Recent Advances in Urban Mobility: Analyzing Efficiency, Equity and Safety

Akhil Shetty

Electrical Engineering and Computer Sciences
University of California, Berkeley

Technical Report No. UCB/EECS-2023-106
http://www2.eecs.berkeley.edu/Pubs/TechRpts/2023/EECS-2023-106.html

May 11, 2023

Recent Advances in Urban Mobility: Analyzing Efficiency, Equity and Safety

by

Akhil Shetty

A dissertation submitted in partial satisfaction of the

requirements for the degree of

Doctor of Philosophy

in

Engineering - Electrical Engineering and Computer Sciences

in the

Graduate Division

of the

University of California, Berkeley

Committee in charge:

Professor Kameshwar Poolla, Chair
Professor Munther Dahleh
Professor Roberto Horowitz
Professor Claire Tomlin

Spring 2023

Recent Advances in Urban Mobility: Analyzing Efficiency, Equity and Safety

Abstract

Recent Advances in Urban Mobility: Analyzing Efficiency, Equity and Safety

by

Akhil Shetty

Doctor of Philosophy in Engineering - Electrical Engineering and Computer Sciences

University of California, Berkeley

Professor Kameshwar Poolla, Chair

Transportation systems are the backbone of our modern economy. The ease of urban mobility is crucial for the success of cities. The automobile, which revolutionized urban mobility in the 20$^{\text{th}}$ century, is now struggling to keep up with demands for safety, efficiency, and equity. With more than 36,000 lives lost annually to traffic crashes in the US, and the average American spending 97 hours every year stuck in traffic, there is an imminent need for innovative solutions.

Two recent technologies show promise in delivering the next leap in urban mobility: (a) transportation network services, and (b) automated vehicles. Transportation network services (TNS) are businesses that connect people spread across the transportation network for efficient provision of goods and services. Examples include ride-hailing services such as Uber, and delivery services such as Instacart. Automated vehicles (AVs), popularly known as *self-driving cars*, obviate the need for a human to be in the driver's seat. The combination of AVs and TNS is expected to significantly reduce traffic fatalities, alleviate congestion, and improve accessibility.

Despite their transformative potential, the path to a bright AV-TNS future of urban mobility seems unclear. Companies have realized after years of effort that designing safe AVs is much more difficult than originally anticipated. Recent studies have found that TNS companies have increased congestion in major US cities. Moreover, a significant proportion of TNS workers are unable to earn the prevailing minimum wage in their jurisdictions. This thesis is an attempt to understand why the promised AV-TNS future has remained elusive.

We explore two broad themes along this path: (a) economics of transportation network services, and (b) automated vehicle safety. Our work on these themes constitute the two halves of this thesis. In the first half, we introduce a unified framework to analyze the economics of transportation network services. Using this framework, we analyze the impact of recent labor regulations on the ride-hailing ecosystem and suggest alternatives that can achieve

better outcomes. Adapting this framework to the case of delivery services, we demonstrate how the trade-off between delivery times and the cost of delivery, mediated by the extent of order pooling, dictates their economic viability. In the second half, we use a combination of theoretical models and empirical data to argue why connectivity is indispensable for safe AV deployment. Recognizing that some crashes are inevitable without connectivity, we propose a risk assessment framework that leverages human driving data along with on-road AV testing data to develop insights into their safety capabilities. We discuss several applications of this framework that inform safe AV deployment.

To *Mummy* and *Dodda*, whose sacrifices paved the path for me.

Contents

List of Figures

List of Tables

Acknowledgments

The last six years have absolutely flown by, and I haven't yet fully internalized the fact that I'll be graduating soon. A nice side-benefit of writing up this thesis is the opportunity to reflect on my time so far and truly appreciate the contributions of the people who helped me get to this point.

I feel blessed to have had the opportunity to learn from my advisors, Prof. Kameshwar Poolla and Prof. Pravin Varaiya. Kameshwar has been a wonderful academic advisor, but an even better friend and mentor. His unwavering faith in my abilities, along with his whole-hearted support and encouragement, gave me the confidence to pursue new ideas without worrying about the final outcome. There is never a dull moment when you're around Kameshwar. Time spent with him has always been entertaining and enlightening in equal measure, and I'm deeply grateful for his friendship.

Pravin was the intellectual guiding light for this thesis. Most of the problems I've tackled in this thesis have their roots in our weekly discussions over coffee at V&A. Pravin's superlative intellect and research pedigree are well known. However, his humility, kindness, and inexhaustible curiosity are what I remember him by. Although Pravin knew pretty much everything about everything, he had a way of making you believe that your ideas mattered, even if you were a fresh graduate student who didn't know much about the field. I feel incredibly fortunate for the privilege of knowing Pravin, and tapping into his wisdom.

I am grateful to Prof. Munther Dahleh, Prof. Claire Tomlin, and Prof. Roberto Horowitz for generously offering their time to be a part of my dissertation committee and providing valuable feedback that improved the quality of this thesis. A big thank you to the amazing EECS administrative staff and in particular, Shirley Salanio, for always promptly (and cheerfully!) helping with any administrative queries along the way. I owe a debt of gratitude to my undergraduate research advisors at IIT Bombay – Prof. Vivek Borkar, Prof. Sharayu Moharir, and Prof. D. Manjunath. They played a crucial role in getting me excited about research by providing me with ample guidance and a supportive environment to play around with new ideas. My decision to pursue a Ph.D. was significantly influenced by my interactions with them.

One of the most enjoyable parts of the Ph.D. experience has been the opportunity to work closely with a wonderful group of collaborators, whose smarts are only exceeded by their kindness and humility. I had the privilege of having Sen Li, Hamidreza Tavafoghi, and Junjie Qin as *defacto* mentors during my initial years at Berkeley. I am grateful to them for being incredibly generous with their time and valuable feedback as I was finding my feet in my academic journey. They played a key role in helping me identify interesting research problems in the domain of transportation network services, as well as bringing the projects to fruition. My understanding of automated vehicle safety and associated challenges was shaped by the weekly traffic safety meetings organized by Pravin. Discussions with Alex Kurzhanskiy, Jared Porter, Hamidreza Tavafoghi, Mengqiao Yu, Aditya Medury, Offer Grembek, Edward Kim, and Prof. Sanjit Seshia taught me a lot about this domain and

helped me isolate the problems that I ended up working on. It was a delight to collaborate with Alex, Jared, Hamid, and Mengqiao on many of these problems.

When I left home for graduate school six years ago, I had no idea that I'd be gaining a new family immediately in the form of the BCCI lab. Despite the dull grey walls and absence of any windows, being a part of this lab added a lot of color to my life, all thanks to the BCCI family – Emmanuel Sin, Galaxy Yin, Octavio Narvaez-Aroche, Jared Porter, Sen Li, Hamidreza Tavafoghi, Junjie Qin, Kate Schweidel, Arun Hegde, Pratyush Chakraborty, and Jonathan Mather. Over late nights at the lab scrambling to meet deadlines, dinners at Great China, karaoke nights at Nick's, golf at Tilden, and countless other memorable times, we forged a strong bond that I hope will last for decades to come.

My friends in Berkeley and beyond have been a crucial support system through the highs and lows of Ph.D. life. They have helped me feel at home, far away from home. In hindsight, choosing to stay at I-House turned out to be one of the best decisions I've made in my time in Berkeley. My time at I-House led me to meet an amazing group of people – Monimoy Bujarbaruah, Edward Kim, Łukasz Langer, Linnan Cao, Mick Jermsurawong, and Rama Subramanian, and my life has been infinitely enriched by their presence. I am indebted to Grace Hong for supporting me through the initial Ph.D. years with her calming presence and optimistic outlook. I am thankful to Jacopo Guanetti for the sumptuous meals he treated us to at "People's House", and Monimoy Bujarbaruah, Paolo Micalizzi, Yeojun Kim, Kyle Sefraoui, and Dimitris Papadimitriou for the hilarious conversations while Jacopo weaved his magic. It's been great to share the Ph.D. journey with Vijay Govindarajan, Utkarsha Agwan, Hari Prasanna Das, Kevin Cheang, Sangwoo Park, Conrad Holda, and Siddharth Nair. I am thankful to my undergrad buddies, Aditya Nambiar, Samarth Gupta, and Sagun Pai, and my high school friends, Akash Pandey, Tushar Anchan, and Purvang Thakkar, for staying in touch even though we were in different parts of the globe. Last but certainly not least, I am grateful to Ayushi Aggarwal for brightening up my life with her lively spirit, for being a wonderful sounding board, and for her constant support.

Finally, I would like to thank my family for being there for me *always*. Papa, through his work ethic, drive, and persevering attitude, has been a source of inspiration for me. I am grateful for my brother and my fun band of cousins, who have helped keep me grounded and not take life too seriously. I feel extremely fortunate to have two remarkable women in my life – Mummy and Dodda (my paternal grandmother). I am constantly amazed by Mummy's ability to handle the most stressful of situations with equanimity. She has always put my needs before her own, given me the confidence to pursue my interests without the fear of failure, and gone above and beyond to put me in a position to succeed. I had the unique privilege of being raised by my grandmother since the age of two. My parents being doctors who worked long hours, Dodda decided to uproot her life and social community to move to Mumbai (a city where she didn't even speak the local language), to take care of me. The last few years have been particularly tough on her with me being away for so long, but she has been supportive throughout. I can never hope to repay Mummy and Dodda for the sacrifices they have made, but I'd like to dedicate this thesis to them as a small token of my gratitude.

Chapter 1

Introduction

> *Transportation is the centre of the world. It is the glue of our daily lives. When it goes well, we don't see it. When it goes wrong, it negatively colours our day, makes us feel angry and curtails our possibilities – Robin Chase, CEO of Zipcar.*

Transportation systems are the beating heart of our modern world. At their best, they go largely unnoticed – a background hum in the cacophony of our daily lives. At their worst, they bring society to a standstill. Better transportation systems improve mobility of individuals. Better mobility, in turn, breeds greater connection and collaboration. Distances become shorter, our circle of influence becomes larger. The best transportation systems make our social species more social.

The success of cities is predicated on the ease of mobility within them. From the subways of New York City (NYC) to the gondolas of Venice, urban mobility systems connect people to achieve mutually beneficial outcomes. The speed and accessibility of these systems dictate the size and shape of cities. For instance, cities with good public transit connectivity like Chicago and NYC tend to have a dense urban core. Cities like Atlanta and Houston without such connectivity are characterized by relatively sparse cores and extended urban sprawl. These differences in the urban landscape translate into vastly different economic outcomes.

The Transportation Trifecta

The trifecta of safety, efficiency, and equity is at the core of a well-functioning transportation system. Safety is an essential requirement. In an ideal world, our mobility services would be risk-free. The unfortunate reality is that mishaps do occur due to human error or systemic faults. Examples include traffic crashes due to drunk drivers, and derailment of trains due to mechanical failure of tracks. The risk of such mishaps must be minimized to the extent possible. Efficiency is an equally important consideration. Our mobility services must facilitate the movement of people and goods while minimizing cost and travel time. Since these services are major contributors to pollution, emissions must be curtailed. Fast, low-cost, and clean mobility has a multiplicative effect on economic activity by reducing the

transaction cost of travel. Lastly, our transportation systems ought to work for everyone. This means ensuring accessibility and affordability for the most vulnerable among us. These requirements are often in conflict. Without technological advances, we are left with making hard choices.

From Horses to Automobiles

Urban mobility took a giant leap on all three fronts with the advent of the automobile in the 20$^{\text{th}}$ century. Until the late 1800s, cities were small and compact. This is because horse-drawn carriages were the fastest modes of urban transport [1]. The dominant mode of transport of the time brought its concomitant annoyances. Here is an excerpt from [2] describing the condition of streets in major cities:

> *By the 1890s around 300,000 horses were working on the streets of London, and more than 150,000 in New York City. Each of these horses produced an average of twenty-two pounds (ten kilograms) of manure a day, plus a quart (about a liter) of urine. Collecting and removing thousands of tons of waste from stables and streets proved increasingly difficult.*

Despite the adverse effects on public health, major cities in the US and Europe continued to be heavily reliant on horses well into the start of the 20$^{\text{th}}$ century. Clogged traffic due to incapacitated horses was a common occurrence [3]. Moreover, equine epidemics would bring transportation in cities to standstill, resulting in economic and social upheaval. The situation was evidently unsustainable.

The automobile emerged as a technological savior for cities, transforming the urban landscape. Cities could now be larger, and the economic benefits of urbanization could be enjoyed by a far greater populace. Automobiles occupied significantly less street space compared with horses and their wagons. The cost of transportation reduced dramatically. A progressively larger section of the population was able to enjoy the benefits of faster transportation. While it may seem that horses were a safer mode of transport than automobiles, evidence from the past shows otherwise. For instance, the traffic fatality rate in New York City due to horses in 1900 was almost twice that due to automobiles a century later [3]. The modern road system, that owes its existence to the automobile, would be unrecognizable to someone from the 19$^{\text{th}}$ century.

The Woes of the Present

Today, it is the automobile that appears to be unsustainable. Globally, 1.3M lives are tragically cut short each year due to accidents on the road [4]. The US saw significant reductions in traffic fatalities in the late 20$^{\text{th}}$ century, but these numbers have since plateaued. More

than 36,000 lives are lost each year to traffic crashes [5]. A fifth of these fatalities are vulnerable road users such as pedestrians and bicyclists who do not have safety belts or airbags to protect themselves. The efficiency of mobility services leaves much to be desired. The average American spends 97 hours every year stuck in traffic, contributing to 20% of all greenhouse gas emissions in the US. Moreover, 30% of this traffic on average is cruising in the search for parking [6, 7]. The US lags behind most of the developed world in the affordability and accessibility of its mobility services. The burden of inaccessible transportation is disproportionately felt by our most vulnerable groups – low income communities, people with disabilities, and senior citizens [8, 9].

Recent Advances in Urban Mobility

Two recent technologies offer promise in delivering the next leap in urban mobility: (a) transportation network services and (b) automated vehicles. Transportation network services (TNS) are businesses that connect people spread across the transportation network for efficient provision of goods and services. Ride-hailing services such as Uber, and grocery delivery services such as Instacart are quintessential examples. These services provide the freedom and flexibility of car ownership without the need to own a car. Automated vehicles (AVs), popularly known as *self-driving cars*, obviate the need for a human to drive the car.

TNS and AVs have been motivated by an exciting vision of the future. AVs promise to make traffic crashes obsolete by eliminating error-prone humans from the driver's seat. Over the last decade, more than $100B dollars have been invested in AV companies. They are expected to save more than 600,000 lives and $230B in safety costs by 2045 [10]. Uber, the largest ride-hailing platform, created an estimated $7B of consumer surplus in 2016 [11]. The combination of AVs and TNS is expected to substantially reduce traffic congestion by lowering vehicle ownership and increasing occupancy, thus freeing up millions of hours of commute time. Ride-hailing and delivery services have significantly influenced labor markets across the globe by introducing a novel, flexible source of income. This has important equity implications since a significant fraction of the workforce belongs to minority groups [12]. AV and TNS companies also promise to provide affordable services to under-served groups such as people with disabilities and low-income neighborhoods, thus, enhancing equity in mobility.

Unintended Consequences

AVs and TNS companies seem poised to improve safety, bring efficiency and realize equity in urban mobility. However, present reality stands in contrast to this optimistic vision. Although billions of dollars have been invested in AV companies over the last decade, widespread AV deployment has not yet materialized. Companies have realized after years of effort that designing safe AVs is much more difficult than originally anticipated. John Kraf-

cik, the former CEO of Waymo, has stated that full autonomy in AVs may never be achieved [13]. Studies commissioned by ride-hailing services such as Uber and Lyft have found that they have, in fact, increased congestion in major US cities [14]. The urban core of San Francisco has been the worst hit with a 13% increase in vehicle miles travelled. Moreover, 40% of these vehicle miles are covered in search of new trips. The situation is not rosy on the equity front either. After an initial boom in earnings, the prospects of TNS workers have worsened. Several studies have found that a significant proportion of workers are unable to earn the prevailing minimum wage in their cities [15, 16]. This has forced them to work longer hours to make ends meet.

These observations raise two fundamental questions:

1. *Why has the bright AV-TNS future remained elusive?*
2. *What changes in technology and policy can help realize this future?*

We aim to address these two questions over the course of this thesis.

Thesis Outline

Our work on transportation network services and automated vehicles constitute the two halves of this thesis. In the first half (Chapters 2 - 6), we present our work on the economics of transportation network services such as ride-hailing and last-mile delivery, with a specific focus on efficiency and equity concerns surrounding them. In the second half (Chapters 7 - 10), we focus on the safety implications of automated vehicle deployment. The problems we have chosen to focus on reflect the nature and technological maturity of these services. While safety concerns are paramount for a nascent technology such as automated vehicles, they are less of a concern for transportation network services given their technological maturity. We briefly describe the content of individual chapters in this thesis:

I Economics of Transportation Network Services

- In Chapter 2, we provide an overview of the economics of transportation network services, and describe the key challenges facing them.
- In Chapter 3, we introduce a unified economic framework for modeling the transportation network service ecosystem.
- In Chapters 4 and 5, we show how the unified framework can be used to analyze the impact of labor regulations such as Proposition 22 and Assembly Bill 5 on the ride-hailing ecosystem. These chapters are based on work published in [17, 18].
- In Chapter 6, we analyze the relationship between order pooling and the economic viability of delivery services. This chapter is based on work published in [19].

II Automated Vehicle Safety

- In Chapter 7, we provide an overview of the transformative potential of automated vehicles, and safety concerns associated with their deployment.
- In Chapter 8, we use a combination of theoretical models and empirical data to argue why connectivity is indispensable for safe AV deployment. This chapter is based on work published in [20].
- In Chapter 9, we propose a risk assessment framework that leverages human driving data along with AV on-road testing data to develop insights into the safety capabilities of AVs. We present several applications of this framework in Chapter 10 that inform safe AV deployment. The content in these chapters was published in [21, 22].

Part I

Economics of Transportation Network Services

Chapter 2

Economics of Transportation Network Services: An Overview

We start with an overview of the economics of transportation network services. We discuss their history and evolution over the last decade. We then describe some negative consequences of these services including wage effects, roadway congestion, and emissions, as well as regulatory efforts to contain and remedy these effects.

2.1 What are Transportation Network Services?

Transportation network services (TNS) are businesses that connect interested participants to provide efficient delivery of services over an underlying transportation network. These services are offered through a *platform* run by the service provider. The platform informs and connects participants, sets prices and wages, allows them to enter into agreements for the delivery of services, and provides real-time feedback. Examples include platforms that offer ride-hailing (eg., Uber, Lyft), and last-mile delivery of food (eg., DoorDash, UberEats) and groceries (eg., Instacart, Cornershop). For instance, ride-hailing platforms connect pas-

Figure 2.1: Examples of Transportation Network Services.

sengers who seek to go from one place to another with nearby drivers who are looking to monetize their free time. Grocery delivery platforms connect customers in need of supplies, shoppers seeking a source of income, and stores looking to expand their reach.

2.2 Rise of Transportation Network Services

Transportation network services have attracted millions of customers and garnered skyrocketing valuations over the last decade. Ride-hailing was the earliest service to gain wide acceptance and popular adoption, Uber and Lyft being the first entrants in the US. These platforms offered a disruptive business model for urban mobility by (a) leveraging mobile applications to significantly reduce the search friction for rides, (b) providing real-time feedback to customers, and (c) unlocking the potential of flexible part-time labor. The confluence of these factors contributed to their explosive growth in the subsequent decade. As of 2020, Uber had over 5 million drivers in 900 cities of 93 countries, fulfilling 18.7 million trips per day [23, 24]. Today *Uber* is used as a verb synonymous with travel. This is a testament to the ubiquity of ride-hailing in our daily lives. Many other platforms have also enjoyed rapid growth in recent years, such as Lyft in the US, Grab in Singapore, Ola in India, and Didi in China.

Delivery services for groceries (eg., Instacart, Cornershop) and food (eg., DoorDash, Uber Eats) followed soon after ride-hailing. These platforms served more than 50% of the US population and generated revenues of $25B in 2020 [25, 26, 27]. Delivery platforms aim to solve the *last-mile problem* for customers: how to efficiently transport food (groceries/packages) from the restaurant (grocery store/depot) to a customer's doorstep? The growth of last-mile delivery services was, in particular, accelerated by the COVID-19 pandemic. Online delivery accounted for 8.1% ($92B) of all grocery sales ($1.1T) in 2020 as compared with 3.4% ($34B) in 2019 [28]. Similarly, food delivery sales more than doubled in 2020 [26]. This shift in consumer acceptance of delivery services during the pandemic is expected to continue [28]. As such, transportation network services have become an indispensable part of life in large metropolitan cities.

2.3 Value Proposition

Transportation network services create value in three dimensions: (a) system efficiency, (b) customer satisfaction, and (c) labor opportunities. The profitability of TNS companies is derived from their agility. In real time, they are able to modulate supply (eg., drivers) through wages and demand (eg., customers) through prices. By dynamically matching supply and demand, platforms are able to provide services with much greater efficiency.

What explains the enormous popularity of transportation network services among customers? Two major factors are *price* and *convenience*. Trip fares charged by ride-hailing platforms are significantly lower than their taxi counterparts [29, 30]. The convenience of

using cellphones to book rides and receiving real-time feedback on arrival times is a transformative upgrade from the uncertainty associated with traditional taxi cabs. These services have become indispensable for the disabled and the elderly by providing tailor-made solutions to accommodate them. They have also emerged as a crucial source of income for people from minority communities [12].

Transportation network services have enjoyed great success in attracting workers to their platform: drivers for ride-hailing and shoppers for grocery delivery. Workers are attracted to TNS platforms because they can easily monetize their free time. Anyone with a car and a few hours of spare time could make an extra income, while retaining complete control of their working schedule. Workers are encouraged to commit more hours to these services through performance incentives. Flexible working hours, higher earnings and incentives cajoled many workers to quit their day jobs and work full-time for these services.

The COVID-19 pandemic transformed life as we knew it. Benign activities such as grocery shopping or dining out were potentially life-threatening, magnifying the value of transportation network services. Grocery and food delivery companies provided essential services for high risk groups by reducing contact with the general public. Ride-hailing companies offered a safer mode of transport for people whose only other alternative was public transit. The prophylactic nature of these services made them valuable to a society grappling with the challenge of minimizing pandemic spread and protecting its most vulnerable citizens.

2.4 A Darker Side to the Story

However, the TNS success story belies a darker side. We now discuss three negative consequences that have raised concerns regarding their sustainability: (a) increase in traffic congestion, (b) decline in worker wages, and (c) increase in emissions.

Increase in Traffic Congestion

An early analysis of traffic in New York City (NYC) found that traffic speed dropped by 15%, and vehicle miles traveled (VMT) increased by 36% due to the growth of ride-hailing services [31]. Studies conducted by the San Francisco Country Transportation Authority (SFCTA) estimated that 50% of the increase in traffic congestion in San Francisco was contributed by these services [32, 33]. In response to these findings, Uber and Lyft commissioned a study to estimate their share of VMT in six major metropolitan regions in 2019. The study found that ride-hailing platforms indeed had a significant impact on VMT especially in urban centers, ranging from 1.7% (Seattle) to 12.2% (San Francisco) [14].

Declining Worker Wages

Ride-hailing platforms enjoyed enormous popularity among drivers in the early 2010s in part due to generous subsidies from investors. However, these subsidies proved untenable

and despite their growing market size, most platforms are still struggling today to find a path toward profitability. This has led to increased tensions between platforms and their drivers. In 2018, labor economists James Parrott and Michael Reich were commissioned by the New York City Taxi and Limousine Commission (NYCTLC) to investigate the earnings of platform drivers in NYC [15]. They showed that the mean pre-expense earnings of platform drivers declined from $27.74/hr in 2016 to $24.49/hr in 2017. They also found that a majority of drivers work full-time in NYC, and 85% of the drivers make less than the minimum wage of the city (after accounting for operating expenses such as gas, maintenance, insurance, etc.). Parrott and Reich also examined the pay and working hours of platform drivers in Seattle [16]. They reported that platform drivers in Seattle have hourly earnings of $21.53 before expenses. However, after expenses of $11.80, a driver nets $9.73/hr – considerably lower than Seattle's $16/hr minimum wage. They also reported that 32% of platform drivers work full-time (more than 32 hours per week) and provide 55% of all trips. More than four-fifths of full-time drivers purchased their vehicle primarily or partly to provide platform services and almost three-fourths of full-time drivers rely on driving for platforms as their sole source of income.

Increase in Emissions

The increase in emissions is a serious concern associated with the proliferation of transportation network services. On average, ride-hailing trips contribute 70 percent more climate pollution than the trips they replace [34]. The impact of ride-hailing on the climate can be attributed to two main factors. Firstly, ride-hailing trips tend to be more polluting compared to personal car trips due to the additional miles traveled without passengers (also referred to as *deadheading*). Secondly, ride-hailing does not merely replace personal car trips, but also leads to an increase in the overall number of car trips. In the absence of ride-hailing, many passengers would have opted for alternative transportation methods, such as mass transit, walking, biking, or even canceling the trip altogether.

2.5 A Clarion Call for Regulation

These efficiency and equity concerns have driven cities to regulate ride-hailing platforms. In December 2018, the NYCTLC established a minimum per-trip payment formula that guarantees an equivalent minimum wage of $17.22/hr for platform drivers [35]. A similar wage floor was also adopted by the city of Seattle, which roughly amounts to the city's $16 minimum wage after expenses [36]. The impact of these wage regulations was evaluated by [37], where the authors showed that a wage floor on platform driver earnings will push platforms to hire more drivers and lower overall ride prices, which benefits both passengers and drivers at the cost of reducing platform profits.

Assembly Bill 5

A broader set of protections was adopted by the state of California to protect workers in the gig economy including platform drivers. In September 2019, California legislators passed Assembly Bill 5 (AB5) that requires companies like Uber and Lyft to treat their drivers as employees. Under this regulation, a worker is classified as an independent contractor only if *all* of the following conditions are met:

1. the worker is free from the control and direction of the hirer in connection with the performance of the work, both under the contract for the performance of such work and in fact,
2. the worker performs work that is outside the usual course of the hiring entity's business,
3. the worker is customarily engaged in an independently established trade, occupation, or business of the same nature as the work performed for the hiring entity.

Under AB5, workers who do not qualify as independent contractors are treated as employees and receive the prevailing minimum wage and benefit protections. Ride-hailing platforms such as Uber and Lyft were ordered to classify their drivers as employees under AB5 [38].

Proposition 22

Ride-hailing platforms resisted AB5 and maintained that their drivers were independent contractors as they had the flexibility to choose their work schedule. Labor protection advocates argued these services already exerted substantial control over their drivers by dynamic pricing, which effectively nudges drivers to work only at profitable times with high hourly rates. This sparked a year-long legal battle on the employment status of platform drivers. Platforms crafted and lobbied fiercely for an alternative regulation codified in Proposition 22 (Prop 22).

In November 2020, the State of California passed Prop 22, which exempts platforms from AB5 and allows them to classify their drivers as independent contractors instead of employees. This ballot measure was supported by prominent companies in the gig economy including platforms such as Uber and Lyft, who collectively contributed over $205 million for the "Yes on Prop 22" campaign, making it the most expensive ballot in California history. Under Prop 22, platforms are exempted from providing a full suite of employee benefits to their drivers (e.g., overtime pay, sick leave, bargaining rights, etc). In exchange, they offered to pay at least 120% of the prevailing minimum wage for driver's *engaged time*, which does not include the time spent cruising while waiting for the next trip request. Note that engaged time constitutes less than two-thirds of all driving time in major US cities [14]. Since the driver wage floor is only placed on the engaged time, the effective hourly wage for platform drivers can be significantly lower than the prevailing hourly minimum wage. Such nuances of Prop 22 have to be well understood before it can be reproduced in other states where gig worker classification is under contention.

Congestion Taxes

In order to alleviate the increase in congestion due to ride-hailing platforms, NYCTLC introduced a $2.75 flat congestion charge for any ride-hailing trip starting from, ending in, or passing through the *congestion zone* of the city (south of 96th Street in Manhattan) [35]. Similar regulations have been passed in other cities in the US. The city of San Francisco instituted a congestion mitigation tax ranging from 1.5% (shared rides, zero emission vehicles) to 3.25% (single passenger rides) in January 2020 [39]. Chicago passed a similar tax for ride-hailing trips in its downtown areas [40].

Emissions Standards

In 2021, the California Air Resources Board (CARB) approved a regulation that requires ride-hailing platforms to commence the electrification of their fleets [41]. By 2030, the regulation will require ride-hailing platforms to attain zero greenhouse gas emissions and ensure that 90% of their vehicle miles are fully electric, bringing them in line with other corporate fleet requirements. This move is a significant step towards fulfilling the state's air quality goals, which may have broader implications for other regions seeking to address climate change. In response, companies like Uber and Lyft have announced programs with great fanfare to limit their carbon footprint by embracing electric vehicles in their fleets.

2.6 Unfavorable Economics

Despite more than a decade of being in operation, most of the leading TNS platforms are yet to make a profit. Uber posted a net loss of $9.1 billion in 2022 [42]. Even though revenue for delivery services surged during the pandemic, they still remain loss-making businesses. DoorDash, the largest food delivery service in the US by market share, posted a loss of $468M in 2021 [43]. Instacart, a major grocery delivery platform, recently slashed its valuation by 67%, from $39B to $13B citing challenging market conditions [44]. Worker wages and per-unit transportation costs seem to be the primary drivers of this bleak outlook. Regulations further threaten profitability.

2.7 Pressing Questions

This raises several pressing questions about the economic viability of transportation network services and regulations governing their operations:

- *What explains the unfavorable economics of transportation network services?*
- *What changes in operations/business model/technology can improve their economic viability?*

- *Are current labor regulations effective in protecting TNS workers? Are there alternative regulations that both protect workers and preserve viability of these services?*

We develop tools to address the above questions in the first half of the thesis. Answering these questions requires an analytical framework that captures the complex interactions between participants in the TNS ecosystem: workers, customers, platforms, and the regulatory authority. We present this framework in the next chapter.

Chapter 3

Modeling Transportation Network Services

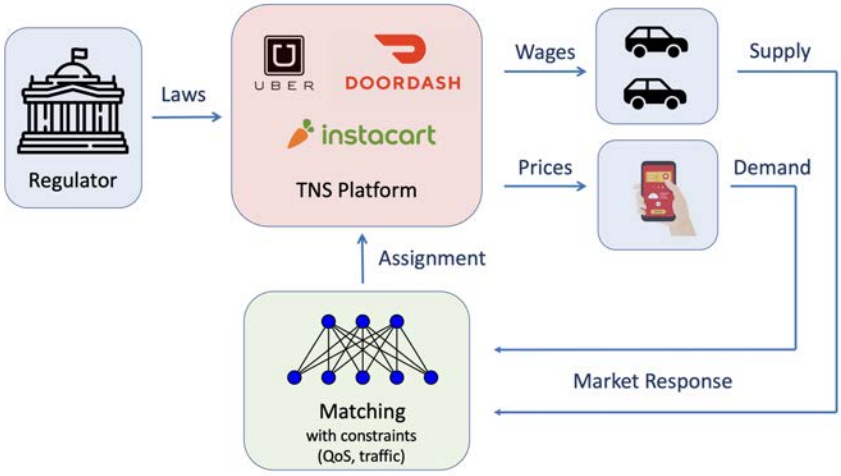


Figure 3.1: Transportation Network Service Ecosystem.

Diverse transportation network services such as ride-hailing and delivery share a common underlying structure, as illustrated in Figure 3.1. There are four key agents in the TNS ecosystem: (a) workers who provide services, (b) customers of these services, (c) the platform that matches workers to customers, and (d) the regulatory authority that aims to safeguard the interests of workers, customers, as well as other transportation network users.

As seen in Figure 3.1, the platform can use the levers of worker wages and service fees to modulate worker supply and customer demand respectively. This induces a certain distribution of workers and customers across the transportation network to join the platform. The platform then matches workers and customers to maximize profit, while ensuring that quality of service (QoS) constraints are respected. The platform can optimize its operations at two levels: (a) *economic*: optimally setting wages and service fees to onboard workers and

customers, (b) *algorithmic*: optimally matching workers and customers given wages, fees, underlying transportation network topology, and traffic conditions. The regulatory authority dictates the constraints within which the platform can operate. We now present a unified economic framework that captures such interactions between agents in the TNS ecosystem.

3.1 A Unified Modeling Framework

We consider a TNS platform matching customer demand with worker supply. For simplicity, we consider a representative hour of operation (for eg., 9 AM - 10 AM on Wednesday), and a particular geographic area of service (for eg., San Francisco, California). We note that parameters in our model such as supply, demand, wages, and fees vary with time and location. In subsequent chapters, we extend this model to account for such variations.

