_{max}, $\bar{x}_R$ and x_C. To characterize x_{max}, we solve the following equation: $\int_0^{x_C} \lambda_R(x)dx + \int_{x_C}^{x_{max}} \lambda_C(x) + \lambda_R(x)dx = N$. In other words, the total number of commuters who leave H during the morning commute must be equal to N. After substituting $\lambda_C(x)$ and $\lambda_R(x)$ with the values derived previously and simplifying the equation, we have $x_{max} = \frac{N(\alpha-\beta)+R_C[p_(x_{max})-p(0)]+[(\alpha-\beta/\delta)R_C+(\alpha-\beta)R_R]x_C-\delta\beta R_R\bar{x}_R}{\alpha R_C+(\alpha-\delta\beta)R_R}$. To characterize $\bar{x}_R$, we find $\bar{x}_R$ that satisfies $\int_0^{\bar{x}_R} \lambda_R(x) - R_R dx = \Gamma$. After substituting $\lambda_R(x)$ with the values calculated above, we have $\bar{x}_R = \frac{(\alpha-\beta)\Gamma}{(\alpha-\beta)R_R+\beta R_C/\delta}$ if $\bar{x}_R < x_C$, and $\bar{x}_R = \frac{(\alpha-\beta)\Gamma}{\alpha R_R} + [\beta/\alpha - \frac{\beta R_C}{\alpha\delta R_R}]x_C$, otherwise. To characterize x_C, we find x_C that maximizes the total profit of the TNC is equal, stated in (3.4), i.e., $\max_{x_C} \int_0^{x_{max}} \lambda_R(x)[u(x)-\alpha\tau_R(x)]dx = \int_0^{x_C}(R_R+\frac{R_C\beta/\delta}{\alpha-\beta})[u(0)-\frac{\beta(R_C-\delta R_R)}{\delta R_R}x]dx+\int_{x_C}^{x_{max}} \frac{\alpha R_R}{\alpha-\beta}[u(0)-\frac{\beta(R_C-\delta R_R)}{\delta R_R}x_C]dx$. To find the maximum, we set the derivative of this equation with respect to x_C equal to zero. After some simplification, we characterize $x_C^* \geq 0$ as follows: $\Big[\delta R_R\{(\alpha R_C - \beta\delta R_R)[R_R(\alpha-\beta)+R_C\beta/\delta]u(0) - [R_C[p(x_{max})-p(0)]+N(\alpha-\beta)](R_C-\delta R_R)\alpha\beta/\delta\}/\{\beta(R_C-\delta R_R)[(\alpha-\beta)R_R+R_C\beta/\delta][\alpha R_C-(\alpha+\delta\beta)R_R]\}\Big]^+$. □

Proof of Proposition 3.2. We first characterize $\lambda_C(x)$ and $\lambda_R(x)$ that minimize the total cost presented in (3.3), and then derive $u(x)$.

By Lemma C.1, $\tau_I(x) + \tau_C(x+\tau_I(x)) = \tau_C(x)$, so we can simplify (3.3) as follows:

$$\int_0^{x_{max}} \lambda_C(x)[(\alpha-\beta)\tau_C(x)+(\alpha'-\beta)t_C+\beta(T-x)] + \lambda_R(x)[(\alpha-\beta)\tau_R(x)+(\alpha'-\beta)t_R+\beta(T-x)+u(x)]dx. \tag{C.2}$$

At any $x \in [0, x_{max}]$, the social planner assigns all commuters who leave H at x to the transportation mode that has a lower cost associated with it. In particular, three different scenarios can happen. In the first scenario, $(\alpha'-\beta)t_C+(\alpha-\beta)\tau_C(x) = (\alpha'-\beta)t_R+u(x)+(\alpha-\beta)\tau_R(x)$, so the social planner assigns commuters to both modes of transportation, i.e., $\lambda_C(x) > 0$ and $\lambda_R(x) > 0$. In the second scenario, $(\alpha'-\beta)t_C+(\alpha-\beta)\tau_C(x) > (\alpha'-\beta)t_R+u(x)+(\alpha-\beta)\tau_R(x)$, so the social planners wants all commuters to use ride-hailing, i.e., $\lambda_C(x) = 0$ and $\lambda_R(x) > 0$. In the third scenario, $(\alpha'-\beta)t_C+(\alpha-\beta)\tau_C(x) < (\alpha'-\beta)t_R+u(x)+(\alpha-\beta)\tau_R(x)$, so the social planners wants all commuters to drive, i.e., $\lambda_C(x) > 0$ and $\lambda_R(x) = 0$. We find the optimal value of $\lambda_C(x)$ for the third scenario. Similarly, we can prove that the optimal value of $\lambda_R(x)$ is equal to R_R for $x \in [0, x_{max}]$ in the first and second scenarios, and the optimal value of $\lambda_C(x)$ is equal to R_C for

$x \in [0, x_{max}]$ in the first scenario. To avoid repetition, these proofs are not provided.