3.1.1 Customer Demand and Worker Supply

The customer demand per unit time λ depends on the fee p per unit of service and quality of service (QoS) Q, i.e.,

$$\lambda = D(p, Q). \tag{3.1}$$

For instance, the likelihood of a customer choosing to travel via a ride-hailing platform depends on both the trip fee p and the time t^{w} they need to wait until the driver's arrival (*waiting time*). In this context, $Q = -t^{\mathrm{w}}$ is a useful QoS metric (lower the waiting time, higher the QoS). The number of workers N providing the required service depends on the wage w offered by the platform, i.e.,

$$N = J(w). \tag{3.2}$$

We refer to the customer demand and worker supply jointly as the *market response*.

3.1.2 Quality of Service

Quality of service (QoS) is a critical consideration for customers. A passenger's decision to choose a ride-hailing service for a trip depends on their *waiting time* (duration between placing the trip request and the driver's arrival). For food and grocery delivery services, *delivery time* (duration between customer placing the order, and order being delivered) is an important QoS metric. In particular, grocery delivery platforms offer differentiated pricing based on the delivery time that a customer is willing to accept.

Recognizing this, the platform guarantees a minimum quality of service $Q^{\min}$ to its customers. The QoS depends on customer demand λ, worker supply N, and the matching algorithm γ used by the platform, i.e.,

$$Q = G(\lambda, N, \gamma), \tag{3.3}$$

$$Q \geq Q^{\min}. \tag{3.4}$$

Observe that the customer demand λ and quality of service Q are coupled through equations (3.1) and (3.3). This makes intuitive sense. Passenger demand in ride-hailing depends on waiting time, while waiting time depends on the number of idle drivers which, in turn, is a function of passenger demand.

3.1.3 Matching Algorithm

The platform matches workers with customers. It uses a matching algorithm γ. The inputs of γ are the market response, and the output is the worker-customer assignment. For example, an intuitive matching algorithm in the ride-hailing case is to match a passenger with the nearest idle driver. Another plausible algorithm would be to pool nearby passengers together and match them to the same driver. Note that there is a trade-off between service cost and QoS in the choice of matching algorithm. For instance, pooling together passengers can result in lower trip cost per passenger but a higher waiting time. The platform must navigate such trade-offs while selecting its matching algorithm.

3.1.4 Regulatory Intervention

The regulatory authority imposes constraints on the platform's operations with the aim of protecting other participants in the TNS ecosystem. A guaranteed hourly minimum wage is an example of a regulation aimed at safeguarding worker interests. This can be modeled as the constraint,

$$w \geq w^{\min}, \tag{3.5}$$

where $w^{\min}$ is the guaranteed hourly minimum wage. More generally, regulations can place constraints on a broader set of model variables such as service fees and worker supply. A general regulation can be expressed as

$$R(p, w, \lambda, N) \geq 0. \tag{3.6}$$

In Chapters 4 and 5, we will discuss how recent legislations in California such as Proposition 22 and Assembly Bill 5 can be modeled within our framework.

3.1.5 Optimal Service Fee and Wage

The platform aims to maximize its profit. For this, it selects the service fee p and worker wage w while ensuring a minimum QoS and regulatory compliance. Thus, the platform's

pricing problem can be expressed as

$$\Pi = \max_{p\geq 0, w\geq 0} \lambda p - Nw \tag{3.7}$$

$$\text{s.t. } \lambda = D(p, Q), \qquad \textit{Customer Demand} \tag{3.7a}$$

$$N = J(w), \qquad \textit{Worker Supply} \tag{3.7b}$$

$$Q = G(\lambda, N, \gamma) \geq Q^{\min}, \qquad \textit{Quality of Service} \tag{3.7c}$$

$$R(p, w, \lambda, N) \geq 0. \qquad \textit{Regulatory Compliance} \tag{3.7d}$$

We now describe how this framework can be used to model the economics of ride-hailing platforms.

3.2 Example: Modeling Ride-hailing Platforms

Ride-hailing platforms such as Uber and Lyft provide an app-based interface that connects passengers and drivers. The platform modulates the passenger demand and driver supply by adjusting trip fares and driver wages.

3.2.1 Passenger Demand and Driver Supply

Passengers have several travel options: ride-hailing, and alternatives such as personal vehicle, public transit, and biking. Let λ^0 denote the total number of passengers considering the platform among other options for completing their trip. The platform is chosen if its trip cost is lower than the cost of alternate modes. From the passenger's perspective, the two main costs associated with a trip are the fare p (*service fee*) charged by the platform, and the time t^{w} that they need to wait between requesting the trip and the driver's arrival (*waiting time*). Recall from our discussion in Section 3.1 that the QoS in this context is given by $Q = -t^{\text{w}}$. The cost of a ride-hailing trip can thus be modeled as $\overline{p} = p + \alpha t^{\text{w}}$. Here α reflects the perceived cost associated with waiting time. We model the total passenger pool as having heterogeneous alternate mode costs. Let $C(\cdot)$ denote the cumulative distribution function (CDF) of the least expensive non-platform mode choice for this population. Then, the total number of passengers choosing the platform at trip fare p and waiting time t^{w} is given by

$$\lambda = \lambda^0(1 - C(p + \alpha t^{\text{w}})). \tag{3.8}$$

Drivers have several employment options. Let N^0 denote the total labor supply considering the platform as one of their work choices. Let w denote the wage offered by the platform. Let $S(\cdot)$ represent the CDF of reservation wages of the heterogeneous supply pool. Then, the driver supply is given by

$$N = N^0 S(w). \tag{3.9}$$

3.2.2 Quality of Service: Waiting Time

Recall that t^{w} denotes the passenger waiting time. Intuitively, the waiting time varies inversely with the number of idle drivers, i.e., those that are not engaged in a trip. Let t^{s} denote the average engaged time per trip, then λt^{s} is the total engaged time across all drivers during an average hour. Since N is the total number of drivers and λt^{s} is the number of drivers engaged in a trip, the number of idle drivers is given by $N - \lambda t^{\mathrm{s}}$. It has been shown that [45, 37] that the passenger waiting time t^{w} is inversely proportional to the square root of the number of idle drivers, i.e.,

$$t^{\mathrm{w}} = \frac{M}{\sqrt{N - \lambda t^{\mathrm{s}}}}. \tag{3.10}$$

Here, M is a constant that depends on city roadway geometry and average vehicle speed. This model agrees well with empirically observed passenger waiting times [46].

3.2.3 Platform's Pricing Problem

The platform aims to maximize its profit by setting trip fare p and driver wage w. This effectively determines the number of trips λ and drivers N based on (3.8) and (3.9). Since λp represents the total fares paid by all passengers and Nw represents the total wages paid to drivers, the platform's profit is given by $\lambda p - Nw$. In the absence of regulation, the platform's pricing problem can be expressed as

$$\Pi = \max_{p \geq 0, w \geq 0} \quad \lambda p - Nw \tag{3.11}$$

$$\text{s.t. } \lambda = \lambda^0 \left(1 - C\left(p + \frac{\alpha M}{\sqrt{N - \lambda t^{\mathrm{s}}}} \right) \right), \tag{3.11a}$$

$$N = N^0 S(w), \tag{3.11b}$$

where profit is maximized by appropriately setting the trip fare p and driver wage w which induce passenger demand λ and driver supply N.

To use this model for analyzing the economics of ride-hailing, we must calibrate it based on the region of operation. In the next section, we present our model calibration approach for the city of San Francisco (SF).

3.3 Model Calibration

The passenger demand and driver supply functions can be estimated using historical data for passenger demand, driver supply, fares, and wages $(\hat{\lambda}, \hat{N}, \hat{p}, \hat{w})$. Since these quantities are not publicly available for SF, we estimate them from a combination of sources.

3.3.1 Estimating Trip Fares and Driver Wages

We use $\hat{\lambda} = 7156$ trips/hr and $\hat{N} = 2781$ drivers/hr for an average weekday based on a study conducted by the San Francisco County Transportation Authority [32]. A report commissioned by Uber and Lyft found that trip time contributed to 67% of all platform vehicle miles travelled in SF [14]. Since drivers can spend their idle time by choosing to either cruise on the roads or wait on the curbside till their next trip request, the driver utilization rate would be less than 67%. Since curbside parking is difficult to find in SF, we assume that the prevailing utilization rate $\hat{\rho} = 0.6$, which amounts to roughly three-fourth of idle time spent cruising. Substituting $\hat{\lambda}, \hat{N}$ and the utilization rate $\hat{\rho} = 0.6$ in (4.2), we get the engaged time $t^s = 14$ min. Based on [14], one-sixth of engaged time is spent picking up a passenger. This amounts to half of all driving time spent transporting a passenger from her pick-up location to drop-off location. Since the average speed in SF is 14 miles/hr [33] and each driver provides $\hat{\lambda}/\hat{N} = 2.57$ trips/hr, the average trip length is $(0.5 \times 14)/2.57 = 2.72$ miles and takes 11.7 minutes. This can be used to estimate the average trip fare $\hat{p}$ based on Lyft's fare card [47]. This includes a base fare of \$2.24, a \$3 service fee, a \$0.93 per mile charge, and a \$0.40 per minute charge. Plugging in the trip length and duration, we estimate the average fare to be $\hat{p} = \$12.44$. Considering that Lyft takes 100% of the service fee and a 25% commission on the rest of the fare, the average (pre-expense) driver earning per trip comes out to \$7.08. Thus, the average hourly driver earning is $\hat{w} = \hat{\lambda}\hat{p}/\hat{N} = \$18.2$.

3.3.2 Accounting for Driver Expenses

There is no consensus on how driving expenses should be accounted for in order to estimate the post-expense earnings of ride-hailing drivers [48, 49, 50]. While operating expenses such as fuel and vehicle maintenance should be taken into account for all drivers, fixed expenses such as vehicle insurance, depreciation, and financing costs depend on whether a driver works full-time for the platform. For instance, drivers who work part-time for the platform are likely to already own a car and would be paying insurance and incurring depreciation costs even if they were not driving for the platform. On the other hand, a significant proportion of full-time drivers own or lease a car for the purpose of driving for the platform and hence, these fixed expenses should be included in their case [16]. Expense estimates range from \$0.19 to \$0.58 per mile depending on whether fixed expenses are considered [49, 48]. We consider ride-hailing drivers in SF as full-time if they work more than 25 hrs/week. Uber has reported that 26% of its drivers in California drive more than 25 hrs/week and provide 58% of all trips [51]. We consider a driver expense of \$0.42 per mile based on a weighted average of the above estimates. Accounting for three-fourth of idle time spent cruising and an average speed of 14 miles/hr, each driver covers a total distance of 12.53 miles in an average hour. This results in the driver expense being D = \$5.26/hr.

3.3.3 Estimating Model Parameters

The model parameters $\{\lambda^0, N^0, C(\cdot), S(\cdot)\}$ can now be estimated using $\{\hat{\lambda}, \hat{N}, \hat{p}, \hat{w}\}$. Constrained by data availability[1], we consider linear CDFs for alternate travel mode costs $C(\cdot)$ and driver reservation wages $S(\cdot)$, i.e.,

$$C(\overline{p}) = \min\{e^{\mathrm{p}}\overline{p}, 1\}, \tag{3.12}$$

$$S(w) = \min\{e^{\mathrm{d}}w, 1\}, \tag{3.13}$$

where e^{p} and e^{d} are the slopes of the linear CDFs. The platform's pricing problem can then be expressed as

$$\Pi = \max_{p\geq 0, w\geq 0} \ \lambda p - Nw \tag{3.14}$$

$$\text{s.t. } \lambda = \lambda^0\left(1 - e^{\mathrm{p}}\left(p + \frac{\alpha M}{\sqrt{N - \lambda t^{\mathrm{s}}}}\right)\right), \tag{3.14a}$$

$$N = N^0 e^{\mathrm{d}} w. \tag{3.14b}$$

Even though the above problem is non-convex, authors in [37] show that first-order optimality conditions are sufficient to find the optimum as long as (3.14) has a feasible solution with strictly positive objective value (please see Proposition 2 and 3 in [37]). Note that these optimality conditions must be satisfied by the historical data $\{\hat{\lambda}, \hat{N}, \hat{p}, \hat{w}\}$, which induces a set of equations involving model parameters $\{\lambda^0, e^{\mathrm{p}}, N^0 e^{\mathrm{d}}, \alpha M\}$. We calibrate our model by solving this set of equations.

[1]The limited data availability to estimate the passenger demand and driver supply functions constrains us to single-parameter CDFs for alternate travel mode costs $C(\cdot)$ and driver reservation wages $S(\cdot)$. In addition, we restrict our attention to linear CDFs for analytical tractability of the time-varying model developed in Section 5.3. The estimated linear passenger demand and driver supply functions can be viewed as local approximations of the actual functions at the prevailing platform operational point.

Chapter 4

Impact of Labor Regulations: Proposition 22

In Chapter 3, we proposed a unified framework for modeling transportation network services and showed how the specific case of ride-hailing can be modeled. Using this model, we can analyze the impact of labor regulations on the ride-hailing ecosystem. In this chapter, we present our analysis of the impact of CA Proposition 22 (Prop 22) in San Francisco, California.

4.1 Introduction

As discussed in Chapter 2, concerns about declining worker wages and increasing congestion have motivated calls for regulation of ride-hailing platforms. In particular, two recent regulations passed in California have garnered significant attention: Assembly Bill 5 (AB5) and Proposition 22 (Prop 22). AB5 classifies drivers (and other gig-economy workers) as employees as opposed to contractors. The implication is that ride hailing companies must pay drivers a minimum wage and associated benefits, and that the drivers work full-time. As an alternative, ride hailing companies lobbied fiercely for Prop 22, a ballot measure which was eventually passed in November 2020. Prop 22 allows ride-hailing platforms to classify their drivers as independent contractors instead of employees. This exempts from providing employee benefits such as overtime pay and sick leave to their drivers. To hold their end of the bargain, they offered to pay at least 120% of the prevailing minimum wage for drivers' *engaged time*, which does not include time spent cruising on the streets while waiting for the next trip request. Note that engaged time accounts for less than two-thirds of all driving time in major US cities [14]. Since the wage floor is placed solely on the engaged time, the *effective hourly wage* floor for ride-hailing drivers turns out to be significantly lower than the prevailing hourly minimum wage.

Understanding the impact of Prop 22 is crucial since similar regulations are being considered across the globe. France's top court recognized ride-hailing drivers as employees in

March 2020 [52]. In February 2021, Britain's Supreme Court ruled that Uber should reclassify its drivers as "workers" – a classification that entitles them to benefits such as a minimum wage, vacation pay and access to a pension plan [53]. European lawmakers are considering similar laws that would apply across the European Union [54]. Recently, U.S. labor secretary Marty Walsh also suggested a policy shift in this direction [55]. In response, platforms have lobbied for regulations modeled around Prop 22 that would maintain the independent contractor status of drivers, claiming that reclassification as employees would preclude drivers from enjoying the flexibility of choosing their work hours and location [56]. Platforms also argue that it would put thousands of drivers out of work [57]. An analysis of Prop 22 and AB5 can provide insights into the potential consequences of reclassification of drivers, and help inform regulations that protect drivers without threatening the viability of the platform business model.

4.2 Summary of Results

We summarize the results of our analysis below:

- We find that the engaged time minimum wage guaranteed by Prop 22 will not increase effective driver wages. This is because the wage premium offered to compensate for idle driver time is not sufficiently high.
- We study the effect of increasing the engaged time minimum wage on effective driver wages. Increasing the engaged time minimum wage initially results in an increased effective wage. Counter-intuitively, further increases can reduce the effective driver wage. In the case of SF, we show that increasing the wage premium to 200% can ensure that drivers make the prevailing minimum wage. However, this comes at the cost of a 17% decrease in utilization rate, thus reducing system efficiency. It would also lead to a 17% increase in fares, and a 22% decrease in platform profit. Thus, we conclude that a minimum wage covering only engaged time is not an effective mechanism for increasing driver wages.
- Recognizing that increased driver wages come at the cost of adverse outcomes for other participants in the ride-hailing ecosystem, we propose a modification to Prop 22. We show that incorporating an idle time wage component in Prop 22 would increase utilization rate and platform profit, and reduce passenger fares.

4.3 Analysis under Proposition 22

Under Prop 22, the platform is required to pay drivers a minimum wage w^{pr} per hour of engaged time. This wage lower bound consists of two components: (i) 120% of prevailing hourly minimum wage, and (ii) \$0.3 per mile to compensate for driver expenses. Let $\hat{w}^{\min}$

and v denote the prevailing hourly minimum wage and average speed respectively. Then,

$$w^{\text{pr}} = 1.2\hat{w}^{\min} + 0.3v. \tag{4.1}$$

Note that Prop 22 guarantees a minimum wage only during engaged time. We can translate such an engaged time minimum wage w^{pr} to a minimum wage w^{eff} per hour of total driving time. We refer to w^{eff} as the *effective minimum wage*.

4.3.1 Driving Time and Utilization Rate

The total driving time for ride-hailing drivers has three components: (i) picking up a passenger (T_1), (ii) taking the passenger from pick-up location to drop-off location (T_2), and (iii) waiting for a trip request (T_3). We use the term *engaged time* to refer to $T_1 + T_2$, i.e., the time between accepting the trip request and completing the trip. The terms *passenger waiting time* and *idle driver time* refer to T_1 and T_3 respectively. The utilization rate ρ is the ratio of engaged time to total driving time. Recall that λt^{s} is the total engaged time across all drivers during an average hour. Since there are N drivers providing trips every hour on average, the total driving time across all drivers during an average hour is also given by N. Thus, we have

$$\rho = \frac{\lambda t^{\text{s}}}{N}. \tag{4.2}$$

Using (4.2), we can express the constraint imposed by Prop 22 as a lower bound on the hourly driver wage w,

$$w \geq \left(\frac{\lambda t^{\text{s}}}{N}\right) w^{\text{pr}}. \tag{4.3}$$

Thus, the platform solves the following optimization problem under Proposition 22,

$$\Pi^{\text{pr}} = \max_{p\geq 0, w\geq 0, N\geq 0} \ \lambda p - Nw \tag{4.4}$$

$$\text{s.t. } \lambda = \lambda^0 \left(1 - C\left(p + \frac{\alpha M}{\sqrt{N - \lambda t^{\text{s}}}}\right)\right), \tag{4.4a}$$

$$N \leq N^0 S(w), \tag{4.4b}$$

$$w \geq \left(\frac{\lambda t^{\text{s}}}{N}\right) w^{\text{pr}}. \tag{4.4c}$$

4.3.2 Impact of Prop 22

Observe that the wage lower bound (4.4c) is the only additional constraint imposed by Prop 22. If the prevailing values of (λ, N, p, w) under no regulation satisfy (4.4c), then the platform does not need to change its pricing decision under Prop 22. Since Prop 22 applies at the

state level, we set $\hat{w}^{\min} = \$14$ based on California's minimum wage. Substituting $w_{\min}$ and the average SF speed $v = 14$ miles/hr in (4.1), the engaged time minimum wage $w^{\text{pr}} = \$21$. Using the utilization rate $\hat{\rho} = 0.6$, we estimate the effective minimum wage under Prop 22 to be $\hat{w}^{\text{eff}} = \hat{\rho} w^{\text{pr}} = \$12.6$. Notice that the prevailing hourly wage $\hat{w} = \$18.2$ is substantially larger than the effective minimum wage under Prop 22. This implies that the solution of the platform's pricing problem (4.4) under Prop 22 would be the same as that under no regulation (3.11). Thus, Prop 22 in its current form will not increase driver wages.

Jacobs and Reich [58] argue that the engaged time wage guaranteed by Prop 22 does not adequately protect driver wages since it under-compensates idle driving time and driver expenses. Although the premium of 120% over the prevailing minimum wage $\hat{w}^{\min}$ in (4.1) is meant to account for idle time, it is insufficient to ensure that drivers effectively earn the hourly minimum wage. Considering the prevailing driver utilization rate $\rho = 0.6$ in SF, a wage premium of $1/0.6 = 1.67$ would correspond to drivers being compensated adequately for idle time. The \$0.3/mile driver expense component in Prop 22's minimum wage does not adequately compensate full-time drivers as suggested by our \$0.42/mile estimate in Section 3.3. Moreover, it only applies during engaged time, thus ignoring cruising miles that account for one-third of total driving miles in SF [14].

The prevailing minimum wage in SF is \$16.07/hr. In addition, drivers will need to pay payroll taxes and cannot avail of employee benefits since they remain classified as independent contractors under Prop 22. Based on [58], the added costs from taxes and missed employee benefits amount to \$1.13/hr. Thus, the independent contractor-adjusted minimum wage in SF for ride-hailing drivers is \$17.2/hr. Thus, drivers must make $w^{\min} = 17.2 + \text{D} = \$22.46$/hr in order to earn the hourly minimum wage in SF after expenses. Note that we use $w^{\min}$ to refer to the pre-expense wage which ensures that drivers make the hourly minimum wage (\$16.07/hr) after expenses. Thus, the driver wage that remains unchanged under Prop 22, $\hat{w} = \$18.2$, is more than \$4 below $w^{\min}$.

4.3.3 Increasing Engaged Time Wage

This raises the question: *How should Prop 22 be modified to ensure that drivers effectively earn the hourly minimum wage?* The obvious approach is to increase the engaged time wage such that idle time and driver expenses are sufficiently compensated. Let us increase the driver expense component in (4.1) to \$0.42/mile, which corresponds to D = \$5.26 as discussed in Section 3.3. Since the driver utilization rate is a function of λ and N that are determined by the platform's pricing decision, the utilization rate can change when the engaged time minimum wage is increased. Therefore, the wage premium of 1.67 based on $\rho = 0.6$ does not guarantee that drivers make minimum wage. The engaged time minimum wage under Prop 22 can be modified as follows:

$$w^{\text{pr}} = \phi w^{\min}, \tag{4.5}$$

where ϕ denotes the constant by which the prevailing pre-expense hourly minimum wage is multiplied to compensate drivers for idle time. We refer to it as the *wage multiplier*. Under

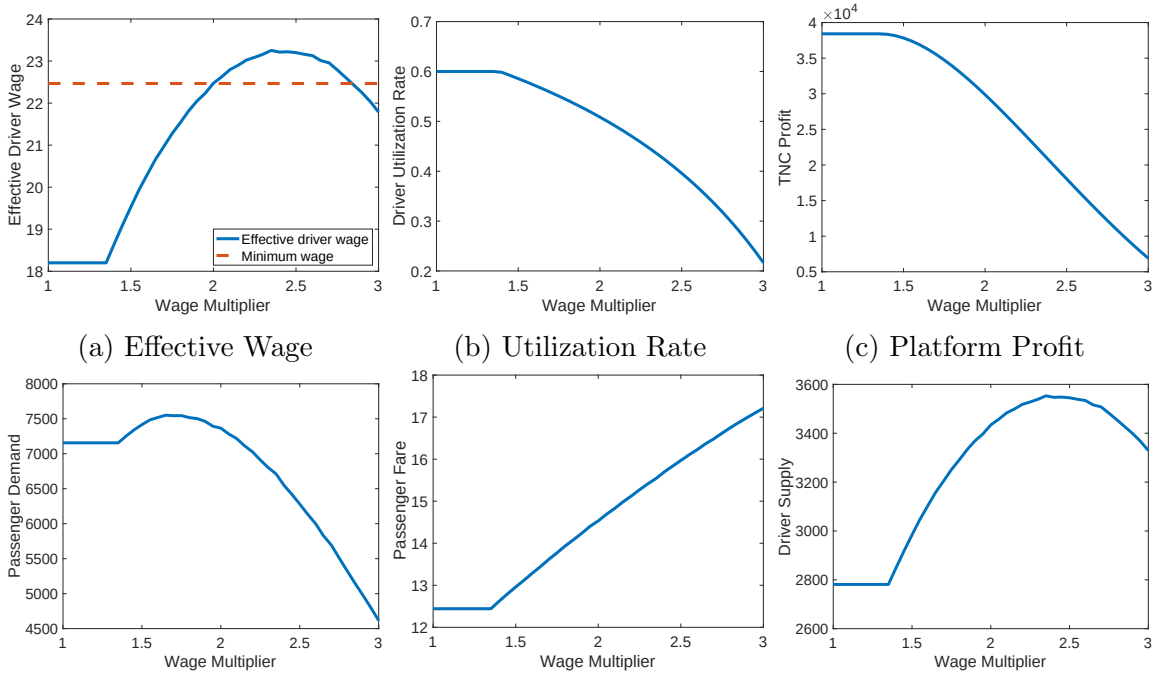
Prop 22, $\phi = 1.2$, although it does not cover driver expenses during idle time. Intuitively, it seems that increasing ϕ should increase the driver wage w. With this objective, we consider the following optimization problem for the platform as a function of the wage multiplier ϕ and observe how w varies with increase in ϕ:

$$\Pi^{\text{pr}}(\phi) = \max_{p\geq 0, w\geq 0, N\geq 0} \lambda p - Nw \tag{4.6}$$

$$\text{s.t. } \lambda = \lambda^0\left(1 - C\left(p + \frac{\alpha M}{\sqrt{N - \lambda t^{\text{s}}}}\right)\right), \tag{4.6a}$$

$$N \leq N^0 S(w), \tag{4.6b}$$

$$w \geq \phi\left(\frac{\lambda t^{\text{s}}}{N}\right) w^{\min}. \tag{4.6c}$$



(a) Effective Wage (b) Utilization Rate (c) Platform Profit

(d) Passenger Demand (e) Trip Fare (f) Driver Supply

Figure 4.1: Ride-hailing ecosystem outcomes under varying wage multiplier ϕ.

4.3.4 Results and Discussion

We find that the variation of effective driver wage w with ϕ can be divided into three regimes as illustrated in Figure 4.1a. Initially, w is constant despite an increase in ϕ since the modified wage constraint (4.6c) is inactive. Once this constraint becomes active at $\phi = 1.4$, the effective hourly wage monotonically increases up to $\phi = 2.35$. Counter-intuitively, we

find that the driver wage decreases with increase in ϕ beyond this point. Since Prop 22 guarantees wages during engaged time rather than total driving time, it incentivizes the platform to reduce the driver utilization rate so that the effective hourly wage it needs to pay drivers remains low. We find that the above trend holds for a wide range of model parameters (see sensitivity analysis in Appendix A). We also find that drivers would make the prevailing minimum wage in SF at $\phi = 2$. However, this comes at the expense of a 16.6% decrease in utilization rate, a 16.8% increase in passenger fares, and a 22.2% decrease in platform profit compared with the no regulation setting. As described above, the main reason behind these adverse outcomes is that the platform is incentivized to increase the fraction of idle driver time since Prop 22 only guarantees driver wages during engaged time. Thus, increasing the engaged time wage alone would have adverse effects on the entire ride-hailing ecosystem. We discuss in Section 4.4 how adding an idle time wage component to (4.6c) can align the platform's incentives towards improving its utilization rate, and hence improve outcomes for all participants in the ecosystem.

4.4 An Alternative Proposal

Our analysis shows that the *status quo* would be maintained under Prop 22, i.e., the significant fraction of drivers who work full-time would continue to make below minimum wage. Increasing the engaged time wage in order to increase driver wages results in adverse outcomes for all participants in the ecosystem. We now present a suitable modification to Prop 22 that would improve outcomes for drivers and passengers without substantially affecting the platform's profitability.

Adding an Idle Time Wage to Prop 22

As discussed in Section 4.3.3, increasing the engaged time wage such that drivers earn the effective minimum wage comes at the cost of adverse outcomes for the entire ride-hailing ecosystem. Since Prop 22 institutes a lower bound on the engaged time driver wage, it allows platforms to keep the effective hourly driver wages low by reducing the driver utilization rate. Thus, a very high engaged time wage ($w^{\text{en}} = \$44.9$) is required to ensure that drivers earn the effective minimum wage. However, this is accompanied by a greater fraction of idle drivers, thereby contributing to increased congestion on the roads. Adding an idle time wage component in Prop 22 can dissuade platforms from reducing utilization rates by making them internalize the cost of idle driving time.

Let w^{en} and w^{id} denote the guaranteed wage during engaged and idle time respectively. Then, the effective hourly wage lower bound under our modified proposal is given by

$$w \geq \left(\frac{\lambda t^{\text{s}}}{N}\right) w^{\text{en}} + \left(1 - \frac{\lambda t^{\text{s}}}{N}\right) w^{\text{id}}. \tag{4.7}$$

Thus, the platform solves the following pricing problem,

$$\Pi^{\mathrm{cr}} = \max_{p\geq 0, w\geq 0, N\geq 0} \ \lambda p - Nw \tag{4.8}$$

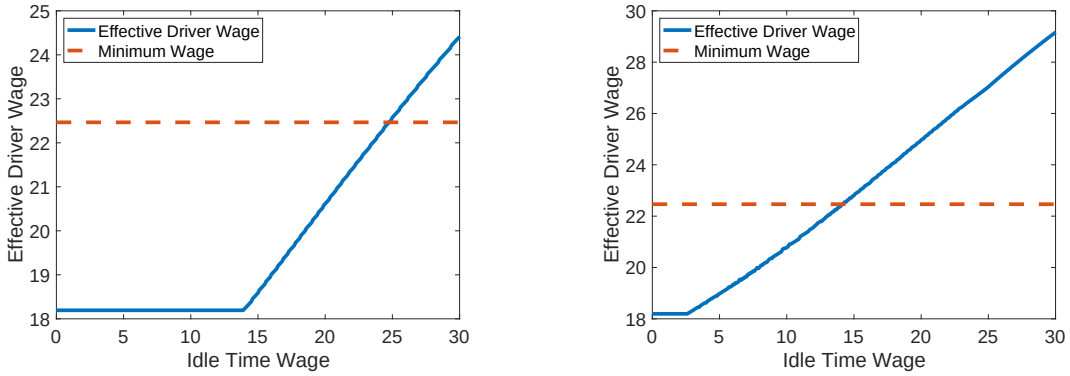
$$\text{s.t. } \lambda = \lambda^0\left(1 - e^{\mathrm{p}}\left(p + \frac{\alpha M}{\sqrt{N - \lambda t^{\mathrm{s}}}}\right)\right), \tag{4.8a}$$

$$N \leq N^0 e^{\mathrm{d}} w, \tag{4.8b}$$

$$w \geq \left(\frac{\lambda t^{\mathrm{s}}}{N}\right) w^{\mathrm{en}} + \left(1 - \frac{\lambda t^{\mathrm{s}}}{N}\right) w^{\mathrm{id}}. \tag{4.8c}$$

Note that Prop 22 can be viewed as a special case of the above framework with $w^{\mathrm{en}} = w^{\mathrm{pr}}$ and $w^{\mathrm{id}} = 0$. Let us set the same engaged time wage as in Prop 22, $w^{\mathrm{en}} = \$21$ and observe how the effective hourly driver wage w varies with w^{id}. As seen in Figure 4.2a, the wage constraint (4.8c) becomes active at $w^{\mathrm{id}} = \$12.9$, and the effective hourly wage w increases with w^{id} from this point. At $w^{\mathrm{id}} = \$24.8$, the driver wage is equal to the effective minimum wage $w^{\mathrm{min}} = \$22.46$. Since $w^{\mathrm{en}} = \$21$ based on Prop 22, this implies that drivers would need to be paid a higher wage during idle time compared with when they are engaged in a trip. Note that a compelling argument in support of an engaged time minimum wage is that it deters drivers from gaming the system by availing of the minimum wage without actually providing rides. A higher idle time wage further incentivizes such adverse driver behavior.

Such a situation can be avoided by making sure that the guaranteed engaged time wage is sufficiently higher than that during idle time. Moreover, since platforms can monitor driver behavior, they can make the idle time wage payable only to drivers that are actively seeking rides and have a high trip acceptance rate. Suppose that setting the engaged time wage to be at least twice the idle time wage would deter drivers from gaming the system. As illustrated in Figure 4.2b, an engaged time wage $w^{\mathrm{en}} = \$28.6$ and idle time wage $w^{\mathrm{id}} = \$14.2$ is sufficient to ensure that drivers make the effective minimum wage while ensuring that driver incentives are aligned with that of the platform. As seen from Table 4.1, the utilization rate under such a wage regulation is 15% higher compared with a purely engaged time based wage floor. Moreover, trip fares are 9.5% lower and platform profit is 14.5% higher under this setting. Thus, incorporating an idle time wage is uniformly better than a purely engaged time wage in terms of outcomes for all participants in the ride-hailing ecosystem.



(a) $w^{\text{en}} = \$21$ (b) $w^{\text{en}} = \$28.6$

Figure 4.2: Effective hourly wages under varying idle time wage.

Wage Regulation	Trip Fare ($)	Driver Utilization Rate	Number of Trips	Platform Profit ($)
No Regulation	12.44	0.60	7,156	38,419
Engaged Time	14.53	0.50	7,366	29,870
Engaged + Idle Time	13.15	0.58	8,470	34,519

Table 4.1: Ride-hailing ecosystem outcomes under no regulation, a minimum wage for only engaged time, and minimum wage for both engaged and idle driver time. In the latter two settings, the minimum wage is chosen such that drivers earn the effective minimum wage $w^{\text{min}} = \$17.2$.

Chapter 5

Impact of Labor Regulations: Assembly Bill 5

In this chapter, we present our analysis of the impact of Assembly Bill 5 on the ride-hailing ecosystem.

5.1 Introduction

Recall from our overview in Chapter 2 that concerns over wages for ride-hailing drivers motivated the introduction of Assembly Bill 5 by the state of California. This bill changed the worker classification status of drivers from independent contractors to employees. The classification of ride-hailing drivers as employees has two significant implications:

(a) The Wage Effect: As employees, ride-hailing drivers would have the right to negotiate employment conditions through a labor union. They would be protected by pre-existing laws that establish minimum wage, overtime compensation, sick leave, and other employee safety nets. Since drivers currently make less than minimum wage on average, platforms would lose a substantial fraction of their profits due to higher driver wages.

(b) The Flexibility Effect: Platforms argue that classifying drivers as employees would force them to impose a traditional shift-based work arrangement [50]. Having a pool of full time drivers reduces the ability of Uber and Lyft to match driver supply to customer demand on a fine temporal scale. Passengers face higher fares and longer wait times due to resulting supply-demand imbalances. Drivers lose the flexibility of choosing when and where they can work. The platform's profit would decrease since it facilitates fewer trips, and it engages more drivers than necessary for certain hours of the day.

5.2 Summary of Results

We summarize the results of our analysis below:

- We propose a novel approach for solving the platform's temporally coupled non-convex pricing problem under AB5. We show that tight bounds on the platform's profit can be derived via the dual formulation of the original problem. The tightness of these bounds suggests that our approach can be of independent interest as a framework to solve problems in similar contexts (eg., food/grocery delivery) where platforms wish to maximize profit under shift-based employment constraints.
- We isolate the impact of the two main consequences of AB5 on platform profit: (i) wage effect, and (ii) flexibility effect. We show that even in the extreme case of drivers working in pre-scheduled shifts, the wage effect dominates in terms of reducing platform profit. We find that the increased driver wage under AB5 reduces platform profit by 18%, whereas the loss of driver supply flexibility results in a profit reduction ranging between 1 - 10% depending on how shifts are designed. We also find that passenger fares are minimally affected by AB5. However, the shift-based work arrangement would preclude many part-time drivers from participation in the ride-hailing ecosystem.
- We propose a hybrid driver supply model consisting of a mixture of employees and part-time drivers that would mitigate the adverse effects of reduced flexibility while safeguarding driver income under AB5.

5.3 Analysis under Assembly Bill 5

We now analyze the effects of reclassifying drivers as employees as mandated by Assembly Bill 5 (AB5). As discussed in Chapter 2, this would result in two main changes to the platform-driver relationship: (a) wage effect: the platform profit decreases because drivers are guaranteed an hourly minimum wage and benefits, (b) flexibility effect: the platform profit decreases since shift-based driver pools make it difficult to balance driver supply and passenger demand.

Our analysis is based on a time-varying version of the equilibrium model developed in Section 3.2. We consider platform operations over one day. Let i denote the hour index starting at 12 A.M. For notational convenience, time-varying variables and model parameters without a subscript represent 24 dimensional vectors with each dimension corresponding to a specific hour of day. For instance, $\lambda = [\lambda_1, \ldots, \lambda_{24}] \in \mathbb{R}^{24}_+$ represents time-varying passenger

demand over a day. The platform's optimal pricing problem can be expressed as

$$\Pi^{\text{nr}} = \max_{p\geq 0, w\geq 0} \sum_{i=1}^{24} (\lambda_i p_i - N_i w_i) \tag{5.1}$$

$$\text{s.t. } \lambda_i = \lambda_i^0 \left(1 - C_i\left(p_i + \frac{\alpha_i M_i}{\sqrt{N_i - \lambda_i t_i^{\text{s}}}}\right)\right), \ \forall i, \tag{5.1a}$$

$$N_i = N_i^0 S_i(w_i), \ \forall i. \tag{5.1b}$$

We refer to the formulation in 5.1 as the *no regulation* setting.

Remark 5.1 *The driver supply in hour i may depend on wages in other hours. Empirical studies provide evidence for both positive and negative inter-temporal elasticity of labor supply with respect to hourly wages [59, 60, 61]. Moreover, both effects have been found in the case of ride-hailing drivers due to the driver pool made up of both part-time and full-time drivers [62]. Due to the limited availability of data, it is challenging to estimate inter-temporal dependencies between hourly wages and driver supply. We therefore assume that the driver supply in any hour depends only on the wage offered in that hour.*

5.3.1 Modeling Impact of Shift-Based Employment

We now model changes in the platform's operation under shift-based employment. Suppose the platform offers J shifts in which the drivers can work. Let S_j denote the number of drivers working in the j^{th} shift and $S = [S_1, \ldots, S_J]^\intercal \in \mathbb{R}_+^J$. The platform decides the shift pattern, i.e., the set of shifts available to drivers. This can be described by the shift pattern matrix $A \in \mathbb{R}^{24\times J}$ such that

$$A_{ij} = \begin{cases} 1, & \text{if shift } j \text{ contains hour } i, \\ 0, & \text{otherwise.} \end{cases}$$

For example, for three non-overlapping traditional eight hour shifts, the shift pattern matrix takes the form,

$$A = \begin{bmatrix} \mathbf{1}_{8\times 1} & \mathbf{0}_{8\times 1} & \mathbf{0}_{8\times 1} \\ \mathbf{0}_{8\times 1} & \mathbf{1}_{8\times 1} & \mathbf{0}_{8\times 1} \\ \mathbf{0}_{8\times 1} & \mathbf{0}_{8\times 1} & \mathbf{1}_{8\times 1} \end{bmatrix}. \tag{5.2}$$

Let $N = [N_1, \ldots, N_{24}]^\intercal \in \mathbb{R}_+^{24}$ represent the time-varying hourly driver supply. This is related to the number of drivers in shifts by

$$N = AS. \tag{5.3}$$

We assume that the platform does not face any driver supply constraint as long as it pays its drivers a shift wage equivalent to an hourly minimum wage, i.e., $w_j^{\text{sh}} = w^{\min} \sum_{i=1}^{24} A_{ij}$.

Remark 5.2 *Due to drivers working in shifts and being guaranteed a minimum wage under AB5, platforms would be offering a mode of employment that is similar to other minimum wage jobs. The city of SF currently has 142,000 workers making minimum wage [63]. Since the number of drivers on the streets of San Francisco does not exceed 6000 even during peak hours [32], we assume that the platform does not face any driver supply constraint for full-time drivers as long as it pays its drivers a shift wage equivalent to an hourly minimum wage.*

5.3.2 Platform's Pricing Problem under AB5

Under AB5, the platform aims to maximize its profit while having its drivers work in pre-scheduled shifts and paying a shift wage equivalent to an hourly minimum wage. The platform's pricing problem in this setting is

$$\Pi^{\text{ab}} = \max_{p\geq 0, N\geq 0, S} \sum_{i=1}^{24} \lambda_i p_i - \sum_{j=1}^{J} S_j w_j^{\text{sh}} \tag{5.4}$$

$$\text{s.t. } \lambda_i = \lambda_i^0 \left(1 - C_i\left(p_i + \frac{\alpha_i M_i}{\sqrt{N_i - \lambda_i t_i^{\text{s}}}}\right)\right), \ \forall i, \tag{5.4a}$$

$$N = AS, \qquad \textit{Shift Constraint} \tag{5.4b}$$

$$S \geq 0. \tag{5.4c}$$

We refer to 5.4 as the *AB5* setting.

Remark 5.3 *Labor protection advocates argue that AB5 does not explicitly require platforms to hire full-time drivers and make them work in shifts. However, platforms argue that if drivers are to be paid an hourly minimum wage regardless of the number of trips serviced, they would need to exert greater control over when and where their drivers could operate [50]. As a case in point, the introduction of NYC's minimum pay standard led platforms to force drivers to work only in pre-scheduled time slots [64]. A similar situation was reported in Geneva after Uber Eats couriers were required to be classified as employees [57]. Thus, we consider a setting in which all drivers work full-time for the platform in pre-scheduled shifts. As we will find in Section 5.4.2, the platform can mitigate the flexibility effect to a large extent even in this extreme case in which all drivers work in shifts. We also discuss an alternative setting in Section 5.5 in which a mixture of part-time and full-time drivers service trips.*

5.3.3 Wage and Flexibility Effect

Recall that the difference between the platform's profit under no regulation and AB5 is explained by two factors: (a) the wage effect, and (b) the flexibility effect. We would like to decompose the impacts of these two effects on the platform business model. For this, we

consider an intermediate setting that isolates the wage effect. In this setting, drivers are paid the minimum wage on an hourly basis and are not constrained to work in shifts. Thus, the platform's hourly pricing decisions decouple.

These pricing decisions are solved by

$$\Pi^{\min} = \max_{p\geq 0, N\geq 0} \sum_{i=1}^{24} (\lambda_i p_i - N_i w^{\min}) \tag{5.5}$$

$$\text{s.t. } \lambda_i = \lambda_i^0 \left(1 - C_i\left(p_i + \frac{\alpha_i M_i}{\sqrt{N_i - \lambda_i t_i^{\mathrm{s}}}}\right)\right), \ \forall i. \tag{5.5a}$$

We refer to 5.5 as the *minimum wage* setting. Since the minimum wage and AB5 settings differ only in the shift constraints, the difference between the two captures the flexibility effect. Similarly, the difference between the no regulation and minimum wage setting isolates the wage effect. We now formalize these ideas.

The Wage effect (W) is defined as the difference in platform profit between the no regulation (5.1) and minimum wage (5.5) setting as a fraction of the profit in the no regulation setting, i.e.,

$$\mathrm{W} = \frac{\Pi^{\mathrm{nr}} - \Pi^{\min}}{\Pi^{\mathrm{nr}}}. \tag{5.6}$$

The Flexibility effect (F) is defined as the difference between the platform's profit under the minimum wage (5.5) and AB5 (5.4) settings as a fraction of the profit in the no regulation setting, i.e.,

$$\mathrm{F} = \frac{\Pi^{\min} - \Pi^{\mathrm{ab}}}{\Pi^{\mathrm{nr}}}. \tag{5.7}$$

5.4 Model Calibration and Simulation Study

We calibrate our model using time-varying data for trips, drivers, fares and wages. We use data for number of trips and drivers operating in SF in each hour from [32]. Since time-varying data for fares and wages is unavailable, we estimate them from a combination of sources. We estimate trip fares by multiplying the trip fare $\hat{p}$ derived in Section 3.3 with the fare variation across time from [65]. Similarly, driver wages are estimated by multiplying the wage $\hat{w}$ estimated in Section 3.3 with the wage variation across time based on [66]. As in the static case in Section 3.3, we consider linear CDFs for the alternate mode costs $C_i(\cdot)$ and driver reservation wages $S_i(\cdot)$ in hour i. These time-varying data are then used to estimate time-varying model parameters $\{\lambda_i^0, e_i^{\mathrm{P}}, N_i^0 e_i^{\mathrm{D}}, \alpha_i M_i\}$ using the same approach as in the static case.

We assume that as employees, all drivers work full-time in 8 hour shifts. Since fixed expenses must be taken into account for full-time drivers, we assume a driver expense of

\$0.58/mile. This results in an hourly driver expense $D = \$7.27$, hourly minimum wage $w^{\min} = \$24.47$ and an equivalent shift wage $w_j^{\text{sh}} = \$195.76$. In order to understand the extent to which shift design influences the flexibility effect, we consider two extremes in terms of shift design: (i) Three shifts, i.e., 8 A.M. - 4 P.M., 4 P.M. - 12 A.M., 12 A.M. - 8 P.M., and (ii) 24 staggered shifts separated by an hour , i.e., 8 A.M. - 4 P.M., 9 A.M. - 5 P.M., ..., 7 A.M. - 3 P.M. We refer to the former as the *traditional* shift pattern, and the latter as the *staggered* shift pattern.

5.4.1 Estimating the Flexibility Effect

Even though the platform's pricing problems in the no regulation (5.1) and minimum wage setting (5.5) are non-convex, they can be solved efficiently since they are hourly separable, i.e., each hourly problem can be independently solved by grid search as in the case of (3.11). On the other hand, the platform's pricing problem in the AB5 setting (5.4) is temporally coupled, i.e., it cannot be decomposed into independent hourly problems. Due to the large number of decision variables, finding the optimum using grid search is intractable. We now discuss how upper and lower bounds can be computed efficiently for the platform's profit under AB5.

A lower bound on the platform's profit can be readily obtained from a feasible point or a local optimal point computed using a standard nonlinear programming solver. Substituting this lower bound in (5.7) provides us an upper bound $\overline{F}$ for the flexibility effect. Since (5.4) is a maximization problem, we can get an upper bound for the platform's profit by solving its dual problem. The lagrangian for the platform's pricing problem (5.4) under constraints (5.4b) and (5.4c) is given by

$$\mathcal{L}(p, \lambda, N, \eta, \beta) = p^\intercal \lambda - w^{\text{sh}\intercal} S - \eta^\intercal (N - AS) + \beta^\intercal S, \tag{5.8}$$

where $\eta \in \mathbb{R}^{24}$ and $\beta \in \mathbb{R}^J_+$ denote the lagrange multipliers associated with the shift constraints (5.4b) and (5.4c) respectively. The primal problem can be equivalently expressed as

$$\max_{\lambda \geq 0, N \geq 0, S} \; \min_{\beta \geq 0, \eta} \; p^\intercal \lambda - \eta^\intercal N + (A^\intercal \eta + \beta - w^{\text{sh}})^\intercal S \tag{5.9}$$

$$\text{s.t. } \lambda_i = \lambda_i^0 \left(1 - e_i^{\text{p}} \left(p_i + \frac{\alpha_i M_i}{\sqrt{N_i - \lambda_i t_i^{\text{s}}}} \right) \right), \; \forall i. \tag{5.9a}$$

Since we dualized constraint (5.4c), it follows that $A^\intercal \eta + \beta - w^{\text{sh}} = 0$. Combined with $\beta \geq 0$, we have $A^\intercal \eta \leq w^{\text{sh}}$. This results in the following dual formulation,

$$\min_{A^\intercal \eta \leq w^{\text{sh}}} \; \max_{\lambda \geq 0, N \geq 0} \; \sum_{i=1}^{24} (\lambda_i p_i - N_i \eta_i) \tag{5.10}$$

$$\text{s.t. } \lambda_i = \lambda_i^0 \left(1 - e_i^{\text{p}} \left(p_i + \frac{\alpha_i M_i}{\sqrt{N_i - \lambda_i t_i^{\text{s}}}} \right) \right), \; \forall i. \tag{5.10a}$$

Given lagrange multipliers η_i, the inner maximization problem resembles the minimum wage setting (5.5) with the fixed hourly wage $w^{\min}$ replaced by the time-varying η_i. As in the minimum wage setting, it can be decomposed into hourly problems and solved optimally using grid search. We can equivalently express the dual problem as

$$\min_{A^\intercal \eta \leq w^{\text{sh}}} \quad G(\eta), \tag{5.11}$$

where $G(\eta)$ is the lagrangian dual function defined as the value of the inner maximization problem above, i.e.,

$$G(\eta) = \sum_{i=1}^{24} g_i(\eta_i), \tag{5.12a}$$

$$g_i(\eta_i) = \max_{\lambda_i \geq 0, N_i \geq 0} \lambda_i p_i(\lambda_i, N_i) - N_i \eta_i, \ \forall i, \tag{5.12b}$$

$$p_i(\lambda_i, N_i) = \frac{1}{e_i^{\text{p}}}\left(1 - \frac{\lambda_i}{\lambda_i^0}\right) - \frac{\alpha_i M_i}{\sqrt{N_i - \lambda_i t_i^{\text{s}}}}, \ \forall i. \tag{5.12c}$$

Solving the dual problem (5.11) is non-trivial since $G(\eta)$ cannot be expressed analytically and is non-convex in η. However, notice that any dual feasible point $\tilde{\eta}$ automatically provides an upper bound $G(\tilde{\eta})$ for platform profit. Finding dual feasible points that minimize the dual objective further will provide better upper bounds for platform profit and as a consequence, better lower bounds for the flexibility effect. We now present an algorithm to find better dual feasible points using projected gradient descent.

First, we start with an initial dual feasible point η^0 (for eg., lagrange multipliers at the local optimum found by nonlinear programming solvers). Note that by weak duality, $\Pi^{\text{ab}} \leq G(\eta^0)$. Since an analytical expression for $G(\eta)$ is unavailable, we need to rely on numerical approximations to estimate gradients. Thus, we update the lagrange multipliers η^k at the k^{th} step as follows (with small enough $\delta^k, \epsilon^k > 0$ that ensures descent),

$$\hat{\eta}_i^{k+1} = \eta_i^k - \epsilon_i^k \left(\frac{g_i(\eta_i^k + \delta_i^k) - g_i(\eta_i^k - \delta_i^k)}{2\delta_i^k} \right), \forall i, \tag{5.13a}$$

$$\eta^{k+1} = \text{Proj}(\hat{\eta}^{k+1}), \tag{5.13b}$$

where $\text{Proj}(\hat{\eta}^{k+1})$ denotes the projection of $\hat{\eta}^{k+1}$ onto the half space $A^\intercal \eta \leq w^{\text{sh}}$ under the Euclidean distance metric, i.e.,

$$\text{Proj}(\hat{\eta}^{k+1}) = \min_{A^\intercal \eta \leq w^{\text{sh}}} \|\eta - \hat{\eta}^{k+1}\|_2. \tag{5.14}$$

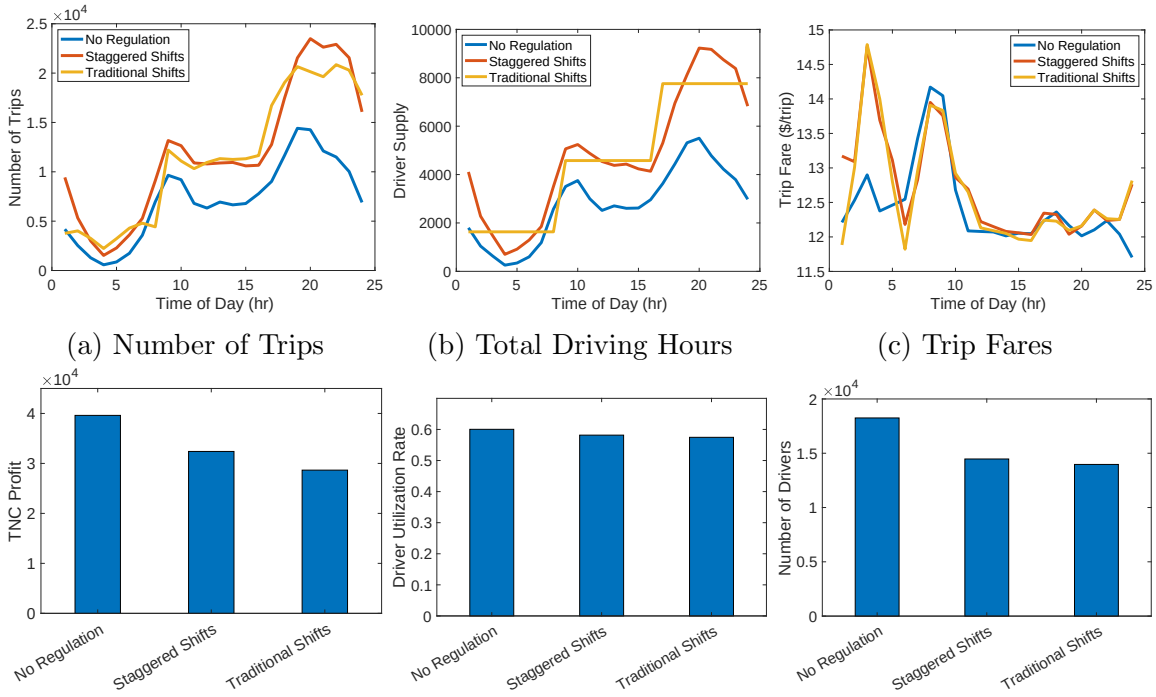
Proposition 5.1 *Suppose the projected gradient descent scheme (5.13) is run for K steps. Then, $G(\eta^K)$ is an improved upper bound for platform profit under AB5 and consequently, the flexibility effect is lower bounded by*

$$\underline{F} = \frac{\Pi^{\min} - G(\eta^K)}{\Pi^{\text{nr}}}. \tag{5.15}$$

Proof *Since* η^K *is a dual feasible point, it follows from weak duality that* $G(\eta^K) \geq \Pi^{\text{ab}}$. *Moreover, the projected gradient descent scheme ensures that* $\Pi^{\text{ab}} \leq G(\eta^K) \leq G(\eta^0)$, *i.e., we have an improved upper bound for platform profit. Substituting* $G(\eta^K)$ *in place of* Π^{ab} *in* (5.7) *results in an improved lower bound for the flexibility effect.*

As we shall see in the next section, the above approach tightly bounds the flexibility effect for both traditional and staggered shift patterns.

5.4.2 Results and Discussion



(a) Number of Trips (b) Total Driving Hours (c) Trip Fares

(d) Platform Profit (e) Driver Utilization Rate (f) Number of Unique Drivers

Figure 5.1: Ride-hailing ecosystem outcomes under no regulation, staggered and traditional shifts.

We find that the flexibility effect $F \in [8.9\%, 9.9\%]$ in the case of traditional shifts, while $F \in [0.1\%, 0.5\%]$ for staggered shifts. On the other hand, the wage effect contributes to the platform losing 17.8% of its original profit. While reduction in platform profit is an inevitable consequence of increased driver wages, the platform can mitigate the flexibility effect to a large extent by suitable shift design (a sensitivity analysis can be found in Appendix B).

Under a shift-based work arrangement with only full-time drivers, drivers who benefit from the part-time, flexible nature of the prevailing platform work arrangement will be excluded. In order to estimate the loss of work opportunities under a shift-based work

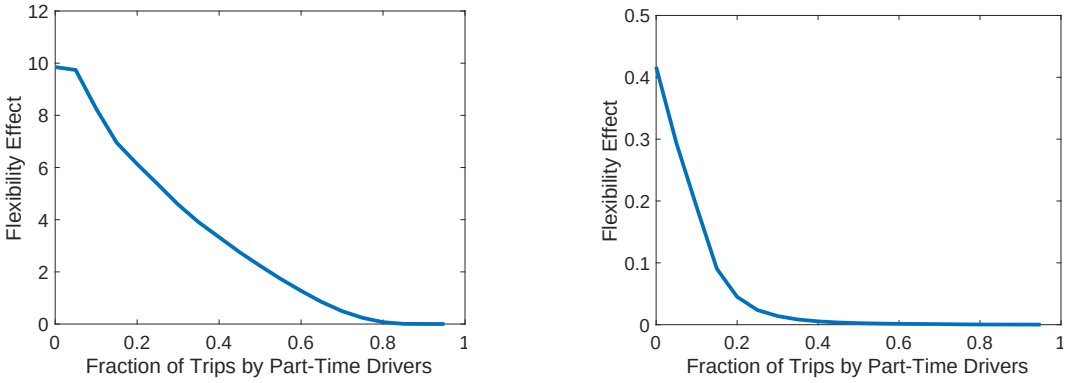
Driver type (hrs/week)	Percentage of trips served
<10	10%
10 - 25	32%
25 - 40	33%
>40	25%

Table 5.1: Distribution of platform trips across driver types.

arrangement, we need to first estimate the total number of unique drivers operating in SF in the no regulation setting[1]. Note that $\sum_i N_i$ is the total driving time supplied by all drivers on the platform across all hours of the day. Table 5.1 provides a distribution of trips across drivers classified by the number of hours they work for the platform. This is based on Uber's data on its drivers in California [51]. Using the total number of driving hours supplied ($\sum_i N_i$) and the distribution of trips across driver types in Table 1, we estimate the total number of unique drivers of each type under no regulation. As illustrated by Figure 5.1f, we find that the total number of drivers would reduce by about one-fourth under both traditional and staggered shifts. This is expected since the part-time drivers who constitute 74% of all ride-hailing drivers, would be replaced by full-time drivers. A more drastic reduction in the number of Uber Eats couriers was observed in Geneva, where only 23% of the prevailing couriers were offered work on the platform post-employee classification [57].

5.5 An Alternative Proposal

As demonstrated in Section 5.3, achieving favorable outcomes for some participants in the ride-hailing ecosystem comes at the expense of worse outcomes for others. While AB5 would ensure that drivers make minimum wage without increasing trip fares, platforms would lose a substantial fraction of their profits due to the wage and flexibility effects. It could also preclude platforms from hiring drivers who benefit from driving part-time and enjoy the flexibility of working hours. This motivates the formulation of an alternative proposal that could improve outcomes for drivers and passengers without substantially affecting the sustainability of the platform business model. We present one such alternative to achieve such a compromise.



(a) Traditional Shifts (b) Staggered Shifts

Figure 5.2: Flexibility effect (%) under varying fraction of trips served by part-time drivers (γ).

Allowing Part-Time Drivers under AB5

In Section 5.3, we assumed that all drivers would be classified as employees under AB5 and work full-time shifts for platforms. However, there is a significant fraction of drivers who benefit from the opportunity to drive fewer hours of their own choosing. As discussed in Section 5.4.2, part-time drivers make up 74% of the driver pool and provide 42% of all platform trips in California. Such drivers can help mitigate the loss of flexibility in driver supply owing to platforms moving towards a shift-based work arrangement under AB5.

We consider a setting in which part-time drivers can serve a maximum fraction γ of all platform trips in each hour. Let N_i^{p} and N_i^{f} denote the number of part-time and full-time drivers respectively in hour i such that the total number of drivers $N_i = N_i^{\mathrm{p}} + N_i^{\mathrm{f}}$. Then, the number of part-time drivers hired by the platform in hour i must satisfy

$$N_i^{\mathrm{p}} \leq \frac{\gamma}{1-\gamma} N_i^{\mathrm{f}}. \tag{5.16}$$

As in Section 5.3, we assume that full-time drivers work in shifts, i.e., $N^{\mathrm{f}} = AS$. Since we would like to isolate the effect of part-time drivers on the flexibility effect, we assume that part-time drivers earn the same wage per hour as full-time drivers. Let w_i^{p} denote this equivalent hourly wage corresponding to the full-time shift wage w^{sh}. Thus, the platform

[1]The data for number of trips and drivers operating in each hour of the day does not directly convey how many unique drivers provide trips during the course of the day. For instance, $N_1 = 10$ and $N_2 = 10$ could result from both (i) 20 unique drivers with 10 drivers working in hour 1 and the rest in hour 2, and (ii) 10 unique drivers working both hours.

Type of Regulation	Trip Fare ($)	Utilization Rate	No. of full-time drivers	No. of part-time drivers	Platform Profit ($/hr)
No regulation	12.38	0.60	5,474	12,777	39,612
Traditional shifts	12.45	0.57	13,957	0	28,661
Staggered shifts	12.50	0.58	14,456	0	32,397
Traditional shifts + part-time drivers	12.49	0.58	11,299	11,232	31,286
Staggered shifts + part-time drivers	12.51	0.58	13,656	3,045	32,559

Table 5.2: Ride-hailing ecosystem outcomes under no regulation, all drivers working full-time in shifts, and a mixture of full-time and part-time drivers. For the hybrid driver supply settings, we consider $\gamma = 0.4$, i.e, at most 40% of all trips are serviced by part-time drivers.

solves the following pricing problem under this hybrid driver supply setting,

$$\Pi^{\text{ab}} = \max_{p\geq 0, N^{\text{p}}\geq 0, S} \sum_{i=1}^{24}(\lambda_i p_i - N_i^{\text{p}} w_i^{\text{p}}) - \sum_{j=1}^{J} S_j w_j^{\text{sh}} \tag{5.17}$$

$$\text{s.t. } \lambda_i = \lambda_i^0 \left(1 - e_i^{\text{p}}\left(p_i + \frac{\alpha_i M_i}{\sqrt{(N_i^{\text{p}} + N_i^{\text{f}}) - \lambda_i t_i^{\text{s}}}}\right)\right), \ \forall i, \tag{5.17a}$$

$$N_i^{\text{p}} \leq \frac{\gamma}{1-\gamma} N_i^{\text{f}}, \ \forall i, \qquad \textit{Cap on part-time drivers} \tag{5.17b}$$

$$N^{\text{f}} = AS, \tag{5.17c}$$

$$S \geq 0. \tag{5.17d}$$

Intuitively, as the proportion of part-time drivers γ increases, the flexibility effect decreases. Figure 5.2 illustrates how the flexibility effect varies with increase in γ for traditional and staggered shift patterns. We find that allowing a small fraction of part-time drivers under AB5 can facilitate a substantial reduction in the flexibility effect in both cases. For instance, allowing 40% of all trips to be serviced by part-time drivers would reduce the flexibility effect from 9.9% to 3.2% under traditional shifts. Doing the same in the case of staggered shifts is sufficient to completely eliminate the flexibility effect. Thus, a hybrid model of driver supply with only a small fraction of part-time drivers is sufficient to mitigate the decrease in platform profit due to the shift arrangement-induced loss of flexibility. As can be seen from Table 5.2, it would also ensure that prevailing part-time drivers who enjoy the flexible work schedule offered by platforms can continue to participate in this ecosystem. Thus, a hybrid driver supply under AB5 would ensure that drivers who work full-time for the platform and serve a majority of the trips can enjoy employee protections while at the same time, the flexibility of driver supply that undergirds the platform business model is not mitigated.

5.6 Concluding Remarks

Although we present results based on data from SF, our generic framework can be adapted to analyze the impact of similar regulations currently under consideration in other cities. We now briefly discuss how such regulations can be interpreted using our model and the need for data disclosure by platforms for a more detailed analysis. We also highlight key unaddressed issues which merit careful consideration for a comprehensive understanding of the impacts of the above regulations.

5.6.1 Pay Standard in New York City and Seattle

As discussed in Chapter 2, a minimum pay standard has been instituted for ride-hailing drivers in New York City (NYC) and Seattle. Much like Prop 22, these pay standards put in place a floor on driver earnings during each trip. However, it crucially differs from Prop 22 in how it compensates idle driver time. The NYCTLC imposed minimum pay per trip is as follows:

$$\text{Driver pay per trip} = \left(\frac{\$0.631 \times \text{Trip Miles} + \$0.287 \times \text{Trip Minutes}}{\text{Platform Utilization Rate}}\right) + \text{Shared Ride Bonus}, \tag{5.18}$$

where the time and distance based payments are set such that drivers make the independent contractor equivalent minimum wage of \$17.22/hr. Notice that the trip based payment is normalized by the platform's utilization rate. The wage multiplier in this case is a function of the platform's utilization rate ρ, unlike the fixed 120% premium in the case of Prop 22. Therefore, the platform cannot lower driver wages by reducing its utilization rate as in the case of a fixed wage multiplier. Let $w^{\min}_{\text{NYC}} = 0.631 \times \text{Trip Miles} + \$0.287 \times \text{Trip Minutes}$. Viewed from the lens of our model, the constraint on the effective hourly wage is given by

$$w \geq \rho\left(\frac{1}{\rho}\right) w^{\min}_{\text{NYC}}, \tag{5.19}$$

$$= w^{\min}_{\text{NYC}}. \tag{5.20}$$

Thus, the above pay standard places a lower bound on the effective hourly wage earned by drivers.

5.6.2 Data Disclosure

Trip level data such as pick-up and drop-off locations, driver pay and passenger fares can provide valuable insights into the impact that platforms have on the transportation systems within which they operate. Over the years, there have been calls from regulatory authorities in several cities for platform data to be made public. Platforms argue that disclosing trip level data would reduce their competitive edge [67]. Moreover, they claim that passenger details can be inferred even from anonymized trip data. While the latter concern can be

handled by disclosing only approximate locations such as the nearest intersection, the former will be mitigated by the fact that all platforms are required to disclose data, thus creating a level playing field. Nevertheless, Uber and Lyft have lobbied intensely to keep their data confidential.