In the third scenario, since $\lambda_R(x) = 0$, the total cost function that the social planner minimizes is as follows:

$$\min_{R_C \leq \lambda_C(x) \leq N} \int_0^{x_{max}} \lambda_C(x)[(\alpha - \beta)\tau_C(x) + (\alpha' - \beta)t_C + \beta(T - x)]dx. \tag{C.3}$$

By substituting $\tau_C(x)$ with $\frac{\int_0^x [\lambda_C(u) - R_C]du}{R_C}$ in (C.3), the objective function becomes minimizing $\int_0^{x_{max}} \lambda_C(x)[(\alpha - \beta)\frac{\int_0^x \lambda_C(u)du}{R_C} - \alpha x + \beta T + (\alpha' - \beta)t_C]dx$. This is a control theory problem, where $\lambda_C(x)$ is the control variable (i.e., the variable that we can change to control the value of the objective function) and $\int_0^x \lambda_C(u)du$ is the state variable (i.e., the variable that changes as we change the control variable). We use the maximum principle approach to find the optimal value of the control variable $\lambda_C(x)$.[2] Let $f(x)$ represent the derivative of the state variable, i.e., $f(x) = \frac{\partial[\int_0^x \lambda_C(u)du]}{\partial x} = \lambda_C(x)$, and $g(x)$ denotes the negative of the objective function, i.e., $g(x) = -\lambda_C(x)[(\alpha - \beta)\frac{\int_0^x \lambda_C(u)du}{R_C} - \alpha x + \beta T + (\alpha' - \beta)t_C]$. The Hamiltonian equation for this problem is defined as follows: $H = \rho(x)f(x) + g(x) = \rho(x)\lambda_H(x) - \lambda_C(x)[(\alpha - \beta)\frac{\int_0^x \lambda_C(u)du}{R_C} - \alpha x + \beta T + (\alpha' - \beta)t_C]$, where $\rho(x)$ is the marginal return vector. The variable $\rho(x)$ also satisfies the following two equations: $\frac{\partial \rho(x)}{\partial x} = -\frac{\partial H}{\partial[\int_0^x \lambda_C(u)du]} = (\alpha - \beta)\frac{\lambda_C(x)}{R_C}$, and $\rho(x_{max}) = \frac{\partial g}{\partial \lambda_C(x)}(x_{max}) = (\alpha - \beta)\frac{N}{R_C} - \alpha x_{max} + \beta T + (\alpha' - \beta)t_C$. From these two equations, we get $\rho(x) = (\alpha - \beta)\frac{\int_0^x \lambda_C(u)du}{R_C} - \alpha x_{max} + \beta T + (\alpha' - \beta)t_C$ for $x \in [0, x_{max}]$. We find the optimal value of $\lambda_C(x)$ that maximizes H. Since H is a linear function of $\lambda_C(x)$, if $\rho(x) - (\alpha - \beta)\frac{\int_0^x \lambda_C(u)du}{R_C} + \alpha x - \beta T + (\alpha' - \beta)t_C$ is positive, then $\lambda_C(x) = N$ maximizes H, otherwise $\lambda_C(x) = R_C$ is the optimal solution. Substituting $\rho(x)$ with $(\alpha - \beta)\frac{\int_0^x \lambda_C(u)du}{R_C} - \alpha x_{max} + \beta T + (\alpha' - \beta)t_C$, we have $\rho(x) - (\alpha - \beta)\frac{\int_0^x \lambda_C(u)du}{R_C} + \alpha x - \beta T + (\alpha' - \beta)t_C = -\alpha(x_{max} - x)$. Since $x \leq x_{max}$, $-\alpha(x_{max} - x)$ is negative, and the optimal value of $\lambda_C(x)$ for $x \in [0, x_{max}]$ is R_C. Since $\lambda_C(x) = R_C$, we characterize x_{max} as N/R_C, and T as $x_{max} + t_C$. Similarly, x_{max} (resp., T) is equal to $\frac{N}{R_C + R_R}$(resp., $x_{max} + t_C$) and N/R_R (resp., $x_{max} + t_R$) for the first scenario and the second scenario, respectively.

Next, considering the game between the social planner and the TNC, we find the value of $u(x)$ for the first two scenarios. We then show that the third scenario does not happen as it results in a higher total system cost and a lower profit for the TNC than the first scenario. In the first scenario,

[2]For further information on the maximum principle approach see Sethi & Thompson (2000).

since $\tau_C(x) = \tau_R(x) = 0$, the condition $(\alpha' - \beta)t_C + (\alpha - \beta)\tau_C(x) = (\alpha' - \beta)t_R + u(x) + (\alpha - \beta)\tau_R(x)$ can be simplified as follows: $(\alpha' - \beta)t_C = (\alpha' - \beta)t_R + u(x)$. After rearranging this equation, we have $u(x) = (\alpha' - \beta)(t_C - t_R)$. In this scenario, the total system cost is equal to $N[\alpha' t_C + \frac{\beta N}{2(R_C+R_R)}]$ and the TNC's profit is equal to $\int_0^{x_{max}} \lambda_R(x)u(x)dx = \frac{NR_R}{R_R+R_C}(\alpha' - \beta)(t_C - t_R)$. In the second scenario, since $\tau_C(x) = \tau_R(x) = 0$, the condition $(\alpha' - \beta)t_C + (\alpha - \beta)\tau_C(x) > (\alpha' - \beta)t_R + u(x) + (\alpha - \beta)\tau_R(x)$ can be simplified as follows: $(\alpha' - \beta)(t_C - t_R) > u(x)$. Since for the social planner's perspective the cost associated with driving remains constant, the cost associated with ride-hailing must stay constant too, i.e., $u(x) = u$, where $u > 0$ is a constant value. To find the value of $u(x)$, we find the total system cost and the TNC's profit. In this scenario, the total system cost is equal to $N[\alpha' t_R + u + \frac{\beta N}{2R_R}]$ and the TNC's profit is equal to $\int_0^{x_{max}} \lambda_R(x)u(x)dx = Nu$. On the one hand, since the total system cost is a linearly increasing function of u, the social planner wants to set u as low as possible. On the other hand, the TNC wants to maximize its profit, which is also a linearly increasing function of u. Comparing the TNC's profit in this scenario with the previous scenario, we show that at $u = \frac{R_R(\alpha' - \beta)(t_C - t_R)}{R_R + R_C}$ the TNC's profit is equal under both scenarios. Thus, the value of $u(x)$ in scenario two is equal to $\frac{R_R(\alpha' - \beta)(t_C - t_R)}{R_R + R_C}$. Finally, under the third scenario $\lambda_C(x) = R_C$ and $\lambda_R(x) = 0$. This means that the total system cost is equal to $N[\alpha' t_C + \frac{\beta N}{2R_C}]$, which is more than that under the first scenario, and the TNC's profit is equal to zero, which is less than that under the other two scenarios.