In the state of California, platforms are required to submit data to the California Public Utilities Commission (CPUC). However, CPUC has refused to share this data with the San Francisco County Transportation Authority (SFCTA) [68]. This compelled SFCTA to conduct its own study to estimate the impact of platforms in its jurisdiction [32]. The insights gained from such estimates can only be as good as the accuracy of the estimates. In the absence of accurate, publicly available data, there is a lack of clarity in most jurisdictions about fundamental questions such as whether drivers make minimum wage, if platforms have increased congestion or whether platforms provide equitable access to their service. At present, NYC, Chicago and Seattle are the only cities that have made platform data public. Notably, this data has been used to inform minimum pay standards for drivers in NYC [15, 16]. NYC's pay standard has led to higher driver wages, lower waiting times and higher utilization rates at the expense of only a slight increase in passenger fares. While it is too early to estimate the impact of Seattle's pay standard, a similar result is expected. Thus, requiring platforms to disclose trip level data can play a significant role in informing regulations that improve outcomes for all participants in the ride-hailing ecosystem.

5.6.3 Spatial Coverage and Equity

A crucial consideration in the analysis of platform labor regulations is how they impact service coverage and transportation equity across different neighborhoods in a city. Areas with sparse demand such as low income neighborhoods and rural communities are particularly vulnerable to regulatory changes. This is because balancing time-varying trip demand and driver supply is more challenging in such areas. Recent studies have found higher fares and waiting times in such neighborhoods, which often have limited access to alternative modes of public transit [69, 70]. Mandating a high utilization rate can alter the incentives for platforms to operate in low demand density neighborhoods and as a result, significantly influence the price and accessibility of their services.

The potential impact of regulations on transportation equity must be carefully examined so that they do not result in adverse outcomes for vulnerable communities. While our current analysis does not capture this issue, we plan to extend our analysis in future work to address it following a framework similar to [71].

5.6.4 Platform Competition

We do not explicitly model competition among platforms in our analysis. An important issue that remains unaddressed is how regulations impact the competitive advantage of one platform over others. For instance, there are claims that the use of utilization rate as an instrument to regulate drivers' wage may give an edge to larger platforms in one area since

they can naturally achieve a higher utilization rate because of their larger market share. This issue was the subject of a recent lawsuit filed by Lyft and Juno in NYC claiming that the NYC minimum pay standard discussed in Section 5.6.1 placed them at a disadvantage compared to their significantly larger rival Uber [72]. While we do not evaluate such claims in our analysis, we believe that concerns about regulations increasing barriers to entry need to be examined carefully to ensure that the ride-hailing ecosystem remains competitive in the long term.

5.6.5 Post-COVID Economy

As a result of the COVID-19 pandemic, many fundamental factors affecting the ride-hailing ecosystem have significantly changed. We do not explicitly consider such changes in our model since enough empirical data is not available yet to quantify them. Nonetheless, we briefly discuss some of these changes. Most importantly, the pandemic disturbed the delicate balance between passenger demand and driver supply. It is not yet clear whether the new equilibrium between these two sides will be similar to the one realized before the pandemic that was achieved with an expectation of much higher driver earnings.

Since the pandemic drastically reduced passenger demand, many drivers were forced to seek alternative employment opportunities. A significant fraction of erstwhile ride-hailing drivers moved to working for delivery platforms that saw a considerable increase in demand. Moreover, the pandemic has led to a shift in consumer behavior that might have a lasting effect. For instance, passenger preference for shared ride services such as Uber Pool is expected to have significantly lowered because of the increased risk of infection. There is recent news that platforms such as Uber and Lyft are facing a driver shortage. This suggests that a significant proportion of drivers who worked for these platforms pre-pandemic prefer continuing the alternative jobs they were forced to take up during the pandemic [73]. Moreover, minimum wage for alternative job opportunities (e.g. Amazon warehouse workers) that drivers may consider has increased to $15/hr. This may significantly impact their reservation wages and as a result, willingness to work as ride-hailing drivers at current wages.

The enduring effects of the COVID-19 pandemic on the ride-hailing ecosystem remain to be seen. Moving forward, it is necessary to account for the changes brought about by the pandemic in any analysis of the impact of labor regulations.

Chapter 6

Value of Pooling in Last-Mile Delivery

In previous chapters, we have analyzed the economics of ride-hailing. We now turn our attention to other popular transportation network services such UberEats for food, and Instacart for groceries. The last-mile delivery problem is critical for these services. As discussed in Chapter 2, the high cost of delivery is a fundamental roadblock to the profitability of these services. In this chapter, we discuss the pivotal role of order pooling in governing the economic viability of delivery services.

6.1 Introduction

The Challenging Economics of Last-Mile Delivery

Despite their immense popularity, most last-mile delivery services are not profitable [74, 75]. Instacart, a major grocery delivery platform, slashed its valuation by 67%, from a high of $39B to $13B in 2022 citing challenging market conditions [44]. DoorDash, the largest food delivery service in the US by market share, posted a $1.3B loss in 2022 [76]. The high cost of delivery is widely recognized as a major obstacle to profitability [75, 77]. Even though millions of customers are accustomed to using such services, only a small fraction among them are willing to shoulder the full cost of delivery [25, 75]. To analyze the long-term viability of such services, it is crucial to understand what factors cause high delivery costs.

Rise of On-demand Delivery

The past decade has witnessed the rise of *on-demand* delivery services [78], offering order fulfillment within a few hours of receiving the request. This is in contrast to the predominant paradigm of *scheduled* delivery wherein customers chose a time window for delivery a few days in advance. Delivery companies offer on-demand services differentiated by the *delivery time guarantee*, i.e., the time taken to fulfill orders. For instance, Instacart offers 1, 2, and 4 hour delivery options and differentially prices these services. There has been a recent emergence of rapid grocery delivery platforms such as Gorillas and Getir that promise order

fulfillment within 30 minutes [79]. A failure to meet the promised delivery deadlines results in disgruntled customers who are placated by discounts or rewards.

A Time-Cost Trade-off

Although customers are willing to pay more for faster deliveries [75], *the cost of delivery also increases with lower delivery times.* A key source of efficiency in last-mile delivery is the pooling together of multiple orders for delivery in the same trip [80, 81]. Lower the delivery time promised, lower the possibility of pooling orders, and as a result, higher the cost of delivery. Thus, the extent to which the value of pooling is leveraged can significantly influence the cost of delivery.

6.2 Summary of Contributions

In this chapter, we model last-mile delivery services within the unified framework introduced in Chapter 3 to analyze the value created by pooling orders together. We demonstrate how the trade-off between delivery times and the cost of delivery, mediated by the extent of pooling, dictates which delivery services will be economically viable in the long run. Our main contributions are as follows:

- We develop a queuing-based spatial model for the last-mile delivery process that captures the dependence of delivery cost on customer demand, delivery time, extent of pooling, and other region-specific parameters such as travel speed and delivery radius. To the best of our knowledge, our work is the first to model diverse last-mile delivery contexts within a unified framework and illustrate analytically how quantities of interest such as delivery cost, service rate and delivery time vary with the extent of pooling.
- We introduce the notion of *value of pooling* to characterize the efficiency gains from pooling orders that can be leveraged under a certain delivery time guarantee. We derive upper and lower bounds on achievable service rates arising from *service efficiency* and *unit economics* respectively. Using these bounds, we derive necessary conditions for a last-mile delivery service to be economically viable.
- We present a simulation study for a grocery delivery service operating in Los Angeles (LA), California. We show that a delivery time of 2 hours or less is not economically viable in this region. We also find that driver pay accounts for 90% of the delivery cost.

Our analysis suggests that rapid grocery delivery (< 1 hr) is unprofitable in all but the most dense cities in the US, unless there is a substantial increase in the customer willingness to pay for faster delivery. Since driver wages accounts for a bulk of the delivery cost, technological innovations such as delivery via automated vehicles or drones can be pivotal in making

last-mile delivery viable. This also explains why delivery platforms have lobbied heavily for regulations (eg., Proposition 22) that maintain contractor status of their drivers, which is necessary to keep wages low [48].

6.3 Pooling in Last-Mile Delivery

6.3.1 The Need for Pooling

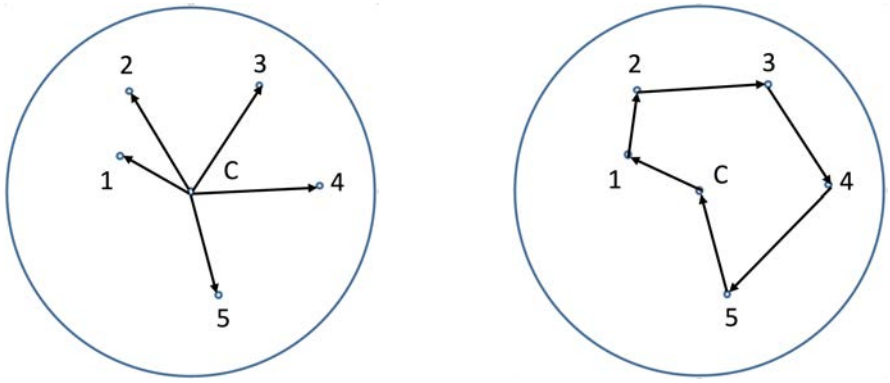
We start with a simple example. Consider a grocery store in Los Angeles (LA), California offering delivery services within a 5 mile radius. Assuming customer demand uniformly distributed over the delivery region, the average distance of a customer from the store is 3.3 miles. Thus, the total distance travelled by a driver to fulfill the order (from the store to the customer and back) is 6.6 miles on average. Since the average travel speed in LA is 22 miles/hr, it takes about 18 min to deliver the order. Considering optimistic estimates for time taken to load (10 min) and drop off orders (2 min) brings the total delivery time to 30 min per order [82]. Thus, a driver can fulfill 2 orders per hour at peak efficiency. Since the minimum wage in LA is \$16.04/hr, the driver pay per order must be at least \$8. Accounting for fuel expenses of \$0.3/mile, the total cost of delivery amounts to \$10 per order. However, surveys of grocery delivery customers have shown that most customers are unwilling to pay more than \$5 for delivery [75, 25]. Thus, delivering one order at a time would result in a loss of more than \$5 per order.

A better strategy would be to pool together a group of orders and deliver them on the same trip. Rather than deliver five orders one at a time as in Figure 6.1a, these five orders can be fulfilled in a single trip as illustrated in Figure 6.1b. Clearly, the distance travelled and time taken per order reduces. Thus, an analysis of efficiency gains due to pooling orders is crucial for understanding the viability of last-mile delivery services.

6.3.2 Connections to the Traveling Salesman Problem

A useful starting point to understand the efficiency gains from pooling is the rich literature on the traveling salesman problem. The objective here is to find the minimum length route covering a given set of points (referred to as a *tour*) such that each point is only visited once. Observe that the pooled delivery case in Figure 6.1b can be viewed as a *tour* starting from the store (at the center) covering all order locations before returning to the store. Thus, the problem of minimizing the travel distance in pooled delivery is a specific instance of the traveling salesman problem.

It is well known that the traveling salesman problem is NP-Hard [83]. A celebrated result in the vehicle routing literature states that the expected minimum tour distance for a set of points distributed uniformly at random is directly proportional to the square-root of the number of points and area over which they are distributed [84, 85]. We have reproduced this result below.



(a) One order at a time (b) Pooled delivery

Figure 6.1: Last-mile delivery with and without pooling. The store is located at the center C of the circular delivery region. The five points spread across the delivery region represent order locations.

Theorem 6.1 *Consider k points independently and uniformly distributed over a connected circular region of area A. Then, the expected length D of the shortest length tour covering these points scales as*

$$D \approx \phi\sqrt{Ak}, \ \text{as } k \to \infty, \tag{6.1}$$

where ϕ is a constant.

Interestingly, this analytical approximation is almost exact even for small values of k [85]. A key insight from this result is that pooling k orders together can provide a $\sqrt{k}$-saving in travel distance over delivering orders one at a time. Moreover, the analytical approximation (6.1) allows us to analyze delivery costs as a function of pooling size k.

6.3.3 A Time-Cost Trade-off in Pooled Delivery

Pooling of orders also leads to an increase in *delivery time*, i.e., the time between a customer placing an order request and its fulfillment. In the worst case, the driver visits $k-1$ other locations before servicing a customer, resulting in a $\sqrt{k}$-increase in delivery time. If we account for order pickup and drop-off times, we get a further increase by a factor of k. Evidently, more is not always better when it comes to pooling. The delivery time guarantee offered by the platform dictates the maximum allowed pooling batch size k, which in turn dictates the cost of delivery. The interplay between the delivery time guarantee, pooling size, and the cost of delivery governs the viability of the last-mile delivery business. In the next section, we develop a model that captures this interplay.

6.4 Model

Consider a store offering delivery services to customers within a radius R. Let N be the number of drivers fulfilling deliveries for the store. Similar to [86], we model order arrivals as a Poisson point process with parameter α, i.e., the number of order arrivals Λ arising from any subset of area A within the delivery region over a time period Δt is distributed as,

$$\Lambda(A, \Delta t) \sim \text{Poisson}(\alpha A \Delta t). \tag{6.2}$$

Let λ denote the expected number of orders per hour placed by customers in the circular delivery region, i.e., $\lambda = \alpha \pi R^2$. Let t_{pick} and t_{drop} denote the time taken to pick up an order from the store and drop off the order respectively. Let v be the average last-mile speed in the delivery region. Let $\overline{T}$ be the maximum time within which delivery is guaranteed by the store. We refer to this as the *delivery time guarantee*. Next, we discuss how the delivery process is modeled.

6.4.1 Modeling the Delivery Process

The abstraction of queues provides a useful way to model the arrival and delivery of orders. From a queuing theoretic perspective, the delivery process can be viewed as follows. Each order arrival is placed in a queue awaiting service. There are N drivers (servers) making deliveries. A service period starts once the driver starts picking up a batch of orders from the store and ends when she has returned to the store after dropping off the batch. Once a driver completes a delivery trip, she is assigned a new batch from the queue of orders. The *delivery policy* dictates the number and choice of orders to be pooled together to form a single batch. Thus, a delivery policy can be described by two attributes: (a) *batch size*: number of orders in a batch, and (b) *pooling algorithm*: which orders are chosen to be pooled together. Once a batch of orders is assigned to a driver, it is fulfilled by taking the minimum distance tour covering the order locations.

6.4.2 Sector Pooling Policies

To maximize the value created by pooling, orders that are both temporally and spatially adjacent should be served in the same trip. A common approach used for pooling orders in last-mile delivery is the *sweep algorithm* [87, 88]. Given a set of orders distributed in a circular region along with a constraint on the maximum number of orders that can be fulfilled in each trip, the sweep algorithm partitions the delivery region into sectors such that orders in the same sector are served together. Motivated by this approach, we consider the class of *sector pooling* policies wherein the delivery region is partitioned into P sectors of equal area as shown in Figure 6.2. Any delivery policy in this class can be described by two features: (a) batch size k, and (b) number of partitions P, as we discuss in the next section.

6.4.3 Delivery Time and Service Rate

Consider a sector pooling policy with P partitions and batch size k. The order arrival and delivery process under a (k, P)-sector pooling policy can be modeled as an $M/G^k/N$ queue, where G^k represents the batching and simultaneous fulfillment of k orders. Since the $M/G^k/N$ queue does not admit analytical expressions for service and queuing times, we rely on a much simpler $M/G/N$ queue approximation proposed in [89, 90] which is almost exact for large N. Based on this approximation, the arrival and pooled delivery of orders can be viewed as a two-step process. In the first step, k orders are batched together. In the second step, these k-order batches are fed into an $M/G/N$ queue where each batch is seen as a single arrival and each server works on an entire k-order batch at a time.

The service time for a k-order batch can be broken down into three terms: (a) travel time for the k-order tour, (b) time taken to pickup the batch of orders from the store, and (c) time taken to drop off orders at customer locations. Let $D^{\text{sp}}(k, P)$ be the expected travel distance for a k-order tour. Then, the service time for a batch of k orders is

$$S^{\text{sp}}(k, P) = \frac{D^{\text{sp}}(k, P)}{v} + kt_{\text{pick}} + kt_{\text{drop}}, \tag{6.3}$$

and the service rate of orders per driver is

$$\mu^{\text{sp}}(k, P) = \frac{k}{S^{\text{sp}}(k, P)} = \left(\frac{D^{\text{sp}}(k, P)}{kv} + t_{\text{fixed}} \right)^{-1}, \tag{6.4}$$

where $t_{\text{fixed}} = t_{\text{pick}} + t_{\text{drop}}$ is the component of service time that remains fixed under any delivery policy.

Recall that the delivery time for an order is the time between the customer placing the order request and the order being fulfilled. It can be decomposed into three terms: (a) *batching time*: the time taken to form a k-order batch, (b) *queuing time* for k-order batch in $M/G/N$ queue, and (c) *dispatch time*: the time from start of service to the order being fulfilled. Observe that orders in the same batch will have different expected delivery times conditioned on when they arrive and in what order they are fulfilled. Orders that arrive earlier face a longer batching time, while those that get fulfilled later face a longer dispatch time. The queuing time is the same for all orders in a batch.

Lemma 6.1 *For large N, the maximum expected delivery time for an order in a (k, P)-sector pooling batch satisfies*

$$T^{\text{sp}}(k, P) \approx \frac{(k-1)P}{\lambda} + \frac{kD^{\text{sp}}(k, P)}{(k+1)v} + kt_{\text{fixed}}. \tag{6.5}$$

Proof *See Appendix C.1.*

Note that the delivery time guarantee $\overline{T}$ must be satisfied for all orders in the batch. Thus, the sector pooling policy must satisfy $T^{\text{sp}}(k, P) \leq \overline{T}$. For ease of exposition, we refer to the maximum expected delivery time $T^{\text{sp}}(k, P)$ as the delivery time for a (k, P)-sector pooling policy.

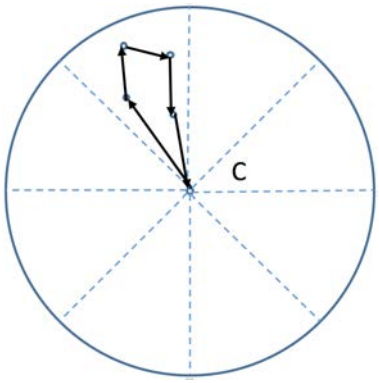


Figure 6.2: Sector Pooling with $P = 8$ partitions. The tour (in black) is an example of batch size $k = 4$.

6.5 System Efficiency in Pooled Delivery

We now analyze how the interplay between delivery time and service rate, mediated by the pooling batch size, influences system efficiency in pooled delivery.

6.5.1 A System Efficiency Upper Bound

Based on [91], the expected tour distance is closely approximated by

$$D^{\text{sp}}(k, P) \approx a(P) + b(P)\sqrt{k}, \tag{6.6}$$

where $a(\cdot)$ and $b(\cdot)$ are functions of the number of partitions P. Thus, the tour distance increases sub-linearly in k (for fixed P). Then, we have the following result.

Lemma 6.2 *For the class of sector pooling policies with a fixed number of partitions P, the service rate $\mu^{\text{sp}}(k, P)$ and delivery time $T^{\text{sp}}(k, P)$ are both increasing in batch size k.*

Proof *See Appendix C.2.*

Lemma 6.2 highlights a fundamental trade-off between the service rate and delivery time while choosing the pooling batch size k. The optimum pooling size is determined by maximizing service rate subject to the delivery time constraint. We have the following:

Theorem 6.2 *Suppose the store fulfills deliveries using a sector pooling policy with P partitions and guarantees delivery within time $\overline{T}$. Then the optimal service rate is achieved at the batch size*

$$k^*(\overline{T}, P) = \max\{k \mid T^{\text{sp}}(k, P) \leq \overline{T}\}. \tag{6.7}$$

Proof *The proof follows from the service rate $\mu^{\text{sp}}(k, P)$ being increasing in k based on Lemma 6.2.*

For orders to not be dropped, the delivery service must ensure that its service rate exceeds the order arrival rate. This imposes the following upper bound on the the order rate,

$$\lambda < N\mu_*^{\text{sp}}(\overline{T}, P), \tag{6.8}$$

where $\mu_*^{\text{sp}}(\overline{T}, P) = \mu^{\text{sp}}(k^*(\overline{T}, P), P)$ is the service rate per driver at the optimum batch size $k^*(\overline{T}, P)$. Equivalently, the optimal service rate $\mu_*^{\text{sp}}(\overline{T}, P)$ can be viewed as an upper bound for the order rate served by a single driver (i.e., λ/N). Since this result places a fundamental limit on the order rate that the last-mile delivery system can possibly serve, we refer to it as the *system efficiency* upper bound.

6.5.2 The No Pooling Policy

As a baseline, we consider a delivery policy that does not pool orders, i.e., each driver fulfills only one order between consecutive visits to the store. Each new order is placed in a queue awaiting service by an idle driver. Note that this setting can be viewed as a sector pooling policy with $(k, P) = (1, 1)$. In this case, the $M/G/N$ queue approximation for the delivery process is exact with zero batching time. Since a single order is fulfilled at a time, the expected tour distance for the no pooling policy simplifies to twice the expected distance between the store and a customer. Thus, the expected tour distance is $D^{\text{np}} = 4R/3$ as orders are distributed uniformly at random across the delivery region. Substituting this in (6.4) and (6.5), we have

$$\mu^{\text{np}} = \left(\frac{4R}{3v} + t_{\text{fixed}}\right)^{-1}, \tag{6.9}$$

$$T^{\text{np}} = t_{\text{fixed}} + \frac{2R}{3v}. \tag{6.10}$$

6.5.3 The Value of Pooling

We know from Lemma 6.1 that the delivery time guarantee places an upper bound on the batch size k, i.e., the extent to which orders can be pooled. Relaxing the delivery time guarantee increases the optimal batch size, resulting in a higher service rate. Note that the delivery time under the no pooling policy is the best that a store can guarantee. Thus, any increase in the delivery time guarantee can be viewed as a means to derive value from pooling. For a given delivery time guarantee $\overline{T}$ and number of partitions P, we define the *value of pooling* as the increase in service rate due to pooling of orders as a fraction of the service rate without pooling, i.e.,

$$V^{\text{sp}}(\overline{T}, P) = \frac{\mu_*^{\text{sp}}(\overline{T}, P) - \mu^{\text{np}}}{\mu^{\text{np}}}. \tag{6.11}$$

As discussed in Section 6.1, a number of delivery services have emerged recently promising ever-decreasing delivery times. Analyzing how the value of pooling $V^{\text{sp}}(\overline{T}, P)$ varies with $\overline{T}$ can help inform whether such services can indeed leverage the value of pooling.

6.6 The Economics of Pooled Delivery

So far, we have derived fundamental limits on the order rates that can be fulfilled by a last-mile delivery service. However, for such a business to be viable in the long run, the economics of delivery must be favorable. We now discuss how the interplay between system efficiency and unit economics governs the viability of last-mile delivery.

6.6.1 The Cost of Delivery

There are two major costs involved in last-mile delivery: (a) driver wages, and (b) travel cost (eg., gas, maintenance). Let w be the hourly wage paid to drivers. Then, the total wages paid to N drivers per hour is Nw and the wage per order is given by Nw/λ. Let c^{pm} be the cost per mile travelled in a delivery trip. For a (k, P)-sector pooling policy, the total travel distance for a batch is $D^{\text{sp}}(k, P)$. Thus, the travel cost per order is given by $c^{\text{pm}} d^{\text{sp}}(k, P)$, where $d^{\text{sp}}(k, P) = D^{\text{sp}}(k, P)/k$ represents the expected travel distance per order. Thus, the delivery cost per order is given by

$$c(\lambda, k, P) = \frac{Nw}{\lambda} + c^{\text{pm}} d^{\text{sp}}(k, P). \tag{6.12}$$

Lemma 6.3 *For a given order rate λ served by sector pooling policies with fixed P, the delivery cost per order decreases as batch size k increases.*

Proof *Since the tour distance $D^{\text{sp}}(k, P)$ increases sub-linearly in k based on (6.6), $d^{\text{sp}}(k, P) = D^{\text{sp}}(k, P)/k$ is decreasing in k. This completes the proof.*

It follows from Lemma 6.3 that for a given delivery time guarantee $\overline{T}$ and fixed P, the optimum batch size $k^*(\overline{T}, P)$ also minimizes the delivery cost. Thus, the store chooses a batch size of $k^*(\overline{T}, P)$ to fulfill orders. For ease of notation, we use $d^{\text{sp}}_*(\overline{T}, P)$ to denote the expected distance per order $d^{\text{sp}}(k^*(\overline{T}, P), P)$.

6.6.2 A Unit Economics Lower Bound

Suppose the store charges a fee p per order for delivery within time $\overline{T}$. This implies an hourly revenue λp and cost $Nw + \lambda c^{\text{pm}} d^{\text{sp}}_*(\overline{T}, P)$. Then, the store's hourly profit is given by

$$\Pi = \lambda p - (Nw + \lambda c^{\text{pm}} d^{\text{sp}}_*(\overline{T}, P)). \tag{6.13}$$

For the delivery business to be profitable (i.e., for $\Pi \geq 0$), the order rate λ must satisfy

$$\lambda \geq \frac{Nw}{p - c^{\text{pm}} d^{\text{sp}}_*(\overline{T}, P)}. \tag{6.14}$$

We refer to 6.14 as the *unit economics* lower bound.

6.6.3 When is Last-Mile Delivery Viable?

While system efficiency constraints impose an upper bound on the order rates that can be served by the platform, the unit economics of delivery dictate a minimum order rate necessary for a profitable delivery business. Combining the system efficiency upper bound (6.8) and the unit economics lower bound (6.14), we have the following result.

Theorem 6.3 *Consider a store fulfilling orders using sector pooling policies and offering a delivery time guarantee $\overline{T}$. Suppose p is the maximum delivery fee that customers are willing to pay for an order. Then, the delivery service is profitable only if*

$$p > \frac{w}{\mu_*^{\mathrm{sp}}(\overline{T}, P)} + c^{\mathrm{pm}} d_*^{\mathrm{sp}}(\overline{T}, P). \tag{6.15}$$

Proof *The proof follows by combining* (6.8) *and* (6.14)*, and rearranging terms.*

Observe that the right-hand side in (6.15) can be recovered by substituting $\lambda = \mu_*^{\mathrm{sp}}(\overline{T}, P)$ and $k = k^*(\overline{T}, P)$ in (6.12). Since the order rate must be strictly lower than the optimal service rate as reflected in (6.8), the above expression is a strict lower bound for the delivery cost.

6.7 Simulation Study

6.7.1 Model Calibration

We calibrate our model based on a typical grocery store operating in the city of Los Angeles (LA), California. The average last-mile speed in LA is $v = 22$ mph [92]. Based on prevailing practice, we set the delivery radius to be $R = 5$ miles which implies a delivery region of size 78.5 square miles. We set the pickup time to be $t_{\mathrm{pick}} = 10$ min based on [82]. Note that this is a best-case estimate which assumes efficient store design and operations. We consider a similar best-case drop-off time estimate $t_{\mathrm{drop}} = 2$ min, which does not take into account routine sources of delay such as unresponsive customers, or parking unavailability. We estimate the expected orders per hour in the delivery region as follows. There are 1.4 million households in LA within a land area of 469 square miles [93], resulting in a household density of 2,985 per square mile. Assuming each household does groceries once a week on average, there are $2985 \times 78.5 = 234,322$ potential order requests arising within the delivery region per week. Since online delivery accounts for roughly 10% of all grocery sales at present [28], we assume that 23,432 of these orders are fulfilled through the online channel. Assuming that the store operates for 12 hours each day of the week, the store serves an order rate of $\lambda = 279$ orders per hour. We set the driver wage to be $w = \$17.2$/hr based on the minimum wage in LA, adjusted for payroll taxes and employee benefits [58].

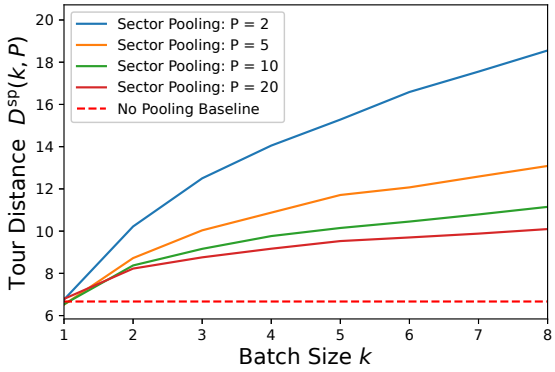


Figure 6.3: Tour Distance vs. Batch Size.

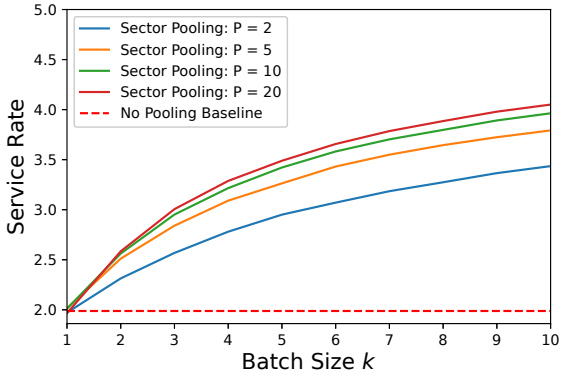


Figure 6.4: Service Rate vs. Batch Size.

6.7.2 Tour Distance and Service Rates

To find service rates, we estimate the tour distance $D^{sp}(k, P)$ using Monte Carlo simulations ($n = 500$) with $P \in \{2, 5, 10, 20\}$. As expected from the analytical approximation (6.6), the tour distance increases sub-linearly with the batch size k in Figure 6.3. Thus, the distance travelled per order $D^{sp}(k, P)/k$ decreases as batch size increases. Observe that the tour distance at $P = 2$ is considerably higher than at $P = 5$, whereas it decreases only marginally beyond $P = 10$.

Figure 6.4 shows that the service rate increases rapidly at small values of k but tapers off at larger values. This suggests that pooling can provide substantial gains even with small batch sizes. For $P \geq 10$, a batch size of 3 results in a 50% increase in service rate. Increasing the number of partitions P uniformly increases the service rate across k. As discussed above, most of the value from spatial partitioning can be derived with $P = 10$ partitions since the travel distance per order decreases marginally beyond this point.

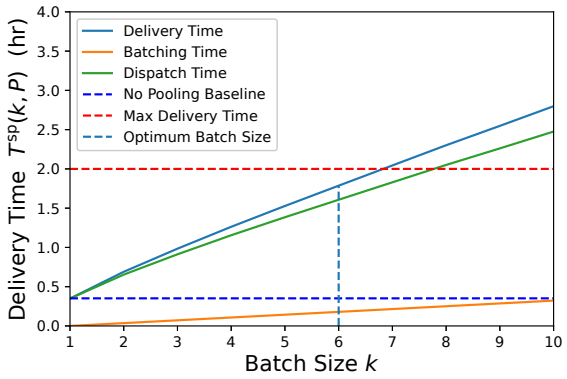


Figure 6.5: Components of Delivery Time.

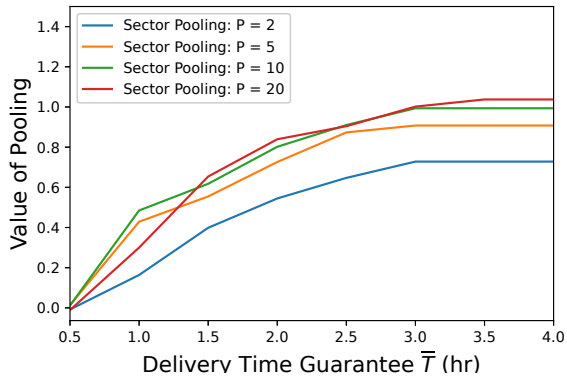


Figure 6.6: Value of Pooling vs Delivery Time.

6.7.3 Delivery Times and Optimum Batch Size

Figure 6.5 illustrates how the delivery time varies with batch size for $(k, 10)$-sector pooling policies. As suggested by Lemma 6.2, the delivery time increases with k. In contrast with the service rate, the delivery time seems to increase almost linearly. The analytical expression for delivery time provided in Lemma 6.1 can help explain this observation. Notice that the first and third terms in (6.5) which represent the batching and fixed time respectively, are both linear in k. The second term represents the travel time which is sub-linear in k. Since the fixed time component dominates the sub-linear term for $k \geq 2$, the total delivery time turns out to be approximately linear. Assuming a delivery time guarantee of $\overline{T} = 2$ hrs, we find the optimum pooling size to be $k^*(2, 10) = 6$ using (6.7).

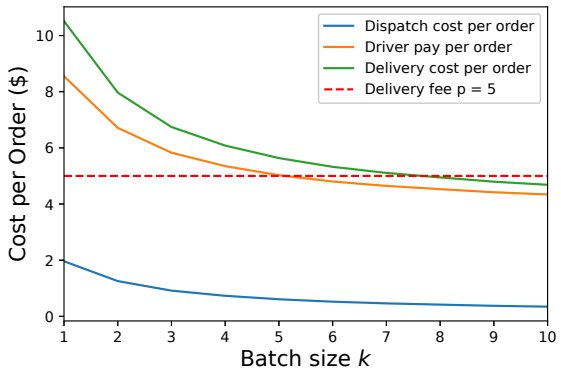


Figure 6.7: Delivery Cost vs. Batch Size.

6.7.4 The Value of Pooling

Recall from our discussion in Section 6.5.3 that the delivery time guarantee $\overline{T}$ dictates the value that can be derived from pooling. We observe from Figure 6.6 that there is no value derived from pooling at $\overline{T} = 30$ min across all sector pooling policies. The optimum pooling size in this case is $k^*(0.5, P) = 1$ for all P. This is because the short time window for delivery precludes the pooling of orders. The value of pooling can be clearly observed for larger $\overline{T}$. For instance, a delivery time guarantee $\overline{T} = 1$ hr is sufficient to obtain a 50% increase in service rate on account of pooling. The value of pooling tapers off beyond $\overline{T} = 2$ hrs.

6.7.5 The Economics of Delivery

Figure 6.7 illustrates how delivery cost per order and its two components (driver pay and travel cost) vary with batch size k. Each of these costs drop sharply up to $k = 6$, after which they decrease slowly. Driver pay makes up a bulk of the delivery cost. At $k = 6$, driver pay accounts for 90% ($4.8) of the total delivery cost ($5.3). Surveys show that most customers are unwilling to pay a delivery fee of more than $5 per order [25, 75]. Notice that the delivery cost is higher than $5 for $k < 8$. Thus, large batch sizes are crucial for the unit economics to be favorable in this setting.

From Figure 6.8, we observe that a delivery time guarantee of $\overline{T} = 2.5$ hrs is necessary to bring down the delivery cost below $5 by allowing for large batch sizes. Thus, grocery delivery with a guarantee $\overline{T} < 2$ hrs is not economically viable in LA. Since LA is one of the denser cities in the US, this result also explains why 1 hr delivery is not offered in most US cities.

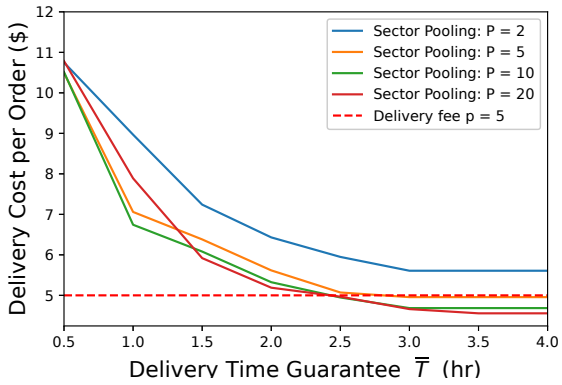


Figure 6.8: Delivery Cost vs. Delivery Time.

6.8 Discussion

6.8.1 Implications for Grocery Delivery

Our analysis in Section 6.7 suggests that the resulting efficiency gains from pooling are at the core of a profitable business model for grocery delivery. As delivery platforms such as Instacart and Cornershop rely on contractors using their own vehicles to fulfill orders, they can realize only a part of the value of pooling. Considering that batch sizes $k > 3$ would not fit in an average sedan, the delivery cost would be more than $7 per order in LA. Moreover, the pickup time would be much longer than our optimistic estimate of 10 min since drivers would need to shop for and load orders themselves. Thus, the unit economics for delivery platforms are particularly unfavorable. Large grocery chains with the means to fulfill order deliveries can leverage the value of pooling to a greater extent by using larger trucks, as well as having dedicated staff for loading of orders.

Our model provides an intuitive lens through which the impact of technological innovations can be interpreted. Since driver pay accounts for most of the delivery cost, fulfilling orders using drones or automated vehicles can significantly improve unit economics. Without driver pay, the delivery cost would go down to less than $2 per order, thus making the delivery business very lucrative. This explains why companies are investing heavily in automated delivery. It is unclear when delivery technologies would become mature enough to replace human drivers. In the near term, optimizing store layout and setting up of automated order picking systems can have an impact by substantially reducing order pickup times (t_{pick}).

6.8.2 A Unified Framework for Last-Mile Delivery

While we have treated the special case of a grocery store, our model can be used to analyze a broader class of last-mile delivery services. For instance, the grocery store can be replaced by a restaurant in the case of food delivery and a depot in the case of package delivery.

Order arrival rates and typical delivery guarantees will also vary according to the context. Our model provides a framework to view the diversity of last-mile delivery services through a unified lens. This allows us to compare the viability of business models both within and across such contexts. For instance, it can explain why food delivery services charge a substantially higher delivery fee as compared to their grocery delivery counterparts. Since speed of delivery is of the essence in the former context, there is little room for pooling orders and as a result, delivery costs are high. On the other hand, last-mile delivery for packages is considerably cheaper since it is not as sensitive to delivery time and hence, allows for a much larger pooling batch size and delivery radius.

Part II

Automated Vehicle Safety

Chapter 7

Automated Vehicle Safety: An Overview

We start the second half of this thesis with an overview of the promise of automated vehicles (AVs), safety concerns about their deployment, and prevailing approaches to alleviate these concerns. We end this chapter with some fundamental questions about AV safety capabilities that we hope to address in the remainder of this thesis.

7.1 The Dream: A Future without Crashes

Traffic crashes are a leading cause of death in the US [94]. In 2018, more than 36,000 lives were lost to mishaps on the road [5]. Moreover, 20% of these fatalities were vulnerable road users such as pedestrians and bicyclists. The economic cost of these crashes is estimated to be $836B annually [95]. With the advent of automated vehicle (AV) technology, the vision of a safe transportation system seemed within reach. This vision captured the imagination of both venture capitalists and established automobile companies. A report by Intel [10] forecasted that AVs would save 600,000 lives and $230B in economic costs by 2045. Within three years from August 2014 to June 2017, the AV industry attracted more than $80 billion dollars [96], so over the decade at least $100 billion dollars have been invested, exceeding the cost of the Apollo program. In 2012, Google co-founder Sergey Brin predicted AVs would be widely available in five years, and AV companies shared the belief that the goal would be realized by 2020 [97, 98, 99, 100].

7.2 The Dream Deferred

While limited testing in selected suburban areas and operational design domains (ODDs) has started in the recent past, it now appears that the goal of widespread deployment of AVs will be deferred for the next few decades [101]. Over time, the overwhelming complexities involved in urban driving have become apparent. In order to be safe, the AV needs

to routinely navigate through hazardous road conditions, unseen obstacles, vehicles with conflicting paths, and inattentive pedestrians. John Krafcik, the CEO of Waymo, has stated that full autonomy may never be achieved [13]. Recent testing offers promise about the ability of AVs to prevent a significant fraction of fatal crashes [102]. At the same time, fatal crashes involving AVs show that faulty technology can end up costing lives [103, 104].

7.3 Not all Crashes due to Human Error

Currently, the central approach followed by AV companies is to break the complex driving task into sub-tasks of sensing, perception, behavior prediction, planning and control. The expectation is that as these sub-tasks are solved with greater precision, we get closer to eliminating all traffic crashes. In particular, AV companies claim that they can eventually eliminate 94% of all crashes caused by human error, quoting a report by the National Highway Traffic Safety Administration (NHTSA) from 2008 [105]. However, the report only states the following:

> The critical reason was assigned to drivers in an estimated 2,046,000 crashes that comprise 94 percent of the NMVCCS crashes at the national level. However, in none of these cases was the assignment intended to blame the driver for causing the crash.

In other words, the 94% of all crashes attributed to human error involve not only impaired or distracted driving, but also such causes as "false assumption of other's actions", "decision error", "recognition error" and "inadequate surveillance". Clearly, not all such crashes are simply a consequence of having reckless and error-prone humans at the wheel. This suggests that the actual fraction of crashes that AVs can hope to eliminate is likely to be much lower than 94%.

7.4 Understanding Crash Causes

In order to understand the types of crashes AVs can or cannot avoid, we must first analyze why they occur. Due to the complex nature of driving, a variety of circumstances can lead to crashes. In addition, crashes are rare events. In the US, a crash occurs once every 500,000 miles on average [106]. Moreover, a fatal crash occurs once every 100 million miles. Since AV fleets have only been tested over tens of millions of miles in limited operational design domains (ODDs), merely analyzing testing data is insufficient for understanding which crashes they can avoid. For instance, Waymo AVs have covered about 20 million miles in on-road testing since 2009 [107]. On the other hand, humans drive billions of miles every year encountering a considerably wider variety of road geometries, weather conditions and neighboring vehicle behaviors than the AVs being tested at present. As a result, they come across a significantly greater diversity of crash modes than their AV counterparts. Analyzing

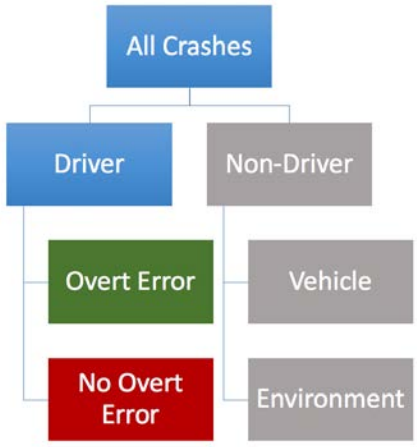


Figure 7.1: A taxonomy of crashes based on crash causes.

human crash data can provide a comprehensive picture of crash causes. NHTSA periodically provides a crash cause based breakdown of a representative sample of national crashes [106, 105]. In order to understand which of these crashes AVs can successfully avoid, we compile the following taxonomy.

7.5 A Crash Taxonomy

As shown in Figure 7.1, crashes are categorized at the first level based on whether they were attributed to driver-related causes. Non-driver related causes included vehicle failures (eg., faulty tires, failed brakes) and challenging environmental conditions (eg., slick roads, fog, bad road design). These non-driver related causes account for 4% of all crashes. Since these crashes were not caused by decisions made by the driver, AVs are unlikely to avoid such crashes at present. Nevertheless, it is conceivable that better automotive technology will eliminate vehicle failure crashes over time. Some environment-related crashes could also be prevented by more cautious driving and better road design.

Driver-related causes account for 94% of all crashes[1]. This statistic is often quoted by AV companies to point to the potential safety benefits that can be brought about by AVs. However, not all driver-related crashes are *caused* by drivers. The NHTSA report explicitly states the following: "Although the critical reason is an important part of the description of events leading up to the crash, it is not intended to be interpreted as the cause of the crash nor as the assignment of the fault to the driver, vehicle, or environment." Thus, drivers are not necessarily at fault for 94% of all crashes. As such, it is not clear whether AVs can

[1]The crash cause could not be identified for 2% of all crashes in studied in [105]. This explains why driver and non-driver related crashes add up to 98% of all crashes.

prevent all of these crashes. This motivates a more refined classification of driver-related crashes based on whether they involved drivers making overt errors.

7.6 Safety Opportunities and Challenges

Overt driver errors include impaired driving, speeding and violating traffic rules. Impaired and distracted driving are involved in 28% and 9% of all fatal crashes respectively [106]. Traffic violators cause 15% of fatal crashes. It seems reasonable to expect that AVs will eventually eliminate such crashes, saving close to 20,000 lives annually. This clearly indicates the immense potential AVs possess to reduce traffic fatalities. Several studies suggest that the fraction of all crashes caused by overt errors is closer to 50% as opposed to 94% [108, 109, 110].

Nevertheless, this implies that a significant fraction of crashes occur despite drivers not making overt errors. An example of this is crashes involving occlusions on the road which prevent even attentive drivers from detecting each other. These crashes are a consequence of the difficulty of driving among other road users with conflicting trajectories, partially observed positions and unpredictable behavior. Scenarios involving such challenges are routinely observed particularly in dense, urban settings. AVs have been known to struggle with maneuvers such as left turns and merging that involve predicting the positions and behavior of other vehicles and pedestrians [111, 112, 113]. This suggests that AVs may not be able to avoid all crashes involving gaps in information about the positions and actions of neighboring road users. We will explore such *information gaps* in greater detail in Chapter 8 and discuss how AVs can deal with them.

Although AVs will not drive impaired or distracted, they consist of hardware and software subsystems that can be faulty and hence, create new crashes in addition to those involving information gaps. The fatal crash involving an Uber AV colliding with a pedestrian in Tempe, Arizona is an important example to keep in mind. In this incident, the AV's Lidar alternated between misclassifying the pedestrian as "vehicle", "unknown object", and "bicycle" in the seconds leading up to the crash [114]. In order to ensure that such errors do not claim more lives, regulators need to ensure that only demonstrably safe AVs are permitted to be deployed in their jurisdictions.

7.7 Scenario Based Safety Assessment

A scenario based analysis of crash causes can provide further insights into the safety capabilities of AVs. A recent NHTSA report [106] provides a crash typology that describes the diversity of hazardous scenarios commonly observed on the roads. It classifies crashes into 36 scenario types based on the set of events leading up to the crash. For example, the crash type *Backing into Vehicle* includes all crashes in which a vehicle collides with another vehicle while backing. The report also contains the statistical distribution of crashes across these

types. This can serve as a useful tool in understanding why crashes occur, how frequently they occur and what type of crashes can be avoided by AVs.

Based on the above NHTSA crash typology, a few crash scenarios are likely to be particularly challenging for AVs. For instance, *Crossing Paths* is a grouping of 6 crash types involving two vehicles moving perpendicular to each other with conflicting paths–a common occurrence at intersections. This group accounts for 19% of all crashes and damages of $135.4B each year. A noteworthy aspect of such crashes is that it is often unclear which vehicle is at fault. There could be static obstacles or surrounding vehicles occluding the field of view of both parties involved so that they do not see each other until it is too late. NHTSA's analysis shows that intersection-related crashes are almost 335 times as likely to have "turned with obstructed view" as the critical reason than non-intersection-related crashes [115]. *Crossing Paths* crashes could also occur due to one or more vehicles violating traffic rules such that their paths intersect. Even if AVs have perfect sensing and perception capabilities, they will not be immune to occlusions and as a result, will end up in such hazardous scenarios. *Changing Lanes* is another crash grouping that accounts for 12% of all crashes and damages of $32.9B every year. Lane changing involves searching for large enough gaps in traffic in order to safely complete the maneuver. Such large gaps may not always be available, especially during peak hours. In such circumstances with small traffic gaps, the lag vehicle either cooperates or is forced to create the required gap so that the merging vehicle can successfully change lanes. Thus, inaccurate predictions of the lag vehicle behavior can lead to dangerous scenarios which ultimately lead to crashes.

Such a scenario-based assessment of safety is reflected in a recent safety report by Waymo [116]. It suggests that while AVs rarely make overt errors, they still get into crashes due to the actions and behavior of neighboring vehicles. At present, it is unclear whether AVs are capable of navigating safely through such precarious scenarios. The typically reported AV safety statistics such as crashes or disengagements per vehicle mile travelled (VMT) are not very informative as they do not provide sufficient insight on how AVs perform in specific crash scenarios. Moreover, the limited on-road deployment of AVs takes place in ODDs and environments where current AV technology is believed to be "safe enough". In such environments, AVs may not encounter many of the above discussed challenging scenarios often enough to gauge their safety performance. For instance, Waymo mainly operates in Phoenix suburbs and only serves 5-10% of ride requests with its fully driverless fleets [117]. The rest of the ride requests are served using fleets with safety drivers. However, hazardous crash scenarios involving occlusions, small traffic gaps and traffic violations are commonplace in dense urban settings. What fraction of such crashes will be eliminated with the introduction of AVs remains an open question.

An alternative technology provides some clues to answer this question. The last couple of decades have seen the rapid adoption of Advanced Driver Assistance Systems (ADAS) that share many features with current AV technologies; for example, Automatic Emergency Braking (AEB), Lane Keeping Assist (LKA), and Adaptive Cruise Control (ACC). Like AV technology, these systems aim to eliminate human error by assisting or warning the driver (e.g., Forward Collision Warning) or by automating driving for a well specified task (e.g.,

Automatic Parking). Each ADAS addresses a certain type of crash scenario. For example, AEB and LKA reduce rear end and lane departure crashes respectively. It seems that for these specific crash scenarios an AV cannot do significantly better than the corresponding ADAS. Thus, ADAS crash reduction studies could provide a good estimate for potential crash reduction due to AV technology. One such study analyzed the field effectiveness of General Motors ADAS based on safety data from 3.7 million vehicles [118]. It found a crash reduction rate ranging from 3% to 81% depending on the ADAS considered. Notably, while ADAS help improve traffic safety, they fail to prevent a significant proportion of system-relevant crashes. This might presage a similar outcome in the future with AVs. Another study predicting AV-related road traffic fatalities using the German In-Depth Accident Study database arrived at the same conclusion [109].

7.8 Pressing Questions

There are several reasons to be optimistic about a future with AVs. They are likely to avoid a significant fraction of crashes that occur due to overt human error. At the same time, it is unclear whether they can avoid the remainder of crashes that occur due to fundamental gaps in information required to ensure safety. Moreover, due to unforeseen vulnerabilities in their software and hardware systems, AVs can be potentially unsafe in routine scenarios that are navigated safely by human drivers. These observations raise some fundamental questions about the safety capabilities of AVs and how they should be deployed:

- *What traffic scenarios are likely to be particularly challenging for AVs?*
- *Is connectivity indispensable for ensuring AV safety?*
- *How does crash risk vary across traffic scenarios and traffic conditions?*
- *How should safety capabilities of AVs be evaluated based on on-road testing data?*

We aim to address these questions in the remainder of this thesis.

Chapter 8

Automated Vehicle Safety: The Vital Role of Connectivity

As discussed in Chapter 7, automated vehicles (AVs) have the potential to eliminate millions of crashes each year that are caused by human error. However, an analysis of recent crashes involving AVs suggests that they are particularly vulnerable in some routine traffic scenarios. A common theme in these scenarios is the presence of *information gaps* that preclude AVs from being safe. In this chapter, we argue that connectivity with infrastructure (I2V) or neighboring vehicles (V2V) is indispensable for safe AV deployment.

8.1 Introduction

Information Gaps: A Fundamental Safety Challenge

A fundamental prerequisite for an AV to be safe on the roads is an accurate understanding of the positions and behavior of neighboring vehicles, both of which might be partly unknown to it. Gaps in such safety critical information can result in crashes despite vehicles not making an obvious error. We illustrate below how *information gaps* can cause crashes through a common on-road scenario.

Consider the scenario depicted in Figure 8.1 where the yellow vehicle is attempting an unprotected left turn. The vehicles in the opposing through-moving lane (left-most lane in Figure 8.1) have right-of-way and hence, the yellow vehicle must yield to them. The vehicles in the middle lane are queued up, waiting for an opportunity to make a left turn. The presence of these vehicles occludes the field of view of the left-turning and through-moving vehicles. As a result, both vehicles keep moving until they see each other, at which point, it is too late to avoid a collision.

This example illustrates how the lack of information about the positions of vehicles with conflicting paths can result in a crash despite the drivers not making an obvious error. In fact, an Uber AV was involved in a similar crash during on-road testing in March 2017

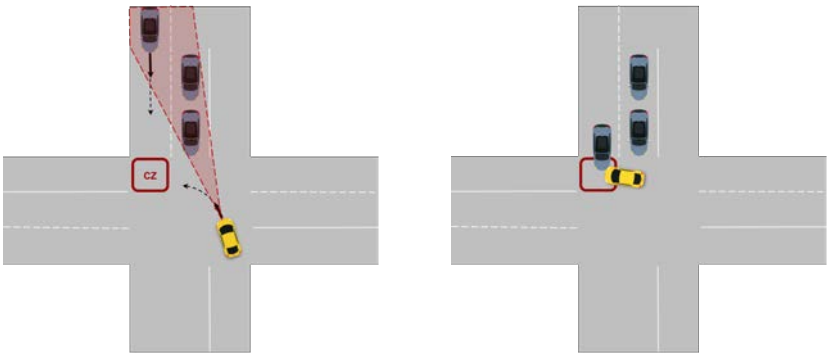


Figure 8.1: (Left): Yellow vehicle making an unprotected left turn with through moving vehicles occluded by queued up vehicles in the middle lane. The left turning vehicle's occluded field of view is illustrated by the shaded red region. The red box *CZ* denotes the conflict zone. (Right): The left turning and through moving vehicles do not detect each other due to the occlusion and hence, cannot stop in time to avoid a crash.

[119]. More generally, knowledge about both the position as well as behavior of neighboring vehicles is a pre-requisite to be safe on the roads. However, we routinely encounter gaps in this knowledge in a variety of driving scenarios. Merging on to a freeway is a scenario which requires a driver to predict how neighboring vehicles will respond to their actions. Red-light violations are an example of our assumption about the rule-following behavior of neighboring vehicles breaking down, resulting in hazardous scenarios. Since safety in these scenarios depends on not just our own behavior but also the position and behavior of other vehicles, it is not immediately clear whether better AV technology can prevent such crashes.

At present, most AV companies aim to develop a *fully autonomous* vehicle that does not communicate with either infrastructure or neighboring vehicles in order to function safely [120]. The lack of connectivity in current AV designs poses a major safety challenge especially in driving scenarios involving information gaps arising from occluded view or unpredictable behavior of other road users. A key question then is: *In the absence of connectivity, can AVs be safe despite information gaps?*

Resolving Information Gaps

To tackle such information gaps, a common approach in the AV industry currently is to to arrive at the best possible estimate for the state and behavior of neighboring vehicles given whatever the AV can observe. While this approach may provide satisfactory performance in many cases, it is unlikely that this is a complete solution to address the challenge of information gaps. Indeed, evidence from AV testing supports this observation [111, 112, 113].

There are two approaches proposed in the AV literature to overcome such information gaps. Researchers from Mobileye, an AV company acquired by Intel, proposed the Respon-

sibility Sensitive Safety (RSS) framework to specifically deal with this challenge [121]. The key idea is to limit an AV's maneuvers such that it is robust to all reasonable states and behaviors that are consistent with its partial observations, thus ensuring safety. However, guaranteeing safety alone is not desirable if it comes at the cost of a sharp drop in throughput. An extreme example that illustrates this point is the recommendation for an AV to not move out of the garage – this is guaranteed to be safe, but is clearly untenable. It is not immediately clear whether the RSS framework can guarantee safety in the presence of information gaps without a substantial reduction in throughput.

Another approach to deal with information gaps is to explicitly bridge them using either infrastructure-to-vehicle (I2V) or vehicle-to-vehicle (V2V) connectivity. In the case of occlusions and traffic violations at intersections, I2V communication would ensure that the involved vehicles detect each other in time [122, 123, 124]. For the lane changing case, V2V communication would ensure that the involved parties are in agreement, thus preventing a potential hazardous situation [125, 126, 127].

Deterrents to Connectivity

There are several deterrents to AV companies adopting I2V/V2V connectivity at present. I2V requires expensive instrumentation. Both I2V and V2V require vehicles to have the required communication hardware installed which further adds to the cost. Moreover, relying on such connectivity limits the deployment of AVs to regions in which the penetration rate of connected vehicles and infrastructure is sufficiently high. Since this is true for only a few regions at present, the region of operation of AVs is constrained. This results in a weak business case. There are also cybersecurity vulnerabilities that arise due to communication with neighboring vehicles or infrastructure [128, 129].

8.2 Summary of Contributions

In this chapter, we develop a framework to study hazardous scenarios involving information gaps that routinely occur in dense urban environments. We study three groups of scenarios with information gaps: (a) occlusions, (b) traffic violations, and (c) behavior prediction uncertainty. We argue that in these scenarios, even with perfect sensing and perception capabilities, AVs would still face significant challenges if they want to guarantee a safety level comparable to human baselines. That is, to ensure such a safety level, AVs must significantly limit their space of maneuvers and sacrifice on traffic efficiency and throughput. We also conclude that the worst-case approach, as laid out by Mobileye's RSS framework, does not provide a practical solution to address the safety consideration in these scenarios. Our analysis suggests that a significant portion of crashes due to information gaps would still persist under the current approach followed by AVs. We discuss how connectivity with infrastructure or surrounding vehicles can alleviate these safety concerns by directly addressing the information gap in these scenarios.

8.3 Occlusions

Let us consider a commonly occurring on-road occlusion scenario and analyze how an AV should act in order to be safe. Consider an AV at a signalized traffic intersection on the left-turning lane as illustrated in Figure 8.2. We assume that the AV is equipped with sensors (e.g., LIDAR) that provide a 360 degree view of its surroundings. There is no protected left-turn phase, so it is waiting for gaps in the on-coming traffic to make a left turn. The left-turn opposing lane is queued up, thus, blocking the AV's view of through-moving vehicles (TMVs) in the opposing lane.[1]

The task of the AV is to choose the right time to make a left-turn so that it does not collide with through-moving vehicles. This scenario falls within *Crash Type 30: Left Turn Across Path/Other Direction (LTAP/OD)* of the NHTSA crash typology and accounts for 5.8% of all crashes. Furthermore, such crossing path crashes have the highest comprehensive costs and equivalent lives lost among all crash groups in the NHTSA report [106]. Recognizing the safety costs of such crashes, fleet operators such as UPS design their routes so that they do not involve left turns [130].

We assume that traffic in the opposing through lane is free-flowing and the arrival of TMVs is modeled as a Poisson process with rate λ. We assume that the through-moving vehicle can make an evasive maneuver (for e.g., hit the brakes) only after it sees the left-turning AV, else it maintains its current speed. As AVs will have to drive among human drivers when they are introduced on the roads, we assume that all other vehicles are human driven with a reaction time $\rho \in [0.7, 2.5]$s[2]. We maintain this assumption for the rest of this chapter. We use the following values for relevant intersection geometry and vehicle parameters:

- Lane width - 4 m,
- Vehicle length - 4 m,
- Vehicle width - 2 m,
- Maximum acceleration rate - 3 m/s^2,
- Maximum deceleration rate - 4 m/s^2.

[1] As the LIDAR is placed on top of the AV, such occluding vehicles may not always be a problem. A LIDAR will not be able to detect the occluded TMVs and pedestrians if the height at which the LIDAR is placed is lower than the height of nearby occluding vehicles. Suppose the AV is a sedan (average height 56 inches) with a LIDAR placed one foot above the roof. This would imply that the AV cannot view beyond any vehicle taller than 68 inches, thus resulting in a potential occlusion. This includes light pickup trucks (average height - 76 inches) as well as heavy vehicles which tend to be even taller. Note that light pickup trucks alone account for $\sim$ 20% of US vehicle market share, thus, such occlusion scenarios are likely to occur frequently.

[2] We use the term *reaction time* to refer to the total time required for perception (mental processing time for recognizing need for evasive action), driver response (time taken to make the evasive maneuver, for e.g., hitting the brakes), and device response (time between driver's action and corresponding vehicle response). This is also referred to as *stopping time* in the literature [131].

8.3.1 Can the AV make an unprotected left-turn with guaranteed safety?

We first consider the conditions under which an AV can make an unprotected left-turn irrespective of the arrival process of TMVs. In this case, the AV's safety depends on whether the TMV can brake in time to avoid the AV[3]. The worst case for the AV in such a situation is when the TMV is just beyond the AV's field of view when it decides to make a left-turn. We use the term *conflict zone* to refer to the region of the intersection where the paths of the left-turning and through moving vehicles intersect – marked in red in Figure 8.2. Let v_{th} and a_{dec} denote the velocity and maximum deceleration rate of the TMVs. Let d_{CZ} denote the distance from conflict zone at which the AV comes into the field of view of the TMV. Based on the above parameters for intersection geometry and vehicle dimensions, d_{CZ} turns out to be 12 m. As the TMV takes time ρ to react to the left-turning AV, its distance from the conflict zone when it starts decelerating is $d_{\text{CZ}} - v_{\text{th}}\rho$. Assuming the TMV brakes at the maximum rate a_{dec}, it can avoid a crash with the AV so long as

$$v_{\text{th}}^2 \leq 2a_{\text{dec}}(d_{\text{CZ}} - v_{\text{th}}\rho). \tag{8.1}$$

Even under the most optimistic choice for TMV reaction time $\rho = 0.7\ s$ [131], an AV can be guaranteed to be safe only if the TMVs are moving at no more than 17 mph. This is much lower than typical through moving vehicle speeds observed at intersections, especially because they have right-of-way in such a setting. Additionally, Figure **??** shows the sensitivity of our analysis with respect to different choices of reaction time ρ and deceleration rate a_{dec}. This suggests that an AV cannot guarantee safety while making an unprotected left-turn under such occlusions.

8.3.2 Can the AV be "safe enough" while making an unprotected left-turn?

Realizing that it is impossible to guarantee safety in this setting, we consider an AV that is willing to accept a probability of collision no greater than p_{coll} under such an occlusion. Note that the arrival rate λ of TMVs is initially unknown to the AV. Let us assume a through-movement velocity $v_{\text{th}} = 25$ mph. Recall that d_{CZ} is the TMV's distance from the conflict zone when the AV comes into its field of view. Let $d_{\text{CZ}}^{\text{min}}$ denote the minimum distance such that the TMV can brake to a stop before reaching the conflict zone. If $d_{\text{CZ}} < d_{\text{CZ}}^{\text{min}}$, then we have a potential collision. Although we account for the TMVs attempting to brake when they see the AV, we do not consider more complicated evasive maneuvers like swerving that could also be used by the TMVs. Such scenarios in which a collision would occur unless evasive maneuvers are carried out by the involved vehicles are called *traffic conflicts*.[4] It is

[3]Refer to Appendix Section D.1 for a detailed analysis of why an AV cannot ensure its safety in this scenario even if it attempts to make an evasive maneuver.

[4]The *traffic conflicts* technique was introduced by Perkins and Harris as a surrogate measure of traffic safety [132]. A traffic conflict is defined as an event that would lead to a crash unless an evasive action

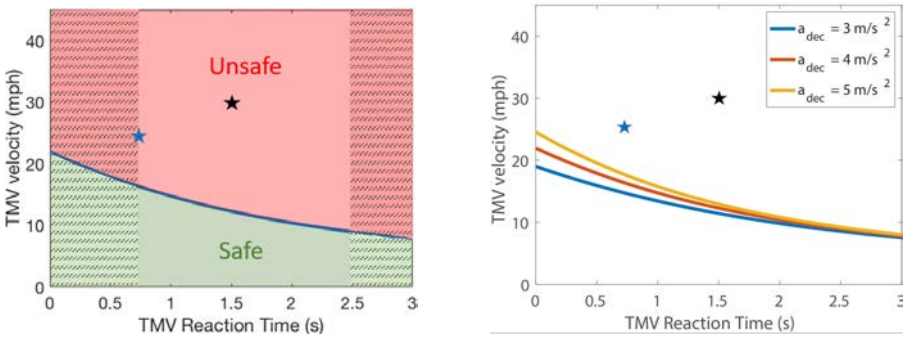


Figure 8.2: Safe TMV velocity as a function of TMV reaction time and deceleration rate. (Left:) Pairs of (v_{th}, ρ) that ensures safety are in green, whereas the unsafe pairs are in red. Typical human reaction times varies between 0.7 s and 2.5 s (unshaded region). The blue star ($\star$) denotes the optimistic ($v_{\text{th}} = 25$ mph, $\rho = 0.7$ s) pair considered in our analysis, while the black star ($\star$) denotes typical observed values ($v_{\text{th}} = 30$ mph, $\rho = 1.5$ s) in practice. Note that both points are in the unsafe region.
(Right:) Change in the safe region boundary (as shown in left) for different values of deceleration rate a_{dec}. The higher the deceleration rate, the higher is the maximum TMV velocity that can still ensure safety. Note that even for higher values of deceleration rate, both the optimistic and typical values are in the unsafe region.

reasonable to expect that at least some of these traffic conflicts will result in crashes. We allow for the possibility of other evasive maneuvers being used by the TMVs by introducing γ — the ratio of traffic conflicts to collisions. We analyze the probability of a traffic conflict and then translate it to a collision probability using the following equation:

$$p_{\text{conf}} = \gamma p_{\text{coll}}. \tag{8.2}$$

Note that the conflict-to-collision ratio γ depends on the scenario considered. For our analysis, we use $\gamma = 1490$ based on an empirical estimate of opposing left-turn conflict-to-collision ratio in [133].[5] Given that the TMV arrivals are Poisson [6], the probability of traffic conflict

such as braking or swerving is taken. As traffic crashes are very rare events, actual crash data is scarce and unreliable as a safety metric. On the other hand, traffic conflicts are abundant and are amenable to empirical estimation. The estimated number of traffic conflicts can then be multiplied by a suitable factor depending on the traffic scenario to get the predicted number of traffic crashes.

[5]Table 8 in [133] contains conflict-to-collision ratios for traffic scenarios such as opposing left-turn, left-turn same direction and through cross traffic under varying traffic volumes. Note that the conflict-to-collision ratio γ is the inverse of the accident/conflict ratio in Table 8.