Lastly, depending on the total system cost, the social planner dictates either scenario two or scenario three. In particular, if the total system cost in the first scenario lower than that in the second scenario, i.e., $N[\alpha' t_C + \frac{\beta N}{2(R_C+R_R)}] < N[\alpha' t_R + u + \frac{\beta N}{2R_R}]$ which can be simplified as $\frac{\beta N R_C}{2(\alpha' R_C + \beta R_R)R_R} - (t_C - t_R) \geq 0$, we observe the first scenario (SO1 in Proposition 3.2). Otherwise, we observe the second scenario (SO2 in Proposition 3.2). □

Proof of Proposition 3.3. (a) The SO1 solution becomes an equilibrium solution, if Conditions 1 and 2 are satisfied.[3] In other words, for SO1 to be an equilibrium, all commuters must incur the same travel cost, regardless of their departure time from H or their mode of transportation. Our approach is to equalize all commuters' costs, by charging the conventional commuters a parking fee, $p(x)$, and the ride-hailing commuters a congestion toll, $\pi(x)$. Next we derive these $p(x)$ and

[3]Note that Condition 3 is already satisfied.

$\pi(x)$ functions.

Under SO1, the total travel cost for the conventional commuters is $C_C(x) = \beta(T-x) + (\alpha' - \beta)t_C + p(x)$ for $0 \leq x \leq x_{max}$ and that for the ride-hailing commuters is $C_R(x) = \beta(T-x) + (\alpha' - \beta)t_R + u(x) + \pi(x) = \beta(T-x) + (\alpha' - \beta)t_R + (\alpha' - \beta)(t_C - t_R) + \pi(x) = \beta(T-x) + (\alpha' - \beta)t_C + \pi(x)$ for $0 \leq x \leq x_{max}$. Note that since there is no passengered congestion time, the congestion in I $\tau_I(x)$, the downtown congestion $\tau_C(x)$, and the drop-off congestion $\tau_R(x)$ are not included in the cost. The commuters who incur the maximum cost are the ones who depart home at $x = 0$. This is because $C_C(x)$ and $C_R(x)$ are linearly decreasing functions of x. In other words, the maximum commuter cost is equal to $\beta T + (\alpha' - \beta)t_C$. As such, the parking fee $p(x)$ (resp., the congestion toll $\pi(x)$) is the difference between $\beta T + (\alpha' - \beta)t_C$ and $C_C(x)$ (resp., $C_R(x)$). So, we have the following: $p(x) = \pi(x) = \beta T + (\alpha' - \beta)t_C - \beta(T-x) - (\alpha' - \beta)t_C = \beta x$ for $x \in [0, x_{max}]$.

(b) The SO2 solution becomes an equilibrium solution, if all ride-hailing commuters incur the same travel cost, regardless of their departure time from H. In addition, no commuter should be able to reduce their cost by choosing to drive instead of using ride-hailing. We derive $p(x)$ and $\pi(x)$ such that these conditions are satisfied.

Under SO2, the total travel cost for the conventional commuters is $C_C(x) = \beta(T-x) + (\alpha' - \beta)t_C + p(x)$ for $0 \leq x \leq x_{max}$ and that for the ride-hailing commuters is $C_R(x) = \beta(T-x) + (\alpha' - \beta)t_R + u(x) + \pi(x) = \beta(T-x) + (\alpha' - \beta)t_R + \frac{R_R}{R_R+R_C}(\alpha' - \beta)(t_C - t_R) + \pi(x)$ for $0 \leq x \leq x_{max}$. The commuters who incur the maximum cost are the ones who depart home at $x = 0$. In other words, the maximum commuter cost is equal to $+(\alpha' - \beta)t_R + \frac{R_R}{R_R+R_C}(\alpha' - \beta)(t_C - t_R) + \pi(x)$. As such, the congestion toll $\pi(x)$ is the difference between this maximum cost and $C_R(x)$. So, we have the following:$\pi(x) = \beta T + (\alpha' - \beta)t_R + \frac{R_R}{R_R+R_C}(\alpha' - \beta)(t_C - t_R) + \pi(x) - [\beta(T-x) + (\alpha' - \beta)t_R + \frac{R_R}{R_R+R_C}(\alpha' - \beta)(t_C - t_R) + \pi(x)] = \beta x$ for $x \in [0, x_{max}]$. Lastly, $p(x)$ should be so high that no commuter chooses to drive. In other words, $p(x)$ should be so high that $C_C(x) > C_R(x)$. By substituting the values of $C_C(x)$ $C_R(x)$ in this inequality we have $C_C(x) = \beta(T-x) + (\alpha' - \beta)t_C + p(x) > C_R(x) = \beta(T-x) + (\alpha' - \beta)t_R + \frac{R_R}{R_R+R_C}(\alpha' - \beta)(t_C - t_R) + \pi(x)$. After rearranging this equation, we get the following equation: $p(x) > -\frac{R_C}{R_R+R_C}(\alpha' - \beta)(t_C - t_R)$ for $x \in [0, x_{max}]$. □