[6]Modeling the traffic flow as a Poisson process is standard in the literature [134, Ch. 9], [135]. In our analysis we assume that the arrival rate of vehicles are constant and does not change over time. As such, we simplify the estimation problem AVs face to a single scalar parameter estimation. We note that one can consider a more complex model to captures the dynamically varying flow rate at an intersection. However, considering such a more complex model requires the estimation of multiple parameters, and thus, more observation time for AV to estimate the model.

can be translated into the probability that there is a Poisson arrival in an interval of t_{conf} sec, where

$$t_{\text{conf}} = \frac{d_{\text{CZ}}^{\min} - d_{\text{CZ}}}{v_{\text{th}}}. \tag{8.3}$$

If the AV knows λ, its decision is straightforward. It should make the left-turn if the probability of at least one TMV arrival in a time interval of length t_{conf} is less than the AV's maximum allowed probability of traffic conflict, i.e.,

$$1 - e^{-\lambda t_{\text{conf}}} \leq p_{\text{conf}}. \tag{8.4}$$

Thus, the AV would make the turn if $\lambda \leq \lambda_{\max}$, where

$$\lambda_{\max} = \frac{1}{t_{\text{conf}}} \log\left(\frac{1}{1 - p_{\text{conf}}}\right). \tag{8.5}$$

However, the AV does not know λ, which raises the question: *In the time the AV waits at the intersection, can it estimate λ with reasonable confidence so that it can make the left-turn when the situation is indeed safe?*

This question can be posed as a hypothesis testing problem [136]:

$$H_0 : \lambda \geq \lambda_{\max}, \qquad H_1 : \lambda < \lambda_{\max}. \tag{8.6}$$

The AV observes through-moving traffic while it waits for t_{obs} seconds at the intersection. It decides to make the left turn if it can reject H_0. For a level-α test, this implies that the AV makes the turn if it sees no TMV arrivals in t_{obs} seconds, i.e.,

$$e^{-\lambda_{\max} t_{\text{obs}}} \leq \alpha. \tag{8.7}$$

Thus, we have

$$t_{\text{obs}} = \frac{1}{\lambda_{\max}} \log\left(\frac{1}{\alpha}\right). \tag{8.8}$$

A suitable choice for an acceptable collision probability p_{coll} is not evident. We derive an upper bound $\bar{p}_{\text{coll}}$ for this quantity by choosing the intersection with the highest rate of opposing left-turn crashes in San Francisco as our baseline – Market Street and Octavia Street with 10 crashes in the period 2011-17 [137]. Then,

$$\bar{p}_{\text{coll}} = \frac{\text{Number of opposing left-turn crashes per year}}{\text{Number of left-turns under occlusion per year}}. \tag{8.9}$$

Assuming that the occlusion scenario occurs during peak hours, we have

$$\text{Number of left-turns under occlusion per year} = \text{Left-turn rate} \times \text{Number of peak hours per day} \times \text{Number of weekdays in a year.} \tag{8.10}$$

Assuming a flow of 1000 vehicles/hr through the intersection with 10% of vehicles turning left on average, gives a left-turn rate of 100 turns/hr. Assuming 4 peak hours/weekday, we get $\bar{p}_{\text{coll}} = 1.4 \times 10^{-5}$. We use this upper bound in our calculations to get a lower bound for t_{obs} using (8.8). We use $\gamma = 1490$ based on [133]. As $p_{\text{conf}} = \gamma \bar{p}_{\text{coll}} = 2.1 \times 10^{-2}$, we choose $\alpha = 10^{-4}$ for the hypothesis testing problem. As $v_{\text{th}} = 25$ mph (11.18 m/s), we get $d_{\text{CZ}}^{\min} = 23.45$ m using (8.1). From (8.8), we see that the AV needs to observe traffic for $t_{obs} = 443$ seconds in order to make its decision to turn. This is prohibitively high. As a result, we conclude that the AV cannot safely make the turn with the required collision probability without drastically reducing throughput, i.e., waiting at the intersection over multiple traffic cycles until it is confident enough to make the left turn.[7]

8.3.3 Occluded Pedestrians

The scenario considered in the previous subsections gets more challenging when pedestrians are involved. Crashes involving pedestrians have the highest fatality rate among all crash groupings in the NHTSA typology [106]. Such crashes account for 58% of all fatal crashes near intersections in the city of San Francisco [137]. In what follows, we investigate the role of occlusions in causing such crashes.

The NHTSA pre-crash scenario typology includes two crash types to classify vehicle-pedestrian crashes based on whether the vehicle is engaged in a maneuver (e.g., passing, turning): *Crash Type 10: Pedestrian/Maneuver* and *Crash Type 11: Pedestrian/No Maneuver* [106]. However, it does not specify if occlusions contributed to the crash. Through a further classification of the above crash types based on possible occlusions, we arrive at three pre-crash scenarios in which pedestrians being occluded by a neighboring vehicle leads to a crash as illustrated in Figure 8.3.

- **Pedestrian/Maneuver Type 1**: The AV is turning left at the beginning of its green phase while pedestrians are trying to cross as soon as possible since the pedestrian phase is changing from the flashing phase to red. The field of view of the left-turning AV is occluded by queued up vehicles on the adjacent right lane. Such a scenario can occur in both protected and permissive left turn signal phases.
- **Pedestrian/Maneuver Type 2**: As the left-turning AV enters the intersection, it cannot detect pedestrians on the crosswalk as its field of view is obstructed by the

[7] As we already account for braking action by TMVs, the conflict-to-collision ratio γ chosen for our analysis is considerably larger than the actual γ for the given setting. Thus, the required waiting time at the intersection is even larger than what we have calculated.

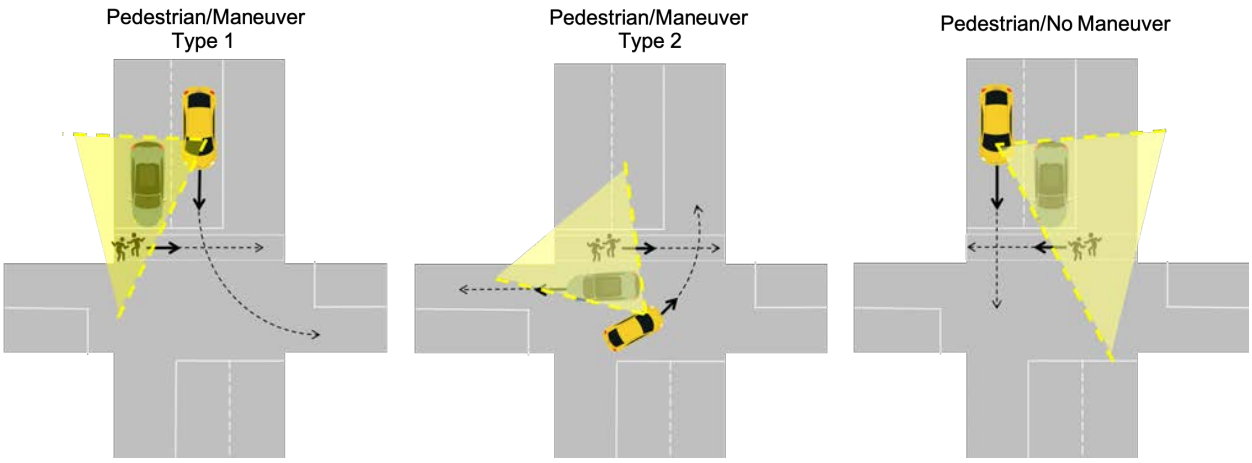


Figure 8.3: Three pre-crash scenarios with occluded pedestrians: AV (in yellow) is occluded by stopped or moving vehicles (in blue), and the AV's occluded field of view is illustrated by the shaded yellow region.

through moving vehicles. This results in a crash if the AV cannot brake in time to avoid the pedestrians. Such a situation can arise during the permissive left-turn phase.

- **Pedestrian/No Maneuver**: The field of view of the through-moving AV is occluded by stopped vehicles on adjacent lanes. At the beginning of the green through phase, the AV might collide with pedestrians who could not finish crossing during the previous phase, especially those who entered the crosswalk during the yellow phase.

We model pedestrians as Poisson arrivals on the crosswalk moving at a fixed velocity. We assume that they are inattentive and as a result, do not attempt to evade the AV. Although this is not always the case, such inattentive behavior is commonly observed during changes in signal phase when pedestrians hastily attempt to cross the road. We use the term *pedestrian conflict zone* to refer to the region of crosswalk intersecting with the AV's path. For our analysis, $t = 0$ is the time at which the AV first detects the pedestrian. We introduce the following notation:

- λ_{ped}: arrival rate (ped/s) of pedestrians,
- v_{ped}: pedestrian speed (m/s) when crossing on yellow/flashing phase/at the end of pedestrian phase,
- $D_{\text{ped\_to\_crash}}$: distance (m) from pedestrian's position at $t = 0$ to the center of pedestrian conflict zone,
- v_{AV}: speed of the AV (m/s),
- $D_{\text{veh\_to\_crash}}$: distance (m) from vehicle's position at $t = 0$ to the pedestrian conflict zone,

- a_{acc}: maximum AV acceleration rate (m/s^2),
- a_{dec}: maximum AV deceleration rate (m/s^2),
- w_{AV}: width of AV (m).

We begin our analysis with the *Pedestrian/Maneuver Type 2* scenario, and then show how this approach can be suitably modified to accommodate the other two scenarios. Notice that the following conditions need to be satisfied for a *Pedestrian/Maneuver Type 2* scenario to end up in a crash:

- **Condition 1:** The AV's view of the crosswalk is obstructed by passing through-moving vehicles when it is making an unprotected left-turn.
- **Condition 2:** When the AV finally detects the pedestrian, it does not have enough time to brake to a stop before the conflict zone.

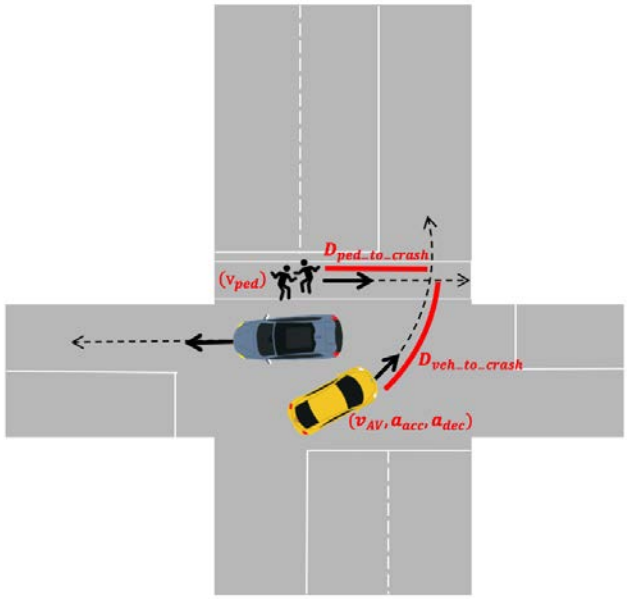


Figure 8.4: Calculating conflict probability for Pedestrian/Maneuver Type 2.

Assuming that both these conditions hold, a crash can occur if the pedestrian's original trajectory intersects that of the AV. When the AV first detects the pedestrian, it has two choices depending on the location of the pedestrian on the crosswalk: decelerate so that the pedestrian can cross the conflict zone before it reaches there or accelerate and cross the conflict zone before the pedestrian arrives there. As we do not account for pedestrians being attentive or more complicated maneuvers such as swerving by the AV, we classify such a scenario as a traffic conflict, as in Section 8.3.2. Let δ denote the time required by pedestrian

to cross the conflict zone, i.e., $\delta = w_{\text{AV}}/v_{\text{ped}}$. Let t_{acc} denote the time taken by the AV to reach the conflict zone while it accelerates. Then,

$$D_{\text{veh\_to\_crash}} = v_{\text{AV}} t_{\text{acc}} + \frac{1}{2} a_{\text{acc}} t_{\text{acc}}^2.$$

Taking the positive root of this quadratic equation gives

$$t_{\text{acc}} = \frac{\sqrt{2a_{\text{acc}} D_{\text{veh\_to\_crash}} + v_{\text{AV}}^2} - v_{\text{AV}}}{a_{\text{acc}}}, \tag{8.11}$$

and a conflict occurs if the pedestrian arrives at the center of the conflict zone in the time interval $\mathcal{T}_{\text{acc}} = [t_{\text{acc}} - \delta/2, t_{\text{acc}} + \delta/2]$.

On the other hand, let t_{dec} denote the time taken by the AV to reach the conflict zone while it decelerates. Then, instead of (8.11), we have

$$t_{\text{dec}} = \frac{v_{\text{AV}} - \sqrt{v_{\text{AV}}^2 - 2a_{\text{dec}} D_{\text{veh\_to\_crash}}}}{a_{\text{dec}}}, \tag{8.12}$$

and a conflict occurs if the pedestrian arrives at the center of the pedestrian conflict zone in the time interval $\mathcal{T}_{\text{dec}} = [t_{\text{dec}} - \delta/2, t_{\text{dec}} + \delta/2]$. Thus, a conflict is unavoidable if the pedestrian arrives in the time interval $\mathcal{T}_{\text{acc}} \cap \mathcal{T}_{\text{dec}} = [t_{\text{dec}} - \delta/2, t_{\text{acc}} + \delta/2]$. This translates into the following condition on $D_{\text{ped\_to\_crash}}$ [8]:

$$D_{\text{ped\_to\_crash}} \in [(t_{\text{dec}} - \delta/2)v_{\text{ped}}),\ (t_{\text{acc}} + \delta/2)v_{\text{ped}}]. \tag{8.13}$$

The probability of this event can be interpreted as the probability of arrival of at least one pedestrian in a time interval of length $(t_{\text{acc}} - t_{\text{dec}} + \delta)$. As the pedestrian arrival process is Poisson with rate λ_{ped}, we have

$$P_{\text{scenario2}}(\text{Conflict} \mid \text{Conditions 1 and 2 hold}) = 1 - e^{-\lambda_{\text{ped}}(t_{\text{acc}} - t_{\text{dec}} + \delta)}. \tag{8.14}$$

Based on intersection geometry and commonly observed values for model parameters, we set these values: $v_{\text{ped}} = 2$ m/s, $\lambda_{\text{ped}} = \frac{1}{60}$ ped/s, $w_{\text{AV}} = 2$ m, $v_{\text{AV}} = 15$ mph (6.71 m/s), $D_{\text{veh\_to\_crash}} = 4$ m, $a_{\text{acc}} = 3$ m/s^2, and $a_{\text{dec}} = 4$ m/s^2. Then, $\delta = w_{\text{AV}}/v_{\text{ped}} = 1$ sec. Using (8.11), (8.12), (8.13), and (8.14), we get

$$\begin{aligned} &D_{\text{ped\_to\_crash}} \in [0.55, 2.07] \text{ m}, \\ &P_{\text{scenario2}}(\text{Conflict} \mid \text{Conditions 1 and 2 hold}) = 0.0125. \end{aligned} \tag{8.15}$$

The above analysis can be conveniently modified for the other two scenarios. For the *Pedestrian/No Maneuver* scenario, Condition 1 should be rephrased as: *The AV's view of the crosswalk is occluded by stopped vehicles in the adjacent lane when it is passing through the*

[8] Please refer to the Appendix Section D.2 for a more detailed explanation.

intersection at the beginning of the green phase. Condition 2 remains the same. As in the *Pedestrian/Maneuver Type 2* scenario, the AV can either accelerate or decelerate to avert a crash. Notice that the expressions (8.11) and (8.12) for t_{acc} and t_{dec} respectively remain the same for this scenario. The main difference is in the relevant intersection and vehicle parameters. As the AV is through moving and has right of way, we have $v_{\text{AV}} = 25$ mph (11.18 m/s) and $D_{\text{veh\_to\_crash}} = 4$ m. The rest of the parameters remain the same as above. Using (8.11), (8.12),(8.13), and (8.14), gives

$$\begin{aligned} &D_{\text{ped\_to\_crash}} \in [0, 1.68] \text{ m}, \\ &P_{\text{scenario3}}(\text{Conflict} \mid \text{Conditions 1 and 2 hold}) = 0.0158. \end{aligned} \tag{8.16}$$

Similarly, for *Pedestrian/Maneuver Type 1*, we set $v_{\text{AV}} = 6.71$ m/s and $D_{\text{veh\_to\_crash}} = 3$ m. Then, we have

$$\begin{aligned} &D_{\text{ped\_to\_crash}} \in [0.063, 1.82] \text{ m}, \\ &P_{\text{scenario1}}(\text{Conflict} \mid \text{Conditions 1 and 2 hold}) = 0.0145. \end{aligned} \tag{8.17}$$

So far, we have calculated conflict probabilities for the three scenarios which are in the range $[0.0125, 0.0158]$. Moreover, Figure 8.5 shows how the conflict probabilities change under varying AV speed v_{AV}. As in (8.2), this conflict probability can be translated into a collision probability as follows:

$$P(\text{Collision} \mid \text{Conditions 1 and 2 hold}) = \frac{P(\text{Conflict} \mid \text{Conditions 1 and 2 hold})}{\gamma},$$

where γ is the conflict-to-collision ratio for the corresponding pedestrian occlusion scenario. Although we are not aware of any empirical estimates for γ specific to scenarios involving pedestrians, it is reasonable to expect that γ should be lower than that for the various scenarios considered in [133]. There are two reasons for this: (i) the empirical estimates are not conditioned on occlusions, and (ii) we account for basic evasive maneuvers such as braking by the AV. Even if we set γ as the largest observed value across all scenarios in [133], we have

$$P(\text{Collision} \mid \text{Conditions 1 and 2 hold}) \approx 2.8 \times 10^{-6}, \tag{8.18}$$

which is significantly high considering that such occlusions are commonplace in urban driving situations. For these reasons, we can expect this collision probability to be considerably higher. This strongly suggests that such occlusions are a major cause of crashes involving pedestrians.

8.3.4 Preventing Crashes with I2V/V2V Connectivity

We saw that under the occlusion scenarios discussed so far, an AV cannot guarantee safety even if it is willing to accept a crash risk comparable to empirically observed crash rates.

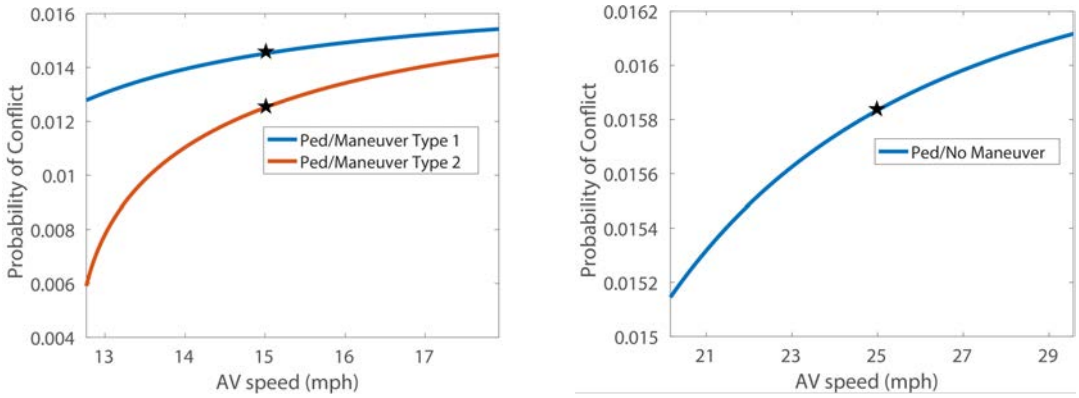


Figure 8.5: Conflict probability as a function of AV's speed for the three occluded pedestrian scenarios. The black star ($\star$) denotes the typical values used in our analysis.

The underlying cause of such crashes is the inability to detect vehicles or pedestrians on conflicting paths in time to take evasive action. Such occlusion scenarios are common during peak hours in dense urban environments. As such, waiting for the occlusion to clear before making the maneuver results in a substantial loss in throughput. Thus, mitigating such crashes without sacrificing throughput requires additional communication between vehicles and the infrastructure so that this critical information is relayed to the necessary parties in time. The vehicle-vehicle scenario illustrated in Figure 8.2 can be prevented by placing a sensor on the through-moving lane at a sufficient distance from the conflict zone depending on the through moving vehicle speeds. For example, if the speed limit on the through moving lane is 30 mph and the worst case human reaction time $\rho = 2.5$ s, a sensor placed at 56 m from the conflict zone will provide enough time for the TMV to prevent the impending crash [123]. In the vehicle-pedestrian scenario illustrated in Figure 8.4, an additional sensor detecting pedestrian movement on the crosswalk would be required to prevent the crash.

8.4 Traffic Violations

Traffic violation is one of the leading causes of crashes on the roads, accounting for 32% of all crashes [106]. Red light running is responsible for a large number of crashes each year. Figure 8.6 depicts the setup for a conflict resulting from this violation. Vehicle V (violator) is going with speed v_{V} from south to north and runs the red light. Vehicle E (ego-vehicle) has the right of way traveling from west to east. In the following analysis, we compute the probability of a conflict between vehicles E and V. Recall that a conflict does not mean crash, but rather a hazardous situation that may lead to a crash, as discussed in Section 8.3. We consider the violation scenario from the point of view of the ego-vehicle. Therefore, we compute the conflict probability under the condition that vehicle E is present at the intersection.

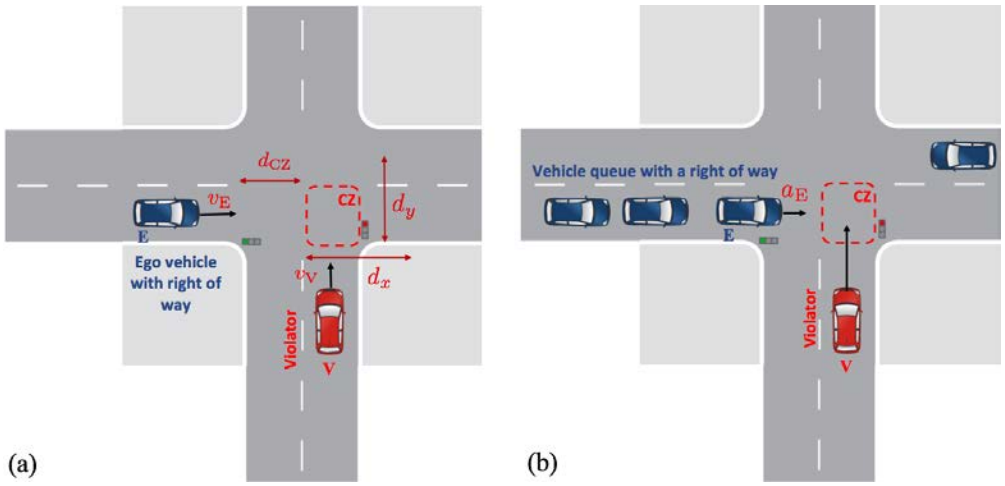


Figure 8.6: Two scenarios of red light violations: (a) vehicles with a right of way randomly arrive during green time; (b) vehicles with a right of way wait in queue and slowly start moving as the light changes to green.

As discussed in [138], at intersections equipped with stop bar detectors and a conflict monitoring card that provides programmatic access to a signal phase, it is possible to monitor red light violations and collect corresponding statistics. An example of such data, collected at the intersection of Montrose Parkway and E. Jefferson Street in North Bethesda, MD, is presented in Figure 8.7.[9] Let $\nu_{\rm A}(t)$ denote the expected number of violations for a given approach A and a given time t. On the south-to-north (northbound – NB) approach, we have $\nu_{\rm NB} = 0.67$ violations between 4:00 and 4:15 A.M., and $\nu_{\rm NB} = 1.91$ violations between 12:00 and 12:15 P. M.. These values for the given time intervals are averaged over a period of one year, from 02/01/2019 through 01/31/2020.

We assume that a red light violation occurs shortly after the signal phase change: when the green (or yellow) light for vehicle V switches to red.[10] Suppose, an average signal cycle length is T_c, so over a period ΔT we expect to have $\Delta T/T_c$ green-to-red switches. Then, the probability of a violation from approach A during a green-to-red switch is

$$p_{\rm A}^{\rm V}(t) = \frac{T_c \nu_{\rm A}(t)}{\Delta T}. \tag{8.19}$$

Thus, given the cycle length $T_c = 150$ seconds at our intersection of Montrose Parkway and East Jefferson Street, and $\Delta T = 900$ seconds, we get $p_{\rm NB}^{\rm V} = 0.67/6 = 0.11$ between 4:00 and 4:15 A.M. and $p_{\rm NB}^{\rm V} = 1.91/6 = 0.32$ between 12:00 and 12:15 P.M.

[9]Note the spike of violations in the east and less in the west directions between midnight and 6 A.M. This could be explained by the fact that the southbound direction leads to Kaiser hospital, and during the night the south-to-north approach has practically no traffic. Hence, eastbound and westbound violators feel relatively safe, not expecting danger from that approach during night hours. Shortly before 6 A.M. the situation changes sharply, as traffic to the hospital starts to increase.

[10]Violating late in the red phase would imply a malicious intent, a rather rare occasion, which we do not consider.

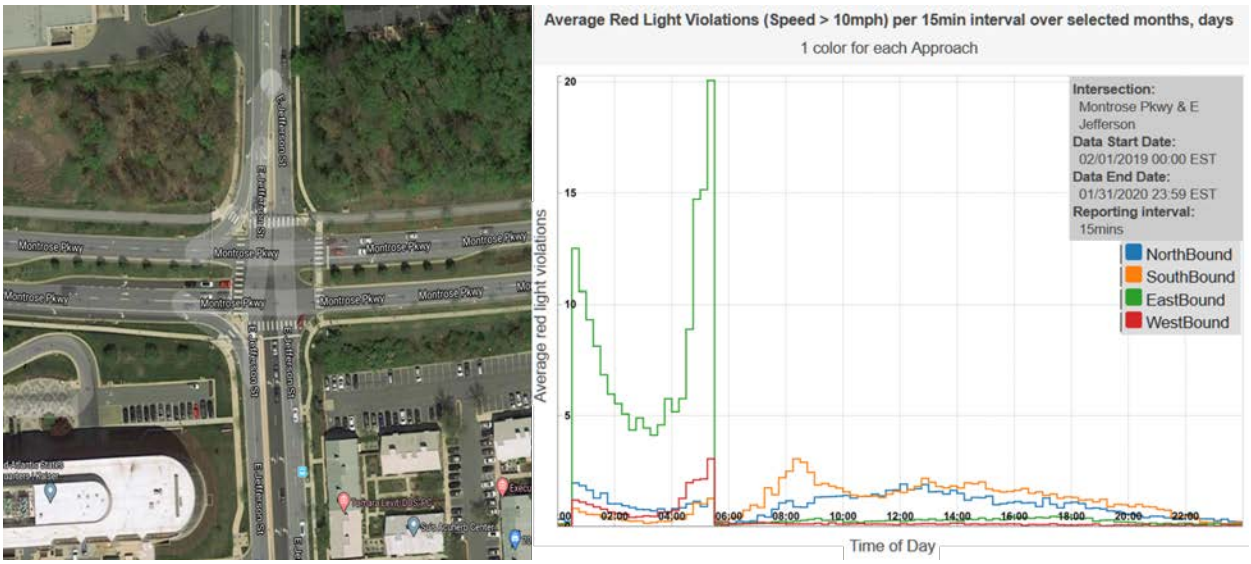


Figure 8.7: Intersection of Montrose Parkway and E. Jefferson Street in North Bethesda, MD: Statistics of red light violations collected over one year from 02/01/2019 through 01/31/2020.

Let us now discuss the probability of a conflict between vehicles E and V. Such a conflict, if it occurs, happens in the conflict zone CZ, indicated by the dashed box in Figure 8.6. Let d_y denote the south-to-north size of this conflict zone – i.e., the distance vehicle V has to cover being exposed to a conflict with vehicle E. This distance includes the expected length of vehicle V itself. In our sample intersection, $d_y = 17$ m (includes the length of vehicle, taken to be 4 m). The time taken by vehicle V to cross the conflict zone is

$$t_{\text{cross}} = t_d + d_y/v_{\text{V}}. \tag{8.20}$$

Here, t_d is the time interval between the signal phase switch that includes red clear time[11] and the instant when vehicle V reaches the conflict zone CZ. The arrival of vehicle E at the intersection during the violator's crossing can happen in two ways:

(a) Vehicle E drives through the intersection at the moment of the signal phase switch and, as its light turns from red to green, at a constant speed without stopping. This case is depicted in Figure 8.6(a).

(b) Vehicle E is already there – waiting at the red light. And, as soon as the light turns green, it starts moving. This case is depicted in Figure 8.6(b).

Since we compute the conflict probability under the assumption that the ego-vehicle is present in the intersection, for both cases (a) and (b), $p^{\text{E}}_{\text{EB}} = 1$.

[11] Red clear time is when the signal light for all the intersection movements is red. It is activated during the signal phase switch and exists for safety purposes – to allow everyone who is already inside the intersection to complete their maneuvers before movements from different approaches are given the right of way. The typical duration of the red clear period is 2-3 seconds.

It remains to compute the probability of a conflict between vehicles V and E under the condition that vehicle V runs the red light and vehicle E is also present at the intersection. Let d_{CZ} denote the distance from the stop bar of the west-to-east approach to the conflict zone CZ, and d_x – the size of the conflict zone CZ in the west-to-east direction. Distance d_x includes an expected length of vehicle E. In our sample intersection, $d_{\text{CZ}} = 16$ m and $d_x = 16$ m (accounts for the length of vehicle E, taken to be 4 m).

Denote the red clear time t_{rc}, and set

$$t_d^* = t_d - t_{\text{rc}}.$$

Here, we assume $t_{\text{rc}} = 3$ seconds. A conflict exists when both vehicles, V and E, are present in the conflict zone CZ simultaneously. This happens if the following condition holds:

- Case (a):
$$\frac{d_{\text{CZ}}}{t_d^* + d_y/v_{\text{V}}} < v_{\text{E}} \le \frac{d_{\text{CZ}} + d_x}{t_d^*}, \tag{8.21}$$
where v_{E} is the speed with which vehicle E arrives at and goes through the intersection. The left part of this inequality states that vehicle E reaches conflict zone CZ before the violator E manages to cross CZ. The right part of this inequality implies that vehicle E would not be able to leave conflict zone CZ before the violator V reaches this conflict zone.

- Case (b):
$$2\frac{d_{\text{CZ}}}{(t_d^* + d_y/v_{\text{V}})^2} < a_{\text{E}} \le 2\frac{d_{\text{CZ}} + d_x}{t_d^{*2}}, \tag{8.22}$$
where a_E is the acceleration with which vehicle E starts moving once its light turns green. Conditions described by the left and the right sides of this inequality are the same as in (8.21).

It is possible to estimate the probabilities of conditions (8.21) and (8.22) for different delay periods t_d, if we have some idea about the range of values for v_{V}, v_{E} and a_{E}.

At the intersection of Montrose Parkway and E. Jefferson Street the speed limit in direction south-to-north is 25 mph (11.18 m/s). To be more conservative, we will assume the speed of the violator, $v_{\text{V}} = 10$ m/s. The speed of vehicles crossing this intersection in the west-to-east direction measured by the detectors is presented in the histogram in Figure 8.8. The typical vehicle acceleration from the intersection stop bar after its light turns green ranges between 1 and 2 m/s^2. Assuming that $a_{\text{E}} \sim \mathcal{N}(1.5, 0.25)$, we can estimate the conditional probabilities of the conflict between vehicles E and V for cases (a) and (b) for different values of t_d – see Figure 8.9.

Finally, we can express the probability of a conflict between vehicles E and V:

$$p_{\text{conflict}} = \begin{cases} p_{\text{NB}}^{\text{V}} p\left(\frac{d_{\text{CZ}}}{t_d^* + d_y/v_{\text{V}}} < v_{\text{E}} \le \frac{d_{\text{CZ}} + d_x}{t_d^*}\right), & \text{case (a)}, \\ p_{\text{NB}}^{\text{V}} p\left(2\frac{d_{\text{CZ}}}{(t_d^* + d_y/v_{\text{V}})^2} < a_{\text{E}} \le 2\frac{d_{\text{CZ}} + d_x}{t_d^{*2}}\right), & \text{case (b)}, \end{cases} \tag{8.23}$$

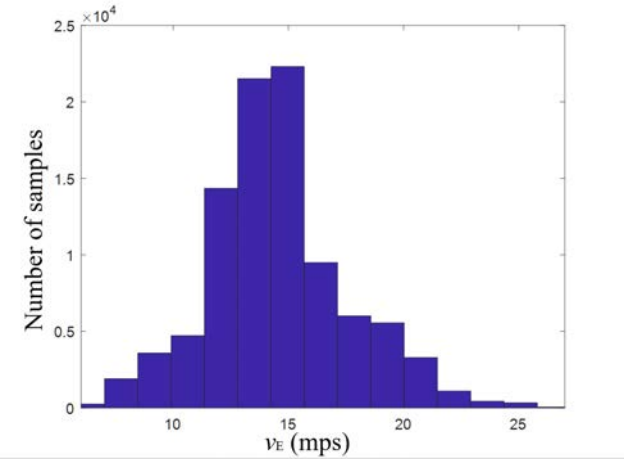


Figure 8.8: Distribution of speeds v_{E}, with which vehicles cross the intersection of Montrose Parkway and E. Jefferson Street in the west-to-east direction. The speed values are taken over a random week of 2019.

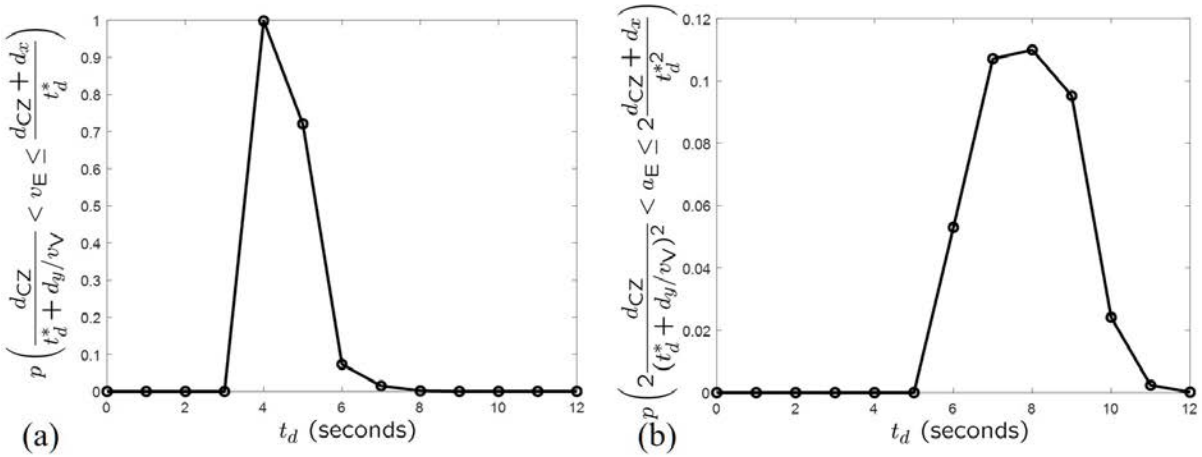


Figure 8.9: Probabilities of a conflict under the condition that both vehicles, V and E, arrive at the intersection of Montrose Parkway and E. Jefferson Street at the same time – cases (a) and (b).

where $p_{\mathrm{NB}}^{\mathrm{V}}$ is determined from (8.19).

Given the parameters of our intersection, these probabilities are presented in Figure 8.10 for two time periods: 4:00-4:15 A.M. and 12:00-12:15 P.M..

These conflict probabilities can be translated into collision probabilities using (8.2) with $\gamma = 2040$ estimated for the through-cross traffic scenario in [133]. Based on a similar argument as in Section 8.3, this suggests that such crashes will occur with a considerably high probability, which might explain why traffic violations are involved in one-third of all

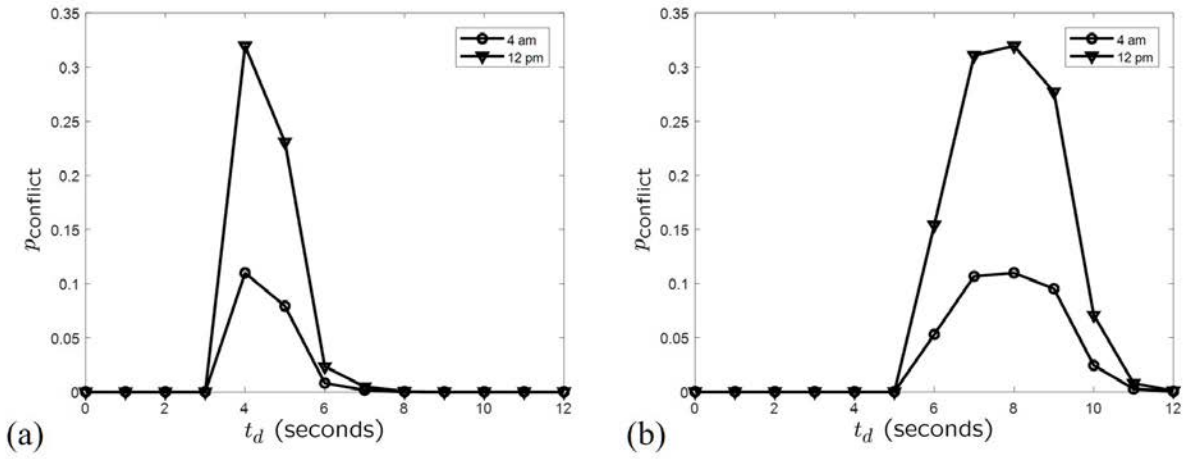


Figure 8.10: Conflict probability computed using data from intersection of Montrose Parkway and E. Jefferson Street for cases (a) and (b).

crashes. If we think of the ego-vehicle as an AV attempting to pass the intersection, it is evident that such traffic violations will jeopardize its safety.

The probability of such a conflict could be reduced by I2V technology, provided that vehicle E and the roadside infrastructure were *connected*. The violator V would be detected by combining stop bar sensor activation with a signal phasing. Then a notification about a violation in a certain direction would be broadcast by the roadside equipment. Intersection conflict avoidance (ICA) defined in the SAE J2735 standard [124] could be used for this purpose. If vehicle E receives this broadcast before it reaches the conflict zone CZ, it should brake. If, on the other hand, it is already inside CZ, it should accelerate to clear the conflict zone faster.

8.5 Behavior Uncertainty

One of the most challenging aspects of driving on the roads is that our safety depends on the behavior of surrounding vehicles. As humans, predicting the actions of vehicles around us is a key component of driving skill. Attempts are being made to imbue AVs with the same capability by training systems with large amounts of driving data. As there is a vast multitude of driving scenarios and driver behaviors, it is unclear how well such systems generalize to unseen situations. What makes such a prediction task even more challenging is that one's own actions affect surrounding vehicle behavior [139, 140]. For AVs that depend on such behavior prediction modules for planning, inaccuracies in prediction can result in hazardous scenarios. An alternate approach that circumvents the challenge of behavior prediction is to plan based on the worst-case behavior of surrounding vehicles [121]. Although this seems appealing from a safety perspective, it is unclear whether such an

approach is feasible in the real world. If not, uncertainty in behavior prediction will indeed result in crashes. In this section, we investigate whether this is the case using empirical data for a common on-road scenario - merging on to a freeway.

Consider an AV attempting to merge from an on-ramp on to a freeway. As in Section 8.3, all other vehicles are human driven. The AV decides to merge if the gap between the lead and lag vehicle on the merging lane is large enough. This scenario falls within *Type 16 – Changing Lanes/Same Direction* in the NHTSA pre-crash scenario typology and accounts for 6.2% of all crashes. Moreover, this scenario is particularly relevant for AVs as there have been several accounts of them having difficulties merging into traffic [141].

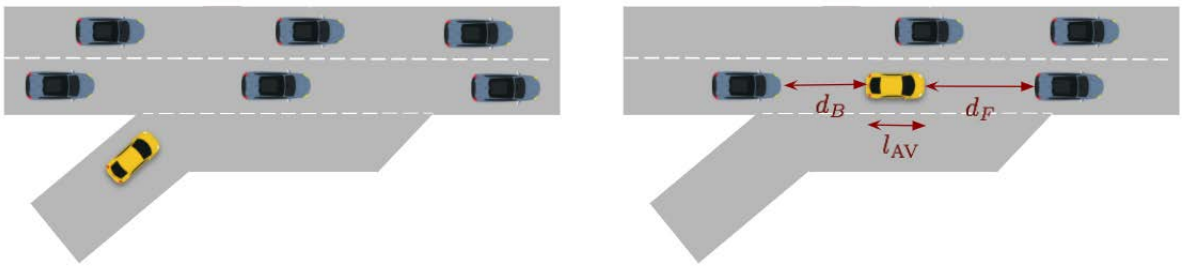


Figure 8.11: An AV (in yellow) merging from on-ramp on to freeway.

We assume that vehicles on the freeway respond to changes in velocity of vehicles in front of them and take evasive action to avoid a crash, albeit with a reaction time. Let ρ_{AV} and ρ_{B} denote the reaction times of the AV and lag vehicle respectively. Let $v_{\text{AV}}, v_{\text{F}}$, and v_{B} denote the speeds of the AV, lead and lag vehicles at the time of the AV's merging decision. Let d_{F} and d_{B} denote the gaps between the lead vehicle and AV, and lag vehicle and AV respectively. We assume that all vehicles have the same maximum acceleration and deceleration rates, a_{acc} and a_{dec} respectively. In order to guarantee safety, the AV must be able to safely evade any potential collision with the lead or lag vehicle regardless of their actions. This implies that the AV must be safe in the following events:

1. Lead vehicle decelerates to a stop while AV merges,
2. Lag vehicle accelerates while AV merges.

We first consider the worst case wherein both events occur simultaneously. We then analyze the case in which only the first event occurs – referred to as the *Single Event* case.

In the worst case, the safe distance for merging d_{safe} can be decomposed as follows,

$$d_{\text{safe}} = d_{F,\text{safe}} + d_{B,\text{safe}} + l_{\text{AV}}, \tag{8.24}$$

Time	Merging Lane Speed (mph)	Observed Gap (m)	Worst Case: Safe Gap (m)	Single Event: Safe Gap (m)
7:50 - 8:05	29.73	43.19	93.79	52.30
8:05 - 8:20	25.43	35.63	84.22	46.36
8:20 - 8:35	21.65	27.44	75.90	41.22

Table 8.1: Observed and safe merging gaps for the NGSIM US-101 dataset.

where $d_{F,\text{safe}}$ and $d_{B,\text{safe}}$ denote the minimum distance required to avoid a collision in events 1 and 2 respectively, and l_{AV} denotes the length of the AV. In event 1, the lead vehicle brakes to a stop and the AV is able to react only after ρ_{AV} seconds, before which it maintains its initial velocity v_{AV}. The lead vehicle is initially at a distance d_{F} from the AV and further traverses a distance $v_{\text{F}}^2/2a_{\text{dec}}$ before it comes to a stop. The AV travels $v_{\text{AV}}\rho_{\text{AV}}$ until it starts decelerating and then travels a distance $v_{\text{AV}}^2/2a_{\text{dec}}$ till it stops. Therefore, the AV can safely evade a collision as long as

$$d_{\text{F}} + \frac{v_{\text{F}}^2}{2a_{\text{dec}}} \geq v_{\text{AV}}\rho_{\text{AV}} + \frac{v_{\text{AV}}^2}{2a_{\text{dec}}}. \tag{8.25}$$

This gives

$$d_{F,\text{safe}} = \max\left\{v_{\text{AV}}\rho_{\text{AV}} + \frac{v_{\text{AV}}^2 - v_{\text{F}}^2}{2a_{\text{dec}}}, 0\right\}. \tag{8.26}$$

In event 2, the lag vehicle accelerates for ρ_{B} until it realizes that the AV is merging, after which it decelerates to avoid crashing into the AV. The worst case in such a setting is if the AV is forced to brake to a stop because of the lead vehicle doing the same. The AV is initially at a distance d_{B} from the lag vehicle and travels $v_{\text{AV}}^2/2a_{\text{dec}}$ further before it comes to a halt. The lag vehicle traverses a distance $v_{\text{B}}\rho_{\text{B}} + \frac{1}{2}a_{\text{acc}}\rho_{\text{B}}^2$ before it starts braking, at which point its speed is $v_{\text{B}} + \rho_{\text{B}}a_{\text{acc}}$. It then travels $(v_{\text{B}} + \rho_{\text{B}}a_{\text{acc}})^2/2a_{\text{dec}}$ before it comes to a stop. Thus, a collision can be avoided if

$$d_{\text{B}} + \frac{v_{\text{AV}}^2}{2a_{\text{dec}}} \geq v_{\text{B}}\rho_{\text{B}} + \frac{1}{2}a_{\text{acc}}\rho_{\text{B}}^2 + \frac{(v_{\text{B}} + \rho_{\text{B}}a_{\text{acc}})^2}{2a_{\text{dec}}}. \tag{8.27}$$

Thus, we have

$$d_{B,\text{safe}} = \max\left\{v_{\text{B}}\rho_{\text{B}} + \frac{1}{2}a_{\text{acc}}\rho_{\text{B}}^2 + \frac{(v_{\text{B}} + \rho_{\text{B}}a_{\text{acc}})^2 - v_{\text{AV}}^2}{2a_{\text{dec}}}, 0\right\}. \tag{8.28}$$

We use the NGSIM US-101 dataset [142] to get estimates for vehicle velocities $(v_{\text{AV}}, v_{\text{F}}, v_{\text{B}})$ and observed merging gaps for three 15 minute intervals between 7:50 A.M. and 8:35 A.M. As the AV is accounting for the worst-case, we set the lag vehicle reaction time $\rho_{\text{B}} = 2.5$

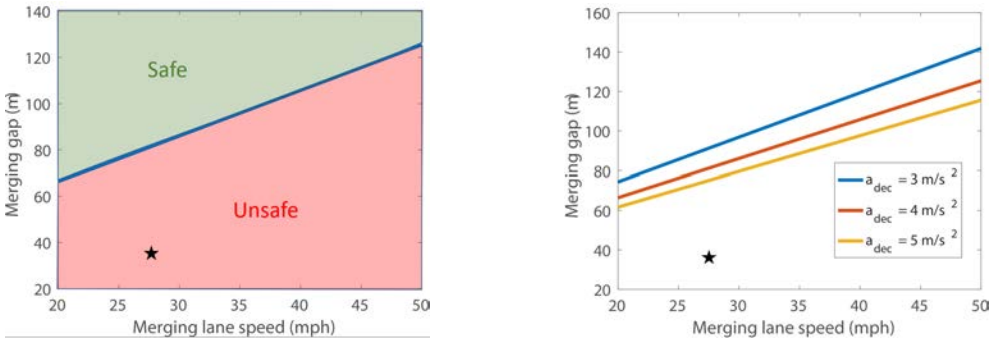


Figure 8.12: Safe merging gap as a function of merging lane speed for the time interval 8:05 AM - 8:20 AM in the NGSIM US-101 dataset. Left: The green (safe) region represents the set of (merging gap, lane speed) pairs for which an AV can merge safely in the worst case, whereas the red (unsafe) region represents unsafe pairs. The black star ($\star$) denotes the observed values of (merging gap, lane speed) for the given time period. Notice that these values are in the unsafe region.
Right: Change in the safe region boundary (as shown on the left) with varying deceleration rates. Note that the observed values from NGSIM remain in the unsafe region for all of the above deceleration rates.

sec according to the American Association of State Highway and Transportation Officials (AASHTO) design specifications based on the 95$^{\text{th}}$ percentile of empirical reaction time estimates - see Table 3 in [143]. We set $\rho_{\text{AV}} = 0.83$ sec, which is the mean of empirically observed AV reaction times - see Table 2 in [144]. Using the following values for the rest of our problem parameters: $a_{\text{acc}} = 3$ m/s^2, $a_{\text{dec}} = 4$ m/s^2 and $l_{\text{AV}} = 4$ m, we compute the safe merging gap d_{safe} for each of the time intervals as seen in Table 8.1. Moreover, Figure 12 shows how the safe merging gaps change with varying merging lane speeds and deceleration rates. It can be observed that the observed merging gaps are considerably smaller than the corresponding safe merging gaps. Thus, the worst-case safe planning approach is not feasible in this scenario. Unless the AV predicts the behavior of vehicles in the merging lane accurately, it cannot merge with guaranteed safety. Thus, behavior prediction uncertainty will indeed result in crashes.

In the above analysis, we considered the worst case of both the lead vehicle braking and lag vehicle accelerating at the same time. We now analyze the *Single Event* case in which only the lead vehicle decelerates to a stop. Observe that the lead gap required to be safe remains the same as in (8.26). However, the lag gap required will be lower in this case as the lag vehicle is not accelerating at the same time. The lag vehicle covers a distance $v_{\text{B}}\rho_{\text{B}}$ before it starts decelerating and further travels $v_{\text{B}}^2/2a_{\text{dec}}$ before it stops. Thus, the expression

for the required lag gap $d_{B,\text{safe}}$ (8.28) changes to the following in this case,

$$d_{B,\text{safe}} = \max\left\{v_{\text{B}}\rho_{\text{B}} + \frac{v_{\text{B}}^2 - v_{\text{AV}}^2}{2a_{\text{dec}}}, 0\right\}. \tag{8.29}$$

As can be seen from Table 8.1, we find that the required gaps even in this case are larger than those commonly observed in the NGSIM dataset.

Deriving estimates for crash probability in this setting is challenging as it requires a model for how vehicles react to surrounding vehicle behavior as well as a probability distribution over possible vehicle behaviors. However, the fact that behavior prediction modules are not always accurate combined with the large number of merging crashes observed each year strongly suggests that such crashes will persist with a significantly high probability.

The fundamental safety challenge in the above merging example is that it requires all the involved vehicles to be in agreement to ensure safety. In the absence of connectivity, predicting vehicle behavior and planning accordingly is the only available recourse. Introducing connectivity between vehicles (V2V) would obviate the need for such behavior prediction and as a result, ensure that the merging maneuver can be executed safely [125, 126, 127].

8.6 Investment in Connected Infrastructure

To summarize, *fully autonomous* AVs would be vulnerable to crashes due to information gaps even if they possessed perfect sensing and perception capabilities. These information gaps can be resolved by incorporating V2V or I2V connectivity in AV design. As discussed in Section 8.1, there are several deterrents to adoption of connectivity by AV companies: (i) high cost of instrumentation, (ii) delay in deployment due to reliance on adoption by other road users, and (iii) cybersecurity vulnerabilities.

At the same time, there is no free lunch as far as connectivity is concerned. Through our analysis in the previous sections, we argue that fully autonomous vehicles cannot guarantee safety in routine traffic scenarios characterized by information gaps. In order to reap the benefits of connectivity, we need to tackle the challenges associated with it. Since more than 50% of all injury/fatal crashes occur at or near intersections [145], a first step could be to focus on making I2V technology at intersections economically viable. A fully instrumented intersection can cost around $25,000 [122]. Due to limited budgets, municipalities cannot instrument all intersections in their jurisdiction. However, crashes are not uniformly spread across all intersections. As observed in our study of intersections in Torrance, California [146], a few intersections account for most of the crashes. Thus, instrumenting a few selected intersections can lead to substantial safety gains. Moreover, public-private partnerships are another increasingly popular option for funding such developments. A recent example is the partnership between Cavnue, a smart infrastructure company, and the state of Michigan to develop a 40 mile connected and automated vehicle corridor [147]. More research and investment is required to mitigate the cybersecurity challenges associated with this technology.

8.7 Concluding Remarks

Automated vehicles (AVs) have the potential to change the transportation landscape and lead us to a safer future. The thousands of lives that are lost every year due to impaired, inattentive or reckless driving could be saved if human drivers are replaced with AVs. However, a significant fraction of crashes cannot be explained by these reasons because of a fundamental aspect of driving on the roads: *Our safety depends on our actions as well as the positions and actions of neighboring vehicles – both of which might be partly unknown.* In the absence of connectivity with neighboring vehicles (V2V) or infrastructure (I2V), it is unclear whether AVs can avoid such crashes. Even so, AV companies maintain that they will eventually eliminate all crashes without relying on connectivity.

One proposed approach is to assume worst case positions and actions of neighboring vehicles and plan to be safe [121]. While this seems appealing as it circumvents the challenges associated with behavior prediction and does not require connectivity, it turns out that such guaranteed safety comes at a crippling cost to traffic efficiency. Our analysis has shown that ensuring guaranteed safety would preclude vehicles from performing basic maneuvers like merging or unprotected left-turns in common traffic scenarios. In the presence of uncertainty about neighboring vehicle positions and behavior, some crashes are indeed unavoidable.

In this chapter, we investigated three crash causes to tease out various aspects that make driving on the roads challenging: (a) Occlusions (unknown vehicle/pedestrian position), (b) Traffic Violations (rule-following assumption violated), and (c) Behavior Uncertainty (inability to accurately predict vehicle/pedestrian actions based on past observations). For each of these cases, we showed that it is impossible to guarantee worst-case safety without connectivity. We also provided some estimates for the probability of a crash in such scenarios.

Recognizing that such situations are inevitable while driving on the roads, the objective of an AV should be to manage crash risk while maintaining efficiency. Assessing crash probabilities in commonly occurring risky scenarios is an important step in this process. As we assume perfect sensing and perception capabilities for AVs and idealized scenarios for our analysis, our estimates can be considered as lower bounds for actual crash probabilities. More realistic bounds can be derived by considering imperfections in road-user behavior and AV capabilities, as well as factors such as vehicle failures, and varied road geometry and conditions. Looking through the lens of managing crash risk, AVs can be seen as a collection of Advanced Driver Assistance Systems (ADAS). Thus, empirical estimates of ADAS effectiveness can be used to arrive at better estimates of crash reduction due to AVs [118].

The crashes we have discussed result from lack of observability of neighboring vehicle positions or incorrect assumptions or predictions about vehicle actions. Communication with neighboring vehicles or infrastructure would ensure that all involved vehicles can detect each other and reach an agreement regarding each others' proposed actions [126, 123, 124, 127]. Thus, connectivity can bring about a significant reduction in otherwise unavoidable crashes. However, connectivity comes with its own safety challenges. Relying on information from other vehicles or infrastructure makes one vulnerable to security attacks from malicious road

users that can jeopardize safety. Additionally, installing sensors on the roads and ensuring that all vehicles have the required technology to communicate would require a significant amount of time and economic resources. In order to reap the safety benefits of connectivity, active research is needed to minimize its associated safety risks and economic costs. Our hope for a future with zero crashes depends on it.

Chapter 9

Automated Vehicle Risk Assessment: Modeling

In Chapter 8, we highlighted a set of routine traffic scenarios that are particularly challenging for AVs, and argued why connectivity is crucial for safely tackling them. Recognizing that AVs cannot avoid all crashes, a comprehensive risk assessment is crucial to ensure that they are "safe enough" to be deployed. In this chapter, we present a risk assessment framework that leverages human driving data, in addition to on-road testing data of AVs, to provide insights into their safety capabilities.

9.1 Introduction

Permitting deployment of AVs without an adequate safety assessment might increase crashes. Extensive on-road testing is needed to ensure that AVs bring about the intended safety benefits. However, testing AVs across all possible driving contexts is impractical. Moreover, since crashes are rare events, AV safety assessment requires comprehensive approaches that account for diversity in driving contexts without testing in all possible scenarios.

Current Approaches for AV Testing

The AV industry currently uses three approaches for safety testing: (a) simulation, (b) closed circuit, and (c) on-road. Each of these approaches has its own advantages and limitations. Evaluating an AV's safety performance via simulation allows testing over billions of virtual miles, but the simulator is only an approximation of real-world driving. Crashes being rare events, small differences in the simulated and real-world environments can translate to large errors in the inferences drawn about safety performance. Closed circuit testing aims to overcome this limitation by replicating real-world conditions and testing AVs in challenging scenarios. However, recreating the overwhelming diversity of real-world driving conditions and road-user behavior is impractical and hence, such testing provides an informative, but

incomplete appraisal of an AV's safety. On-road testing is the gold standard for predicting an AV's performance on future deployment on the roads as there is no gap between the conditions in which the AV is tested and deployed.

Several companies are now testing their AVs in limited operational design domains (ODDs) deemed safe enough for operation. The objectives of on-road testing are: (a) improving AV technology based on the experience gained from driving on the roads, and (b) demonstrating that AVs are indeed safe enough to be deployed on a large scale. While AV companies are achieving the first objective, the path to the second seems less clear. Several works argue that AVs cannot demonstrate their safety capabilities solely based on on-road testing [148, 149, 121]. In particular, a study by the RAND corporation [148] suggests that AVs would have to be tested over billions of miles to demonstrate a fatality rate lower than that of human drivers. It would require decades of testing for current AV fleets to cover such a distance.

Not all Miles Driven are Equal

The RAND study [148] solely uses crashes per mile as the safety metric for comparing AVs and human drivers. However, since AVs are currently being tested only in limited ODDs, they are not exposed to the wide variety of driving conditions and road user behavior that human drivers come across in billions of miles of cumulative driving every year. Driving millions of miles on sparse rural highways is very different from covering the same distance in dense urban settings. Clearly, the context in which these miles are driven is crucial for interpreting the results of on-road testing. Thus, crashes per mile is an insufficient metric for judging AV safety.

This raises the question: *How should we assess the crash risk of AVs?* We have argued so far that more information than just the total number of crashes and miles driven is needed. Since the promise of AVs is to make our roads safer, safety metrics must reflect the efficacy of AVs across all driving contexts encountered by humans. Our risk assessment framework for AVs should be such that successful testing on sparse rural highways does not imply that they would be safe while navigating dense urban intersections. Rather than having a single performance metric for assessing risk, it is desirable to have a collection of metrics that capture safety performance across diverse contexts. In order to do so, we must first come up with a suitable level of abstraction at which the diversity of driving contexts can be expressed in a tractable manner.

High Variance in Crash Risk

The crash typology presented by NHTSA in [106] provides a useful starting point. This typology classifies crashes into 36 types based on the maneuvers of involved vehicles leading up to the crash. This typology is motivated by the observation that there is significant variance in crash risk across driving maneuvers. For example, left turns result in twice as many crashes as right turns. This suggests that analyzing performance across diverse driving

maneuvers would be an important first step towards an effective risk assessment framework for AVs.

Another important source of diversity is the environment in which maneuvers are performed. A lane change executed on urban highways with small traffic gaps has a much higher crash risk compared with the same maneuver on sparse rural highways with larger gaps. Foggy weather and slippery roads can also increase crash risk. Thus, the diversity of driving environments is another factor that needs to be captured in our risk assessment framework.

Risk Assessment: A Delicate Balance

Risk assessment of AVs is of particular interest to regulators who seek to balance the safety benefits of this promising technology and its potential pitfalls. Failing to impose the required safety standards before deploying AVs can prove fatal. For instance, the National Transportation Safety Board concluded that the absence of regulatory oversight contributed to the fatal crash involving an Uber AV in Tempe, Arizona [114]. Currently, states are free to set their own standards for AV testing and deployment. To the best of our knowledge, there are no standardized safety performance metrics used by regulators to govern the deployment of AVs in their jurisdiction. The kinds of testing data should be disclosed by AV companies to regulators for risk assessment is an open question.

9.2 Summary of Contributions

In this chapter, we aim to address the following question:

> *How should the diversity of driving maneuvers and environments be taken into account in the crash risk assessment of AVs?*

Our main contributions are as follows:

- We develop a crash risk assessment framework for AVs that provides insights about an AV's safety performance across diverse driving contexts.
- We derive similar crash risk estimates for an average human driver using police crash reports and fine-grained vehicle maneuver data. Since AVs are expected to be significantly better than their human counterparts, these estimates can be used as a baseline to interpret AV safety performance.
- We present several use cases of our risk assessment framework for AV companies and regulators. In particular, we show how AV companies can estimate crash risk over alternative paths so as to make safety-informed routing decisions. We also discuss how policy makers and regulators can use our framework to analyze the safety costs of AV deployment in their jurisdictions.

9.3 AV Crash Risk Assessment

9.3.1 The Challenge of Risk Assessment

Consider an AV making an unprotected left turn at an urban intersection. Suppose we wish to estimate the probability of the AV getting into a crash. Several obstacles could prevent the AV from making this turn safely: (a) through-moving traffic in the opposite direction with a conflicting trajectory, (b) pedestrians on the crosswalk while the AV makes the left turn, (c) weather conditions that affect visibility and control for both the AV and neighboring vehicles (e.g., foggy weather, smoke, night time), and (d) uncertainty in the behavior of neighboring vehicles/pedestrians.

An accurate estimate of the AV's crash probability would have to take all the above factors into account. First, we would need an accurate model of vehicle and pedestrian arrivals at the intersection. Second and arguably more difficult, is the task of probabilistically modeling the response of oncoming traffic to the AV's left turn maneuver. This includes the time taken to detect the AV (reaction time), their decision process (to slow or not to slow down) and possibly even evasive maneuvers (e.g., brake, swerve) if the neighboring vehicle was too close to the intersection. Adverse weather conditions would complicate the situation further by modifying the behavior of each road user.

9.3.2 Quantifying Crash Risk

Arriving at an accurate estimate of the AV's crash probability in this specific scenario using an agent-based model of road user behavior seems intractable. This motivates the use of data-driven statistics to describe crash risk. A common statistic used to describe AV safety performance is crashes or disengagements per mile driven.[1] The appeal of this safety metric lies in its simplicity; all that needs to be known is the number of crashes or disengagements encountered by AVs and the total number of miles driven. Moreover, it provides a seemingly straightforward solution for comparing AV safety performance with that of humans. For instance, the crash rate for human driven vehicles in the US is 1.9 per million miles driven [106]. Considering that AVs have already been tested on the roads over millions of miles, it seems reasonable to compare AV crashes per million miles to the above number.

However, not all miles driven are equal from a traffic safety standpoint. Crash rates differ significantly depending on the type of maneuvers performed while driving. As a case in point, let us compare two maneuvers: (a) lane-keeping, and (b) making a left turn. Rear-ends are the most common cause of crashes while lane-keeping, and can occur anywhere on the lane. As rear-ends account for 30% of all crashes [106], we can conclude that there is a probability of 6.3×10^{-7} of getting into a rear-end crash for every mile of lane-keeping. On

[1] California DMV requires AV companies to report how often their vehicles disengaged from autonomous mode during tests because of technology failure or situations requiring the test driver to take manual control of the vehicle to operate safely. Since each AV company has its own rules to determine when a disengagement is warranted, the disengagement rates among companies are not comparable.

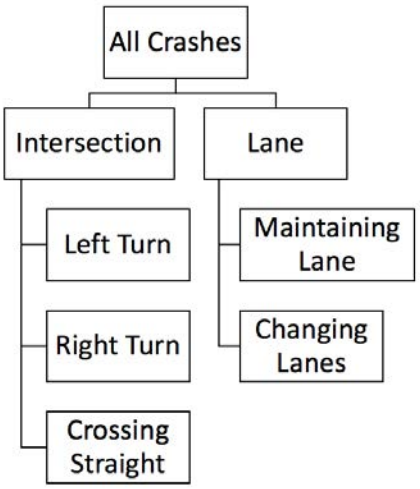


Figure 9.1: Decomposition of all crashes based on maneuvers of vehicles involved.

the other hand, left turn crashes can only occur on a short stretch of road leading up to an intersection. A typical left turn involves covering a distance of less than 20 m. It is reasonable to assume that human drivers make less than one left turn per mile on average across all driven miles. Considering that 7% of all crashes involve left turns [106], we can conclude that the crash probability per left turn mile must be higher than 1.1×10^{-5}. Therefore, even though rear-ends account for 4 times as many crashes as left turns, this argument suggests that the per-mile crash risk during left turns is 170 times larger than that while lane-keeping. Thus, it is clear that the type of maneuver matters while assessing crash risk.

The total number of miles driven alone cannot inform us about the frequency of different types of maneuvers performed. It would be fair to compare AV performance with the human baseline as long as they are tested across all the diverse maneuvers and environments that humans drive in. However, this is far from the case for AVs at present, which have only been tested in environments deemed safe enough for operation. Driving crash-free over millions of miles of rural highways has very different safety implications compared to doing the same in dense urban settings. Thus, focusing on the crashes per mile safety metric can lull us into a false sense of security about the real safety capabilities of AVs.

9.3.3 Maneuver Level Crash Analysis

We saw in the discussion above how pooling together all of driving in the bucket of total miles driven hides away details about an AV's performance over diverse scenarios that are crucial to assess its crash risk. A first step in incorporating such diversity would be to condition driving based on the type of maneuver. The process of driving can be classified into two broad maneuver groups: (a) intersection, and (b) lane maneuvers, as shown in Figure 9.1[2]. The

[2]This is not an exhaustive list of all maneuvers performed while driving. For simplicity of analysis, we do not consider maneuvers such as U-turns, backing and navigating roundabouts which are less frequently

former includes such maneuvers as turns and going straight that are related to navigating through an intersection, while the latter consists of lane-related maneuvers such as staying in lane and changing lanes. The frequency of performing these maneuvers depends crucially on the driving environment. Intersection maneuvers are much more common in urban driving as opposed to freeways. Thus, estimating crash risk for specific maneuvers would enable us to put the safety statistics of AVs into context and provide better insights into whether their performance is likely to generalize to regions they have not been tested in before.