Proof of Corollary 3.2. We find the value of R_R^* for SO2 first and then discuss SO1. The total system cost under SO2 is equal to $C_{SO1} = \int_0^{x_{max}} R_R[(\alpha' - \beta)t_R + \frac{R_R}{R_R+R_C}(\alpha' - \beta)(t_C - t_R) + \beta(T -$

$x)]dx = N\frac{(\alpha' R_C+\beta R_R)t_R+(\alpha'-\beta)t_C R_R}{R_R+R_C} + \frac{\beta N^2}{2R_R}$. To find the value of R_R that minimizes total system cost, we set the derivative of C_{SO1} with respect to R_R equal to zero, i.e., $\int_0^{x_{max}} R_R[(\alpha'-\beta)t_R + \frac{R_R}{R_R+R_C}(\alpha'-\beta)(t_C-t_R)+\beta(T-x)]dx = \frac{N(\alpha'-\beta)(t_C-t_R)R_C}{(R_R+R_C)^2} - \frac{\beta N^2}{2R_R^2} = 0$. If $\beta N - 2(\alpha'-\beta)(t_C-t_R)R_C > 0$ (which is the condition for part (b) of Corollary 3.2), then the derivative is negative, and the total system cost is decreasing in R_R. In this scenario, the optimal value of R_R is equal to its maximum value, which is equal to the inbound highway capacity R_I. Note that R_R cannot exceed R_I because otherwise the drop-off process is no longer a bottleneck. If $\beta N - 2(\alpha'-\beta)(t_C-t_R)R_C \leq 0$ (which is the condition for part (a) of Corollary 3.2), then we find the value of R_R that satisfies $\frac{\partial C_{SO1}}{\partial R_R} = 0$. In particular, we find R_R that satisfies $R_R^2[(\alpha'-\beta)(t_C-t_R)R_C - \beta N/2] - \beta N R_C R_R - \beta N R_C^2/2 = 0$. This quadratic equation has only one positive solution which is equal to $\frac{R_C(\beta N+\sqrt{2\beta N R_C(\alpha'-\beta)(t_C-t_R)})}{2R_C(\alpha'-\beta)(t_C-t_R)-\beta N}$, denoted as R_R^1. To guarantee that this value of R_R is in fact a minimum for the total system cost, we show that the second derivative of C_{SO1} is positive at this value of R_R. In particular, $\frac{\partial^2 C_{SO1}}{\partial R_R^2} = -\frac{N(\alpha'-\beta)(t_C-t_R)R_C}{(R_R+R_C)^3} + \frac{\beta N^2}{2R_R^3} = -\frac{N(\alpha'-\beta)(t_C-t_R)R_C}{(R_R+R_C)^3} + \frac{\beta N^2}{2R_R^3} = -\frac{\beta N}{R_R^2(R_R+R_C)} + \frac{\beta N^2}{2R_R^3} = \frac{\beta N R_C}{R_R^3(R_R+R_C)} > 0$.

Next, the total system cost under SO1, which is equal to $\int_0^{x_{max}}(R_R+R_C)[(\alpha'-\beta)t_C + \beta(T-x)]dx = N\alpha' t_C + \frac{\beta N^2}{2(R_R+R_C)}$, is strictly decreasing in R_R. Hence, as R_R increases the total system cost decreases. However, the condition for observing SO1, i.e, $\frac{\beta N R_C}{2(\alpha' R_C+\beta R_R)R_R} - (t_C - t_R) \geq 0$, is satisfied for $R_R \leq \frac{\alpha' R_C+\sqrt{(\alpha' R_C)^2+2\beta^2 N R_C/(t_C-t_R)}}{2\beta}$. So if we increase R_R beyond this value, denoted as R_R^2, then we observe SO2, for which we showed the optimal value of R_R. Thus, in general the optimal value of R_R is the maximum of R_R^1 and R_R^2, as long as this maximum value is lower than R_I, i.e., $R_R^* = \min\{\max\{\frac{R_C(\beta N+\sqrt{2\beta N R_C(\alpha'-\beta)(t_C-t_R)})}{2R_C(\alpha'-\beta)(t_C-t_R)-\beta N}, \frac{\alpha' R_C+\sqrt{(\alpha' R_C)^2+2\beta^2 N R_C/(t_C-t_R)}}{2\beta}\}, R_I\}$. □

Bibliography

Agarwal, S., Mani, D. & Telang, R. (2019), 'The impact of ride-hailing services on congestion: Evidence from indian cities', *Indian School of Business* .

Aguilar, J. (2018), 'A 10-lane highway and Colorado's first autonomous vehicle lane could be prescription for west-suburban Denver traffic jams', Accessed September 25, 2018, https://www.denverpost.com/2018/01/21/colorado-10-lane-highway-autonomous-vehicle-lane-traffic/.

Albright, J., Bell, A., Schneider, J. & Nyce, C. (2015), 'Market place of change: automobile insurance in the era of autonomous vehicles', Accessed September 25, 2018, https://assets.kpmg.com/content/dam/kpmg/pdf/2016/06/id-market-place-of-change-automobile-insurance-in-the-era-of-autonomous-vehicles.pdf.

Alfa, A. S. & Neuts, M. F. (1995), 'Modelling vehicular traffic using the discrete time markovian arrival process', *Transportation Science* **29**(2), 109–117.

Amoozadeh, M., Raghuramu, A., Chuah, C.-N., Ghosal, D., Zhang, H. M., Rowe, J. & Levitt, K. (2015), 'Security vulnerabilities of connected vehicle streams and their impact on cooperative driving', *IEEE Communications Magazine* **53**(6), 126–132.

Arnott, R., De Palma, A. & Lindsey, R. (1991), 'A temporal and spatial equilibrium analysis of commuter parking', *Journal of public economics* **45**(3), 301–335.

Bai, J., So, K. C., Tang, C. S., Chen, X. & Wang, H. (2019), 'Coordinating supply and demand on an on-demand service platform with impatient customers', *Manufacturing & Service Operations Management* **21**(3), 556–570.

Bando, M., Hasebe, K., Nakayama, A., Shibata, A. & Sugiyama, Y. (1995), 'Dynamical model of traffic congestion and numerical simulation', *Physical Review E* **51**(2), 1035.

Baron, O., Berman, O. & Nourinejad, M. (2018), 'Introducing autonomous vehicles: Formulation and analysis of public policies.', **Working paper**.

Barth, B. (2019), 'Curb Control', Accessed March 31, 2021, https://www.planning.org/planning/2019/jun/ curbcontrol/.

Bayern, M. (2020), 'Autonomous vehicles: How 7 countries are handling the regulatory landscape', Accessed December 07, 2020, https://www.techrepublic.com/article/autonomous-vehicles-how-7-countries-are-handling-the-regulatory-landscape/.

Benjaafar, S., Dooley, K. & Setyawan, W. (1997), *Cellular automata for traffic flow modeling*, Center for Transportation Studies, University of Minnesota.

Benjaafar, S. & 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., Kong, G., Li, X. & Courcoubetis, C. (2018), 'Peer-to-peer product sharing: Implications for ownership, usage, and social welfare in the sharing economy', *Management Science.* **Forthcoming**.

Bergenhem, C., Shladover, S., Coelingh, E., Englund, C. & Tsugawa, S. (2012), Overview of platooning systems, *in* 'Proceedings of the 19th ITS World Congress, Vienna, Austria'.

Besbes, O., Castro, F. & Lobel, I. (2021), 'Surge pricing and its spatial supply response', *Management Science* **67**(3), 1350–1367.

Bierstedt, J., Gooze, A., Gray, C., Peterman, J., Raykin, L. & Walters, J. (2014), 'Effects of next-generation vehicles on travel demand and highway capacity', *FP Think Working Group* pp. 10–11.

Breuer, L. & Alfa, A. S. (2005), 'An em algorithm for platoon arrival processes in discrete time', *Operations Research Letters* **33**(5), 535–543.

Castiglione, J., Chang, T., Cooper, D., Hobson, J., Logan, W., Young, E., Charlton, B., Wilson, C., Mislove, A., Chen, L. et al. (2016), 'Tncs today: a profile of san francisco transportation network company activity', *San Francisco County Transportation Authority (June 2016)* .

Castiglione, J., Cooper, D., Sana, B., Tischler, D., Chang, T., Erhardt, G. D., Roy, S., Chen, M. & Mucci, A. (2018), 'Tncs & congestion'.

Cheah, J. Y. & Smith, J. M. (1994), 'Generalized m/g/c/c state dependent queueing models and pedestrian traffic flows', *Queueing Systems* **15**(1), 365–386.

Chen, D., Ahn, S., Chitturi, M. & Noyce, D. A. (2017), 'Towards vehicle automation: Roadway capacity formulation for traffic mixed with regular and automated vehicles', *Transportation Research Part B: methodological* **100**, 196–221.

Chen, Y., Korpeoglu, C., Korpeoglu, E., Sahin, O., Tang, C., Xiao, S. et al. (2018), 'Innovative online platforms: Research opportunities. working paper. johns hopkins carey business school, baltimore, md, january'.

Daganzo, C. F. (1994), 'The cell transmission model: A dynamic representation of highway traffic consistent with the hydrodynamic theory', *Transportation Research Part B: Methodological* **28**(4), 269–287.

Daganzo, C. F. (2007), 'Urban gridlock: Macroscopic modeling and mitigation approaches', *Transportation Research Part B: Methodological* **41**(1), 49–62.

Daw, A., Hampshire, R. C. & Pender, J. (2019), 'Beyond safety drivers: Staffing a teleoperations system for autonomous vehicles.', **Working paper**.

Del Castillo, J. & Benitez, F. (1995), 'On the functional form of the speed-density relationship-I: General theory', *Transportation Research Part B: Methodological* **29**(5), 373–389.

Dunne, M. C. (1967), 'Traffic delay at a signalized intersection with binomial arrivals', *Transportation Science* **1**(1), 24–31.

Eliot, L. (2019), 'An Inconvenient Truth: Human Drivers and Autonomous Cars Mix Like Oil

And Water', Accessed June 07, 2019, https://www.forbes.com/sites/lanceeliot/2019/05/07/an-inconvenient-truth-human-drivers-and-autonomous-cars-mix-like-oil-and-water/1bd295103b84.

Faggella, D. (2020), 'The Self-Driving Car Timeline – Predictions from the Top 11 Global Automakers', Accessed May 25, 2021, https://emerj.com/ai-adoption-timelines/self-driving-car-timeline-themselves-top-11-automakers/.

Federal Highway Administration (2011), 'Our Nation's Highways', Accessed September 25, 2018, https://www.fhwa.dot.gov/policyinformation/pubs/hf/pl11028/chapter2.cfm.

Federal Highway Administration (2020), '21st century operations using 21st century technologies.', Accessed January 17, 2021, https://ops.fhwa.dot.gov/congestionpricing/cp_what_is.htm.

Garg, N. & Nazerzadeh, H. (2021), 'Driver surge pricing', *Management Science* .

Geroliminis, N., Daganzo, C. F. et al. (2007), Macroscopic modeling of traffic in cities, *in* 'Transportation Research Board 86th Annual Meeting', number 07-0413, No. 07-0413.

Ghiasi, A., Hussain, O., Qian, Z. & Li, X. (2017), 'A mixed traffic capacity analysis and lane management model for connected automated vehicles: A markov chain method', *Transportation Research Part B: Methodological* **106**, 266 – 292.

Goodchild, A., MacKenzie, D., Ranjbari, A., Machado, J. & Dalla Chiara, G. (2019), 'Curb Allocation Change Project', Accessed March 21, 2022, https://depts.washington.edu/sctlctr/sites/default/files/research$_pub_files/Curb_Allocation_Change_Project-UWUrbanFreightLab.pdf$.

Guda, H. & Subramanian, U. (2019), 'Your uber is arriving: Managing on-demand workers through surge pricing, forecast communication, and worker incentives', *Management Science* **65**(5), 1995–2014.

Guzzella, L. & Kiencke, U. (1995), *Advances in Automotive Control*, Elsevier, Ascona, Switzerland.

Harchol-Balter, M. (2013), *Performance modeling and design of computer systems: queueing theory in action*, Cambridge University Press, New York, USA.

Hasija, S., Shen, Z.-J. M. & Teo, C.-P. (2020), 'Smart city operations: Modeling challenges and opportunities', *Manufacturing & Service Operations Management* **22**(1), 203–213.

He, F., Yin, Y., Chen, Z. & Zhou, J. (2015), 'Pricing of parking games with atomic players', *Transportation Research Part B: Methodological* **73**, 1–12.

He, L., Hu, Z. & Zhang, M. (2018), 'Robust repositioning for vehicle sharing', *Manufacturing & Service Operations Management.* **Forthcoming**.

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

He, L., Mak, H.-Y., Rong, Y. & Shen, Z.-J. M. (2017), 'Service region design for urban electric vehicle sharing systems', *Manufacturing & Service Operations Management* **19**(2), 309–327.

He, Q.-M. (2014), *Fundamentals of matrix-analytic methods*, Vol. 365, Springer, New York.

Heidemann, D. (1996), A queueing theory approach to speed-flow-density relationships, *in* 'Internaional Symposium on Transportation and Traffic Theory', pp. 103–118.

Hendricks, D. (2015), '5 reasons you shouldn't get to work early', Accessed January 04, 2021, https://www.businessinsider.com/reasons-you-shouldnt-get-to-work-early-2015-2.

Holtzman, J. M. & Goodman, D. J. (2012), *Wireless Communications: Future Directions*, Vol. 217, Springer Science & Business Media, New York.

Hu, B., Hu, M. & Zhu, H. (2022), 'Surge pricing and two-sided temporal responses in ride hailing', *Manufacturing & Service Operations Management* **24**(1), 91–109.

Ingraham, C. (2019), 'Nine days on the road. Average commute time reached a new record last year.', Accessed December 07, 2020, https://www.washingtonpost.com/business/2019/10/07/nine-days-road-average-commute-time-reached-new-record-last-year/.

INRIX (2017), 'INRIX Global Traffic Scorecard: Interactive Ranking and City Dashboards', Accessed September 25, 2018, http://inrix.com/scorecard/.

Jaffe, E. (2015), 'How Parking Conquered L.A.', Accessed December 07, 2020, https://www.bloomberg.com/news/articles/2015-12-03/how-parking-conquered-los-angeles-in-14-facts-maps-and-figures.

Jain, R. & Smith, J. M. (1997), 'Modeling vehicular traffic flow using m/g/c/c state dependent queueing models', *Transportation Science* **31**(4), 324–336.

Kim, S.-H. & Whitt, W. (2013), 'Estimating waiting times with the time-varying little's law', *Probability in the Engineering and Informational Sciences* **27**(4), 471.

Kuwahara, M. & Newell, G. F. (1987), Queue evolution on freeways leading to a single core city during the morning peak, *in* 'Proceedings of the 10th International Symposium on Transportation and Traffic Theory', pp. 21–40.

Law, A. M., Kelton, W. D. & Kelton, W. D. (2000), *Simulation Modeling and Analysis*, McGraw-Hill, New York.

Lehoczky, J. (1972), 'Traffic intersection control and zero-switch queues under conditions of markov chain dependence input', *Journal of Applied Probability* **9**(2), 382–395.

Li, Z., Hong, Y. & Zhang, Z. (2016), 'Do ride-sharing services affect traffic congestion? an empirical study of uber entry', *Social Science Research Network* **2002**, 1–29.

Lian, Z. & Van Ryzin, G. (2021), 'Optimal growth in two-sided markets', *Management Science* **67**(11), 6862–6879.

Lim, M. K., Mak, H.-Y. & Rong, Y. (2014), 'Toward mass adoption of electric vehicles: impact of the range and resale anxieties', *Manufacturing & Service Operations Management* **17**(1), 101–119.

Liu, H., Xiao, L., Kan, X., Shladover, S., Lu, X., Men, M., Shakel, W. & vanArem, B. (2018), 'Using cooperative adaptive cruise control (cacc) to form high-performance vehicle streams – final report.', **Working paper, University of California, Berkeley**.

Liu, W. (2018), 'An equilibrium analysis of commuter parking in the era of autonomous vehicles', *Transportation Research Part C: Emerging Technologies* **92**, 191–207.

Lucantoni, D. M. (1991), 'New results on the single server queue with a batch markovian arrival process', *Communications in Statistics. Stochastic Models* **7**(1), 1–46.

Mak, H.-Y. (2020), 'Enabling smarter cities with operations management', *Manufacturing & Service Operations Management.* **Forthcoming**.

Mak, H.-Y., Rong, Y. & Shen, Z.-J. M. (2013), 'Infrastructure planning for electric vehicles with battery swapping', *Management Science* **59**(7), 1557–1575.

Meiszner, P. (2019), 'Vancouver proposes drop-off, pick-up fee for Uber and Lyft.', Accessed March 8, 2022, https://www.urbanyvr.com/vancouver-ridesharing-fee/.

Mirzaeian, N., Cho, S.-H. & Qian, S. (2021), 'Can autonomous vehicles solve the commuter parking problem?', *Available at SSRN 3872106* .

Mohajerpoor, R. & Ramezani, M. (2019), 'Mixed flow of autonomous and human-driven vehicles: Analytical headway modeling and optimal lane management', *Transportation Research Part C: Emerging Technologies* **109**, 194–210.

Muoio, D. (2017), 'The 18 companies most likely to get self-driving cars on the road first', Accessed September 25, 2018, http://www.businessinsider.com/the-companies-most-likely-to-get-driverless-cars-on-the-road-first-2017-4/.

National League of Cities (2018), 'Autonomous vehicle pilots across america', Accessed October 26, 2018, https://www.nlc.org/resource/autonomous-vehicle-pilots-across-america.

Neuts, M. F. (1979), 'A versatile markovian point process', *Journal of Applied Probability* **16**(04), 764–779.

Neuts, M. F. & Chakravarthy, S. (1981), 'A single server queue with platooned arrivals and phase type services', *European Journal of Operational Research* **8**(4), 379–389.

NHTSA (2015), 'Why your reaction time matters at speed', Accessed September 25, 2018, https://one.nhtsa.gov/nhtsa/Safety1nNum3ers/.

Nourinejad, M. & Amirgholy, M. (2018), 'Parking pricing and design in the morning commute problem with regular and autonomous vehicles', *York University,* **Working paper**.

Overtoom, I., Correia, G., Huang, Y. & Verbraeck, A. (2020), 'Assessing the impacts of shared autonomous vehicles on congestion and curb use: A traffic simulation study in the hague, netherlands', *International journal of transportation science and technology* **9**(3), 195–206.

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

Parkopedia (2020), ' Pittsburgh Parking Rates', Accessed September 19, 2020, https://en.parkopedia.com/parking/pittsburgh/.

Pelletier, S., Jabali, O. & Laporte, G. (2016), '50th anniversary invited article goods distribution with electric vehicles: review and research perspectives', *Transportation Science* **50**(1), 3–22.

Pi, X., Ma, W. & Qian, Z. S. (2019), 'A general formulation for multi-modal dynamic traffic assignment considering multi-class vehicles, public transit and parking', *Transportation Research Part C: Emerging Technologies* **104**, 369–389.

Pittsburgh Parking Authority (2019), 'Map of Downtown Parking Facilities Facility Rate Chart', Accessed March 22, 2022, https://apps.pittsburghpa.gov/redtail/images/7718$_{M}ap - And - Rates - November - 1 - 2019.pdf$.

Qi, W., Li, L., Liu, S. & Shen, Z.-J. M. (2018), 'Shared mobility for last-mile delivery: Design, operational prescriptions, and environmental impact', *Manufacturing & Service Operations Management.* **Forthcoming**.

Qi, W., Sha, M. et al. (2020), 'When shared autonomous electric vehicles meet microgrids: Citywide energy-mobility orchestration.', **Working paper**.

Qian, Z. & Rajagopal, R. (2015), 'Optimal dynamic pricing for morning commute parking', *Transportmetrica A: Transport Science* **11**(4), 291–316.

Qian, Z. S. & Rajagopal, R. (2014), 'Optimal dynamic parking pricing for morning commute considering expected cruising time', *Transportation Research Part C: Emerging Technologies* **48**, 468–490.

Qian, Z. S., Xiao, F. E. & Zhang, H. (2011), 'The economics of parking provision for the morning commute', *Procedia-Social and Behavioral Sciences* **17**, 612–633.

Qian, Z. S., Xiao, F. E. & Zhang, H. (2012), 'Managing morning commute traffic with parking', *Transportation research part B: methodological* **46**(7), 894–916.

Qom, S. F., Xiao, Y. & Hadi, M. (2016), Evaluation of cooperative adaptive cruise control (cacc) vehicles on managed lanes utilizing macroscopic and mesoscopic simulation, *in* 'Transportation Research Board 95th Annual Meeting', number 16-6384.

Ross, S. M. (2006), *Simulation*, Elsevier Academic Press, San Diego.

Sethi, S. P. & Thompson, G. L. (2000), *Optimal Control Theory Applications to Management Science and Economics*, Springer, New York, NY, USA.

Shaver, K. (2019), 'City planners eye self-driving vehicles to correct mistakes of the 20th-century auto', Accessed May 25, 2021, https://www.washingtonpost.com/transportation/2019/07/20/city-planners-eye-self-driving-vehicles-correct-mistakes-th-century-auto/.

Shladover, S., Su, D. & Lu, X.-Y. (2012), 'Impacts of cooperative adaptive cruise control on freeway traffic flow', *Transportation Research Record: Journal of the Transportation Research Board* (2324), 63–70.

Siciliano, B. & Khatib, O. (2016), *Springer Handbook of Robotics*, Springer, Heidelberg, Germany.

Small, K. A. (1982), 'The scheduling of consumer activities: work trips', *The American Economic Review* **72**(3), 467–479.

Stern, R. E., Cui, S., Delle Monache, M. L., Bhadani, R., Bunting, M., Churchill, M., Hamilton, N., Pohlmann, H., Wu, F., Piccoli, B. et al. (2018), 'Dissipation of stop-and-go waves via control of autonomous vehicles: Field experiments', *Transportation Research Part C: Emerging Technologies* **89**, 205–221.

Strong Towns (2018), 'The Many Costs of Too Much Parking', Accessed December 16, 2020, https://www.strongtowns.org/journal/2018/11/20/the-many-costs-of-too-much-parking: :text=Parking.

Su, Q. & Wang, D. Z. (2019), 'Morning commute problem with supply management considering parking and ride-sourcing', *Transportation Research Part C: Emerging Technologies* **105**, 626–647.

Talebpour, A., Mahmassani, H. S. & Elfar, A. (2017), 'Investigating the effects of reserved lanes for autonomous vehicles on congestion and travel time reliability', *Transportation Research Record: Journal of the Transportation Research Board* (2622), 1–12.

The automated driving community (2018), 'Milestones - The automated driving timeline', Accessed September 25, 2018, https://www.2025ad.com/latest/milestones-the-ad-timeline/.

Tientrakool, P., Ho, Y.-C. & Maxemchuk, N. F. (2011), 'Highway capacity benefits from using vehicle-to-vehicle communication and sensors for collision avoidance', *IEEE Vehicular Technology Conference (VTC Fall)* pp. 1–5.

Tiwari, H. & Marsani, A. (2014), 'Calibration of conventional macroscopic traffic flow models for nepalese roads', *Institute of Engineering, Graduate Conference* .

Transportation Research Board (2000), *Highway Capacity Manual*, Transportation Research Board, Washington, D.C., USA.

Treiber, M., Hennecke, A. & Helbing, D. (2000), 'Congested traffic states in empirical observations and microscopic simulations', *Physical Review E* **62**(2), 1805.

U.S. Bureau of Labor Statistics (2015), 'Careers for night owls and early birds.', Accessed January 31, 2021, https://www.bls.gov/careeroutlook/2015/article/night-owls-and-early-birds.htm.

U.S. Census Bureau (2019), '2014-2018 American Community Survey 5-Year Estimates', Accessed December 07, 2020, https://www.census.gov/programs-surveys/acs/.

Van Woensel, T. & Vandaele, N. (2006), 'Empirical validation of a queueing approach to uninterrupted traffic flows', *4OR* **4**(1), 59–72.

Van Woensel, T. & Vandaele, N. (2007), 'Modeling traffic flows with queueing models: a review', *Asia-Pacific Journal of Operational Research* **24**(04), 435–461.

Vandaele, N., Van Woensel, T. & Verbruggen, A. (2000), 'A queueing based traffic flow model', *Transportation Research Part D: Transport and Environment* **5**(2), 121–135.

Vander Werf, J., Shladover, S., Miller, M. & Kourjanskaia, N. (2002), 'Effects of adaptive cruise control systems on highway traffic flow capacity', *Transportation Research Record: Journal of the Transportation Research Board* (1800), 78–84.

Varaiya, P. (2005), 'What we've learned about highway congestion', *Access* **1**(27), 2–7.

Vickrey, W. S. (1969), 'Congestion theory and transport investment', *The American Economic Review* **59**(2), 251–260.

Virginia DMV (2016), 'The virginia driver's manual', Accessed September 25, 2018, https://www.dmv.virginia.gov/webdoc/pdf/dmv39.pdf.

Wang, H., Rudy, K., Li, J. & Ni, D. (2010), 'Calculation of traffic flow breakdown probability to optimize link throughput', *Applied Mathematical Modelling* **34**(11), 3376–3389.

Wei, K., Vaze, V. & Jacquillat, A. (2021), 'Transit planning optimization under ride-hailing competition and traffic congestion', *Transportation Science* .

Xu, D., Guo, X. & Zhang, G. (2019), 'Constrained optimization for bottleneck coarse tolling', *Transportation Research Part B: Methodological* **128**, 1–22.

Yu, J. J., Tang, C. S., Max Shen, Z.-J. & Chen, X. M. (2020), 'A balancing act of regulating on-demand ride services', *Management Science* **66**(7), 2975–2992.

Zhang, X., Huang, H.-J. & Zhang, H. (2008), 'Integrated daily commuting patterns and optimal road tolls and parking fees in a linear city', *Transportation Research Part B: Methodological* **42**(1), 38–56.

Zhang, X., Liu, W., Waller, S. T. & Yin, Y. (2019), 'Modelling and managing the integrated morning-evening commuting and parking patterns under the fully autonomous vehicle environment', *Transportation Research Part B: Methodological* **128**, 380–407.

Zhao, L. & Sun, J. (2013), 'Simulation framework for vehicle platooning and car-following behaviors under connected-vehicle environment', *Procedia-Social and Behavioral Sciences* **96**, 914–924.