9.4 Modeling Crash Risk

9.4.1 A Simple Maneuver-Level Crash Risk Model

We start by developing a simple model to estimate crash probabilities for specific maneuvers. Suppose each vehicle has a probability p_m of getting into a crash during maneuver type m. Then, the number of crashes C_m over N_m maneuvers of type m is given by

$$C_m \sim \text{Bin}(N_m, p_m), \tag{9.1}$$

where $\text{Bin}(N, p)$ denotes the binomial distribution with N trials. The maximum likelihood estimate for p_m in this case is the empirical crash probability,

$$\hat{p}_m = \frac{C_m}{N_m}. \tag{9.2}$$

Note that the above estimate is similar in spirit to the crashes per mile statistic. The main differences are the conditioning based on maneuver type as opposed to considering all crashes to be equivalent, and the normalization of crashes by number of maneuvers rather than total miles driven.

Remark 9.1 *As discussed in Section 9.1, the driving context includes factors other than maneuver type such as intersection geometry, weather conditions, and traffic density. Let θ represent the set of all such factors that constitute a driving context. Let p_θ denote the probability of a crash in context θ. Then, $X_\theta \sim \text{Bernoulli}(p_\theta)$ is an indicator random variable representing whether a crash occurs during context θ. The number of crashes C_m involving maneuver type m over K maneuvers is given by $C_m = \sum_{k=1}^{K} X_{\theta_k} \mathbb{I}\{m \in \theta_k\}$, where θ_k is the driving context during the k^{th} maneuver and $\mathbb{I}\{m \in \theta_k\}$ is the indicator variable denoting whether maneuver type m is performed in context θ_k. The Binomial model (9.1) can be viewed as an approximation of this general formulation with $p_m = \mathbb{E}[p_\theta | m \in \theta]$, where the expectation is over the driving context θ.*

observed. Nevertheless, the vast majority of all crashes involve one of the maneuvers listed.

9.4.2 Comparing Human and AV Crash Risk

Waymo's recent safety report [116] described the safety performance of its AVs over 6.1 million miles of driving in Maricopa County, Arizona. Notably, it is the first company to provide a comprehensive breakdown of crashes involving its AVs based on maneuver type. Even so, this does not allow us to estimate crash probabilities for maneuvers such as left turns or lane changes since data on the frequency of such maneuvers remains unavailable. However, the crash probability per mile of lane-keeping can be computed based on the given data. This enables us to compare AV and human crash risk for this maneuver.

Four types of crashes can occur during lane-keeping: (a) Rear-end, (b) Other vehicle merging into lane, (c) Road departure, and (d) Head-on with vehicle in opposite direction. Together they accounted for 25 of the Waymo's 47 crashes. Since an overwhelming majority of all miles driven involves the lane-keeping maneuver, the empirical probability of an AV crashing during this maneuver is $25/(6.1 \times 10^6) = 4.1 \times 10^{-6}$ per mile driven. In order to derive a similar estimate for human drivers, we need to know the number of such crashes and the total vehicle miles travelled in the same region. There were 97,105 crashes in Maricopa County over 37.9 billion miles driven in 2019 [150]. Since 70% of all crashes occur while lane-keeping [106], the human crash probability during the lane-keeping maneuver is $(97,105 \times 0.7)/(37.9 \times 10^9) = 1.8 \times 10^{-6}$ per mile. One caveat here is that the human crash statistics are based on police-reports. However, a study by NHTSA [95] estimated that 40% of all crashes go unreported to the police. Accounting for this, the human crash probability estimate turns out to be 3.0×10^{-6} per mile, about 30% lower than the AV crash probability estimate. Note that the above crash risk estimates for AVs and human drivers may not be directly comparable since the miles driven by Waymo AVs are not necessarily representative of typical miles driven by human drivers in Maricopa County. This motivates the need for additional information about the contexts in which AVs drove these miles.

9.4.3 Data Required for Estimation

For estimating maneuver-level crash probabilities based on (9.2), two quantities must be known:

- Number of crashes involving maneuver type m,
- Frequency of maneuver type m.

At present, these quantities have not been made publicly available by most AV companies. Regulations for on-road testing data disclosure vary based on jurisdiction. While California requires AV companies operating in the state to report the circumstances leading to each disengagement, the state of Arizona has no such requirement. Even in California, companies are required to report only the total number of miles driven, not the frequency of maneuvers their AVs were involved in. Moreover, AVs have been tested in very limited on-road environments. AV companies do not have data on how their AVs function across diverse regions

and road conditions that humans drive in. In the absence of this data, it is not possible to estimate maneuver-level crash probabilities using (9.2).

9.5 Leveraging Human Driving Data

Even if the required data becomes available to compute crash risk estimates, there is still a need for baselines to interpret them. The human crash risk for the same maneuvers is an intuitive benchmark for this purpose. Moreover, even though AVs might drive differently compared to humans, their crash risk is likely to be correlated with that of humans due to intrinsic challenges associated with driving on the roads. As a case in point, the left turn example discussed in Section 9.3.1 is challenging for both humans and AVs. For maneuvers and road conditions in which AVs do not have adequate driving experience, the corresponding crash risk for humans is a useful indicator of the AV crash risk.

While on-road testing of AVs is at a nascent stage, humans drive billions of miles on the roads every year across a multitude of road conditions. As a result, much more data is available about human driving performance. Two sources of data are of particular interest in the estimation of maneuver-level crash risk: (a) crash reports, and (b) fine-grained vehicle maneuver data.

9.5.1 Crash Reports

Police reports filed in the aftermath of a crash provide detailed information about the events leading to a crash. As a part of the broad trend towards making governmental data openly accessible, many jurisdictions in the US have uploaded their historical crash data online in an easy-to-query format. We use one such data source, the Transportation Injury Mapping System (TIMS) [137], which is a web-based application that presents crash records for the state of California in a map-centric format. This enables a deeper understanding of traffic safety in a particular jurisdiction beyond just the total number of crashes. For instance, it allows us to query the number of crashes involving a vehicle making a left turn colliding with a pedestrian at a particular intersection over a specified time period.

9.5.2 Fine-grained Vehicle Maneuver Data

As noted in Section 9.4.3, knowledge of crash data alone is not sufficient to estimate maneuver-level crash risk. The frequency of vehicle maneuvers is crucial to put the total number of crashes into context. While vehicle flow and speed are commonly measured for traffic planning purposes, fine-grained maneuver level data such as the number of left turns or lane changes on a particular stretch of road is typically unknown. Recent advancements in technology have led to the emergence of companies collecting GPS traces of vehicles. This has enabled the measurement of previously inaccessible maneuver level data at scale without having to rely on expensive instrumentation of traffic infrastructure. We utilize a trace data

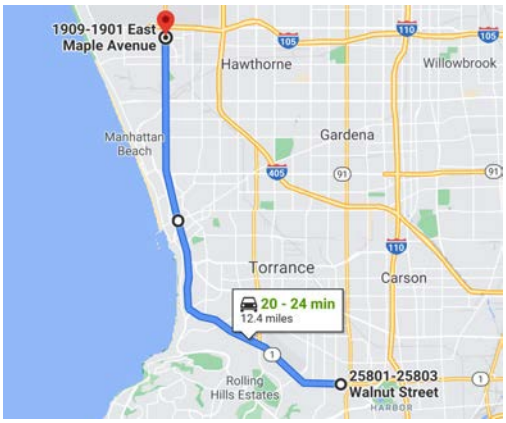


Figure 9.2: Region of study: 12.4 mile stretch of Pacific Coast Highway in Torrance, California (Source: Google Maps).

set provided by Wejo Ltd., a connected vehicle data collection company, for a 12.4-mile stretch along the Pacific Coast Highway in Torrance, California which includes 29 major signalized intersections (shown in Figure 9.2). The data set provides information about (location, speed, heading, accelerations/brakes) of each vehicle at 3 second intervals. These traces were collected from General Motors (GM) vehicles that were equipped with an *enhanced* GPS device with higher accuracy ($\pm$ 1.5 m) that enables measurement of location with lane-level precision. We also employ a data set provided by Sensys Networks Inc. containing measurements by in-ground sensors at intersections along the same stretch. This allows us to extract lane-level information about vehicle flow at 29 intersections along the Pacific Coast Highway.

Using crash reports, vehicle trace data, and vehicle detection measurements, we can estimate maneuver-level crash probabilities for an average human driver.

9.5.3 Deriving Crash Risk Estimates

Suppose we wish to estimate the probability of a crash while making a left turn in the region illustrated in Figure 9.2. Let C_{lt} denote the number of left turn crashes in this region between January 2011 and December 2019, i.e., $T = 9$ yrs. Let R_{lt} denote the left turn rate in this region. Using (9.2), the probability of a left turn crash on the Pacific Coast Highway is estimated as

$$\hat{p}_{\text{lt}} = \frac{C_{\text{lt}}}{R_{\text{lt}} T}. \tag{9.3}$$

Plugging in $C_{\text{lt}} = 115$ and $R_{\text{lt}} = 1868.4$/hr based on crash reports and vehicle maneuver data for this region, we have $\hat{p}_{\text{lt}} = 7.8 \times 10^{-7}$. Figure 9.3 illustrates the number of crashes and estimated crash probability across the maneuvers discussed in Section 9.3.3. Observe that there is significant variation in the estimated crash probability across maneuvers. Although

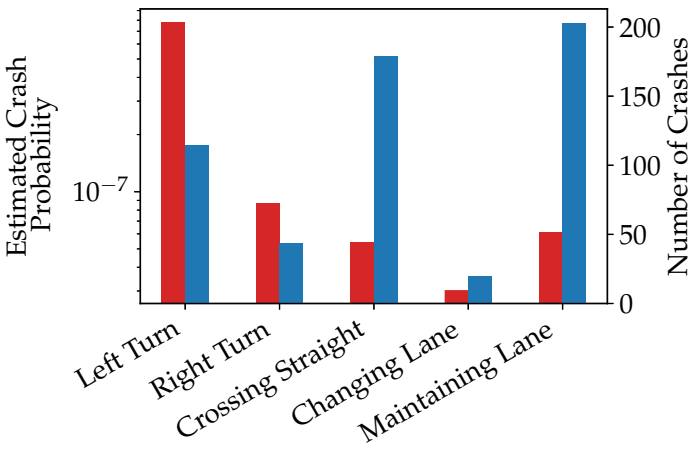


Figure 9.3: Estimated crash probability (red) and number of crashes (blue) across diverse maneuvers. While lane-keeping is associated with the most number of crashes, left turns have the highest crash probability.

Intersection	Left Turn Rate (turns/hr)	Empirical Crash Probability ($\times 10^{-8}$)	95% Confidence Interval ($\times 10^{-8}$)
Diamond Street	39.9	670	(22, 280)
8^{th} Street	15.1	590	(260, 1300)
Rolling Hills Way	46.2	160	(67, 380)
Calle Mayor	130.1	9.8	(0.51, 63)
Palos Verde Boulevard	136.0	9.3	(0.49, 61)
Prospect Avenue	30.2	0	(3.8, 200)

Table 9.1: Confidence Intervals for left-turn crash probability at selected intersections with high (Diamond, 8th, Rolling Hills) and low (Calle, Palos Verdes, Prospect) empirical crash probabilities. Observe that despite Prospect Ave. having zero crashes, its upper confidence bound is higher than that of Palos Verdes Blvd. since it has a considerably lower left turn rate.

most crashes occur while lane-keeping, left turns have the highest crash probability among all maneuvers. This underscores the importance of incorporating frequency of maneuvers for interpreting crash statistics.

9.6 Refined Model

9.6.1 Diversity of Environment

In previous sections, we discussed how crash probabilities can vary significantly across maneuvers. From human driving data, we find that there is considerable variance in crash risk across locations even for the same maneuver. Consider the Diamond Street and Hawthorne Boulevard intersections along the Pacific Coast Highway in Torrance, California. The estimated probability of a crash while making a left turn at Diamond Street is 6.7×10^{-6}, whereas it is 1.7×10^{-7} at Hawthorne Boulevard. In other words, the average human driver making a left turn at Diamond Street is more than 20 times as likely to get into a crash as compared with the same maneuver at Hawthorne Boulevard. Thus, it is important that the environment in which a maneuver is performed is taken into account while assessing crash risk. One way to do this is to model crash risk for (maneuver, intersection) pairs rather than the maneuver alone as in (9.2). Consider a vehicle performing maneuver type m at intersection i. Then, the model in (9.1) can be modified to

$$C_m^i \sim \text{Bin}(N_m^i, p_m^i), \tag{9.4}$$

and the associated maximum likelihood estimate for p_m^i is given by

$$\hat{p}_m^i = \frac{C_m^i}{N_m^i}. \tag{9.5}$$

Such crash probabilities for (maneuver, intersection) pairs can be computed at scale for human drivers using the data sources discussed in Section 9.5. Note that in order to derive meaningful conclusions, these estimates need to be computed over millions of maneuvers at each intersection. While this is true for human driving data which has been collected across millions of drivers over decades, it would take hundreds of years of on-road testing by AV companies to arrive at statistically valid estimates. A potential solution to tackle this dearth of testing data is to model crash risk for broad classes of intersections rather than modeling each intersection individually. We present an example of such a classification based on human crash risk in Section 10.1.1.

9.6.2 Uncertainty about Estimate

While the empirical crash probability estimate in (9.5) can be used to assess risk, it has two major drawbacks. First, it is uninformative for intersections with zero crashes. Moreover, being a point estimate, it does not quantify the degree of confidence associated with the estimate. For instance, we should be less confident about our estimate for an intersection with limited crash and flow data. One way to circumvent both drawbacks is to use confidence intervals of p to assess risk in addition to the estimated crash probability $\hat{p}$. Let $[\hat{p}_{\min}, \hat{p}_{\max}]$ denote the 95%-confidence interval for p based on the available data. Even for intersections having zero crashes, we can be reasonably confident that $\hat{p}_{\max}$ upper bounds the actual

crash probability p. We present such confidence intervals for intersections with high and low empirical left turn crash probabilities in Table 9.1. We use the Wilson score with continuity correction method [151] for deriving these confidence bounds. Notice that even though Prospect Avenue has zero crashes, its upper confidence bound is higher than that of Palos Verdes Boulevard since it has a considerably lower left turn rate.

9.6.3 Context of Maneuver

The crash risk estimates derived so far answer the following question: *What is the probability of a crash while performing maneuver type m at location i?* As such, these estimates do not rely on additional information such as the locations and behavior of neighboring vehicles and pedestrians, traffic density, weather conditions and time of day. Such factors that inform the context in which a maneuver is performed can have a significant influence on crash risk. For example, the presence of occlusions or the observation of risky behavior by neighboring vehicles is an indicator of heightened risk of a crash. Note that at present, information about the position and behavior of neighboring vehicles and pedestrians is not available across all (maneuver, location) pairs. Nevertheless, with increasing installation of cameras at intersections and collection of vehicle and pedestrian trace data, it is plausible that such information would become available in the near future. Moreover, AV companies can already extract such information from on-road testing data. This additional information can be used to refine the estimated crash probabilities based on the specific contexts in which maneuvers are performed.

As an example, consider an AV making an unprotected left turn in the context of its field of view of oncoming traffic being occluded due to a queue of vehicles in the opposing lane. Through on-road testing data, AV companies have access to the number of left turns made in such a context and the resulting number of crashes. This can be used to calculate the empirical probability of a crash in this particular context.

Chapter 10

Automated Vehicle Risk Assessment: Applications

In the previous chapter, we introduced a risk assessment framework for automated vehicles in diverse driving contexts. We now discuss how this framework can inform safe AV deployment. We begin by presenting several use cases for the risk assessment framework developed in previous sections.

10.1 Use Cases

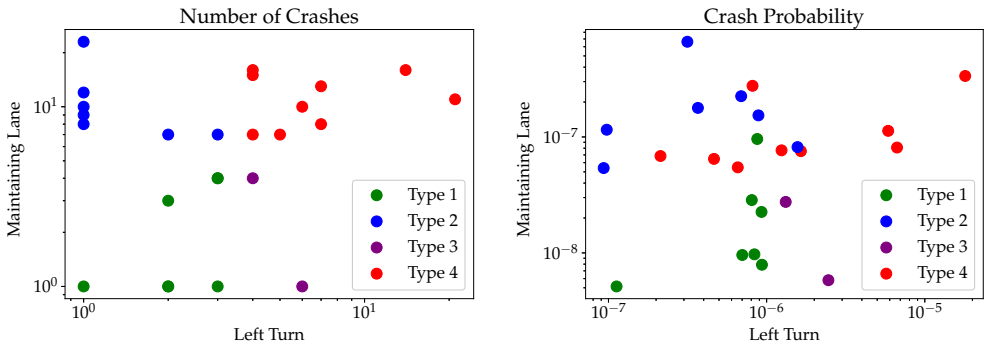


Figure 10.1: Number of crashes and estimated crash probability for intersections along the Pacific Coast Highway (log scale). Each dot represents an intersection. Types 1-4 refer to the classification of intersections based on the number of left turn and lane-keeping crashes occurring at these locations. For eg., Type 2 (blue) intersections have a low number of left turn crashes, but a high number of lane-keeping crashes.

10.1.1 Classifying Intersections based on Crash Risk

As discussed in Section 9.6.1, estimating crash risk for individual intersections solely through on-road testing is impractical for AVs. Therefore, an intermediate level of abstraction for environment diversity is required that navigates the trade-off between accuracy of crash risk estimates and data scarcity. A candidate abstraction is the classification of intersections based on maneuver level crash risk. Such a classification can inform the choice of testing locations and enable AV companies to enhance their crash risk assessment through aggregation of testing data over each class. We provide one such taxonomy for the intersections along the Pacific Coast Highway in Torrance, California. As seen from Figure 9.3, lane-keeping and left turn are the maneuvers with the highest crash probability. Let us first consider a taxonomy of intersections based on the number of left turn and lane-keeping crashes, as shown in Figure 10.1. Observe that Type 4 intersections (red) have a high number of crashes of both types and hence, can be considered as hazardous based on this metric. The crash probabilities for the same intersections are illustrated on the right in Figure 10.1. This plot paints a slightly different picture. Observe that there are Type 4 intersections with relatively low probabilities for both types of crashes. Thus, intersections considered hazardous based on the number of crashes metric may not be so in terms of crash probability. This is particularly true for some busy intersections that witness a high number of crashes but have a low crash probability after normalizing by vehicle flow. The difference between the two metrics is subtle but important. Traffic planners typically rely on the former metric for identifying hazardous intersections as they aim to minimize the total number of crashes in their jurisdiction. On the other hand, crash probabilities are more relevant for AV fleets as they aim to minimize their own crash risk rather than the aggregate crash risk over all road users. Thus, a taxonomy of intersections based on crash probability is more suitable for incorporating environment diversity in AV risk assessment; as discussed in Section 9.6.1, this would enable AVs to aggregate their performance in a context-aware manner based on the above taxonomy. Note that the above example is only one of many ways in which environment diversity can be captured based on crash risk.

10.1.2 Route Risk

While it is conceivable that AVs will eliminate a significant fraction of crashes with improvements in technology, some types of crashes are likely to persist [20, 108]. As a result, AVs need to be able to mitigate risk while driving on the roads. One way to do this is to consider not just delay but also crash risk in its route planning stage. We now discuss how an AV can go about estimating this risk.

A route can be expressed as a sequence of (maneuver, location) pairs $R = \{r_t\}_{t=0}^{L}$, where $r_t = (m_t, i_t)$ denotes the t^{th} pair in the sequence. Let P_t represent the crash probability estimate at the t^{th} step in the sequence. Then, the probability of a crash occurring over the

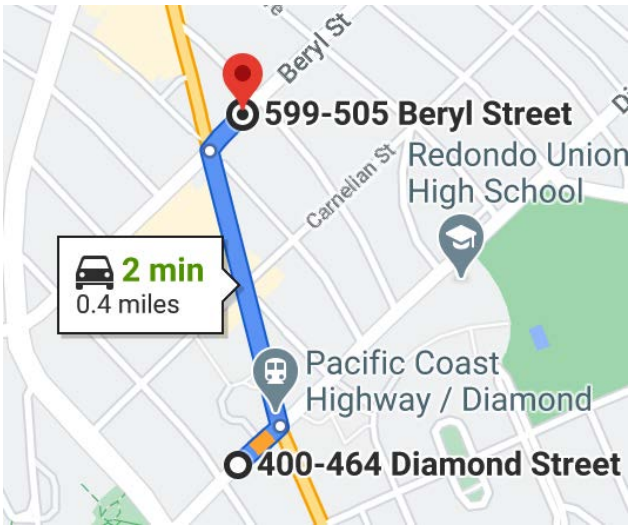


Figure 10.2: Route risk example: Diamond Street to Beryl Street along Pacific Coast Highway (Source: Google Maps).

entire route is given by

$$P(\text{Crash along route } R) = 1 - \Pi_{t=0}^{L}(1 - P_t), \tag{10.1}$$

$$\approx \sum_{t} \hat{p}_{m_t}^{i_t}. \tag{10.2}$$

Let us compute this route risk estimate for an average human driver along the route shown in Figure 10.2. This route involves the following sequence of maneuvers:

1. Left turn from Diamond St. onto Pacific Coast Highway,
2. Staying in lane between Diamond St. and Carnelian St.,
3. Going straight through the Carnelian St. intersection,
4. Lane change leading up to Beryl St. intersection,
5. Right turn from Pacific Coast Highway onto Beryl St.

Plugging in the crash probability estimates for each of the maneuvers, we estimate a crash probability of 6.8×10^{-6} for the given route.

10.1.3 Economic Cost of Crashes

Although the terms crash probability and risk have been used interchangeably in the preceding sections, crash probability is not the only risk metric of interest. In particular, the

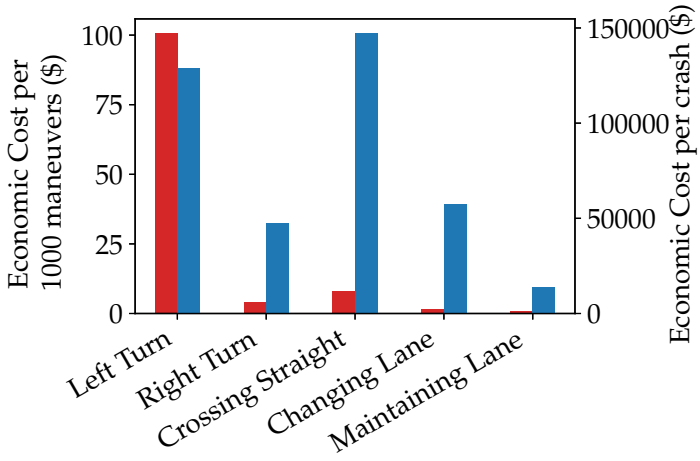


Figure 10.3: Economic cost of crashes per 1000 maneuvers (red) and per crash (blue).

business case for AV fleets being used for ride-hailing and delivery services requires an understanding of the economic costs of crashes. A maneuver level analysis is critical for such an understanding. For instance, although rear-ends are the most common type of crashes, crossing path crashes have the highest economic cost since they cause about twice the damage as rear-ends on average [106]. The crash probability estimate derived in (9.2) can be used to estimate the corresponding economic cost. Let e_m denote the average economic cost of a crash involving maneuver type m. Then, the expected economic cost of a crash involving maneuver type m is given by

$$\hat{E}_m = \hat{p}_m e_m. \tag{10.3}$$

Let us compare the expected economic cost of left and right turns along the stretch of Pacific Coast Highway shown in Figure 9.2. Using (10.3) and estimates for economic cost of crashes from [106], we find that the expected economic cost over 1000 left turns is $100.8, while that for a right turn is $4.2. The huge difference in the economic costs of these two seemingly similar maneuvers suggests that careful route planning can provide substantial safety benefits. This also explains why fleet operators such as UPS plan routes that avoid left turns [130]. Estimates for the expected economic cost for other maneuvers are presented in Figure 10.3. Although the crossing straight maneuver has the highest economic cost per crash, left turns have the highest expected economic cost per maneuver since their crash probability is significantly higher.

10.2 Deploying Automated Vehicles: Policies and Regulation

As the deployment of AVs proliferates, policy makers and regulators need to grapple with the challenge of balancing the risks associated with this burgeoning technology and its potential

to make our roads safer. While AVs can eliminate prevailing crashes involving overt human error such as impaired or reckless driving, faulty technology can cause crashes that wouldn't have occurred otherwise. Since errors made by AVs can result in loss of life, regulators need to make sure that all the necessary checks are carried out before allowing deployment. As a case in point, the National Transportation Safety Board concluded that the absence of regulatory oversight contributed to the fatal crash involving an Uber AV in 2018 [114]. At the same time, certifying that AVs have a fatality rate significantly lower than the average human driver would require billions of miles of on-road testing [148]. However, delaying the introduction of AVs until they are certified safe would result in thousands of traffic fatalities that could have been potentially saved with this technology [152]. We now discuss how the lessons learnt from previous sections can inform policies and regulations governing AV deployment.

10.2.1 Testing Data Disclosure

AV companies collect vast amounts of data during on-road testing. Analyzing this data can provide insights about the current safety capabilities of AVs, which in turn can inform policies regarding their deployment. Regulations regarding data disclosure vary significantly across states. While states like Arizona and Florida do not require any data disclosure, California requires AV companies to disclose every disengagement their fleets were involved in. Most companies are opposed to this requirement [153, 154]. Note that apart from disengagements, companies are only required to report the number of miles driven by their fleets. One cause of concern is that this leads to a comparison of disengagements per mile across companies without any context about the environments in which these miles were driven. This creates perverse incentives for AV companies involved in testing. For instance, companies can test their fleets only on highways in rural California during off-peak hours in order to report a lower disengagement rate compared to other companies that test their AVs in the busy streets of San Francisco. Moreover, what constitutes a disengagement is at the discretion of the company. A focus on disengagement rates incentivizes companies to increase the danger threshold at which they declare a disengagement and hand off control to the safety driver [155]. This jeopardizes the safety of other road users.

Thus, regulators need to take into account the diversity and frequency of contexts encountered by AVs to have a well-informed opinion about their safety capabilities. As discussed in Section 9.4.3, one approach along these lines is to have companies report the frequency of maneuvers and the environments in which they were performed. This allows an apples-to-apples comparison between the performance of various AV companies. Additionally, such a summary of AV safety performance can be complemented with a context-aware analysis of human crashes. This would enable an informed assessment of AV safety capabilities in various driving contexts and how they compare with human drivers. Moreover, it allows us to forecast how AV performance in specific ODDs will translate to jurisdictions where they haven't yet been deployed.

10.2.2 Permitting AV Deployment

Recently, several jurisdictions have allowed the deployment of AVs in specific ODDs. Since AVs will be driving without safety drivers in such regions, faulty technology can prove to be fatal. Thus, regulators need to ensure that due diligence is done before they permit AV companies to operate on their roads. Regulators can gain insights about the safety capabilities of AVs by analyzing on-road testing data. This would inform them about how effective AVs are in safely navigating through a diverse set of traffic scenarios such as left turns at a busy intersection or changing lanes with narrow vehicle gaps. An analogy with human driving tests is instructive here. During a driving test, drivers are required to demonstrate their ability to perform a diverse set of maneuvers such as parking, lane changes and turns. As discussed in Section 9.6.1, there is substantial variation in crash risk across environments even for the same maneuver. Thus, a "driving test" for AVs should involve an assessment of their ability to perform a host of maneuvers at locations known to be hazardous for humans. Such an assessment can be done so long as adequate context is provided in the testing data submitted to regulators, as described in 10.2.1. Allowing AVs to be tested and deployed without a safety driver necessitates greater caution. The rarity of crashes as a fraction of vehicle miles travelled suggests that thousands of miles of testing in a particular region is insufficient to be completely assured about an AV's safety prowess. A risk mitigating strategy here is to undertake a phased deployment of AVs from "easy" to "difficult" regions in the jurisdiction. This will ensure that obvious errors in the technology pipeline are caught early on without inflicting too much damage to other road users. As an example, intersection-related maneuvers are known to cause difficulties for AVs [111, 112]. Thus, it makes sense to start deployment on highways during off-peak hours and gradually move towards allowing AVs to operate near busy intersections.

10.3 Concluding Remarks

Automated vehicles hold great promise for leading us to a future with near-zero crashes. At the same time, deployment of AVs without adequate assessment of their safety capabilities might lead to an increase rather than reduction in crashes. Extensive on-road testing is crucial to ensure that AVs bring about the intended safety benefits.

However, testing AVs in all possible driving contexts faced by humans is impractical. Coupled with the fact that crashes are rare events, this necessitates alternative approaches for evaluating AV safety that account for diversity in driving contexts, yet do not require testing over all possible scenarios. We presented one such risk assessment framework that uses data from human driving, historical crash reports, and on-road testing of AVs to provide insights into their safety performance across diverse maneuvers and environments. We derived human crash risk baselines that can be used to interpret an AV's safety performance over the same maneuvers and environments. Several applications for this framework were also presented.

Our risk assessment framework can be used to inform regulations governing the test-

ing and deployment of AVs. In particular, it suggests that regulators should require AV companies to disclose data not only regarding crashes but also the frequency of the various maneuvers their AVs performed as well as the context in which they were performed. Our framework is just one of the many different ways in which diversity of driving contexts can be accounted for in crash risk assessment. We hope that our work motivates further research along this thread.

Appendix A

Sensitivity Analysis for Results in Chapter 4

In Section 3.3, we derived the model parameters $\{\lambda^0, e^{\mathrm{p}}, N^0 e^{\mathrm{d}}, \alpha M\}$ based on the historical trip data $\{\hat{\lambda}, \hat{N}, \hat{p}, \hat{w}\}$ satisfying first-order optimality conditions. We now analyze the sensitivity of our results to changes in these model parameters. We perturb each of the model parameters by $\pm 20\%$, keeping other parameters fixed to their original estimates.

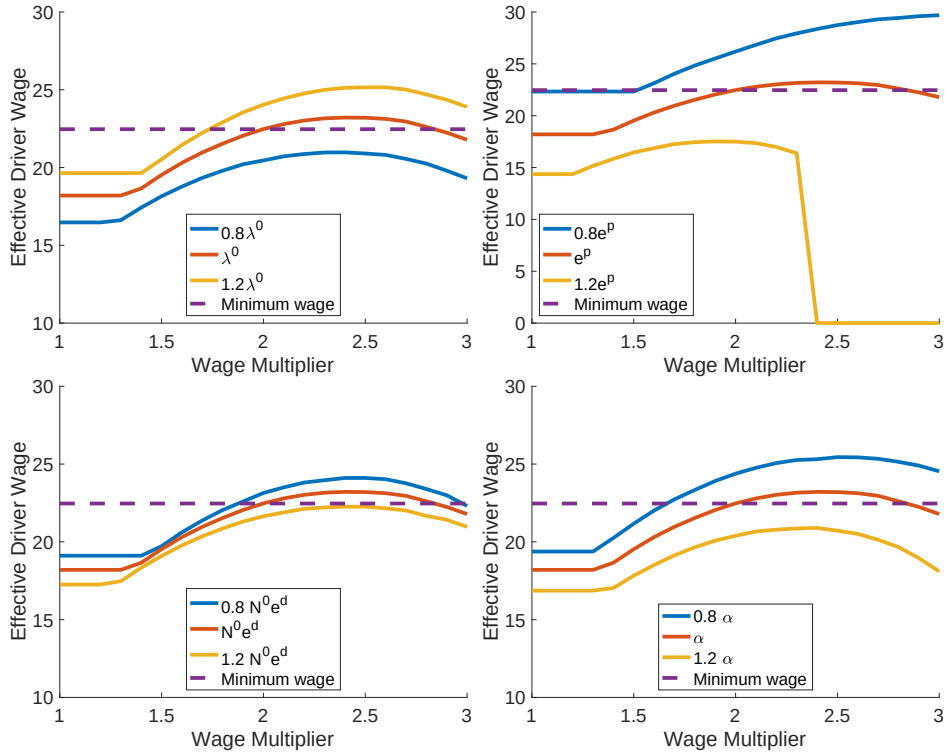


Figure A.1: Sensitivity analysis for effective driver wage under varying engaged time wage multiplier.

We find that the engaged time wage constraint (4.4c) is not active at Prop 22's proposed wage multiplier $\phi = 1.2$ across all cases, and the effective driver wage attains its maximum at an intermediate value of the wage multiplier. Moreover, for $\{0.8\lambda^0, 1.2e^{\mathrm{p}}\}$ (passenger demand lower than estimated), $\{1.2N^0 e^{\mathrm{d}}\}$ (driver supply higher than estimated), and $\{1.2\alpha\}$ (higher perceived cost of waiting time), the prevailing minimum wage cannot be attained under any wage multiplier. The sharp drop in the effective driver wage for case $\{1.2e^{\mathrm{p}}\}$ (low passenger demand) is explained by the platform no longer being profitable with a wage multiplier $\phi \geq 2.4$, which results in $w = 0$.

Appendix B

Sensitivity Analysis for Results in Chapter 5

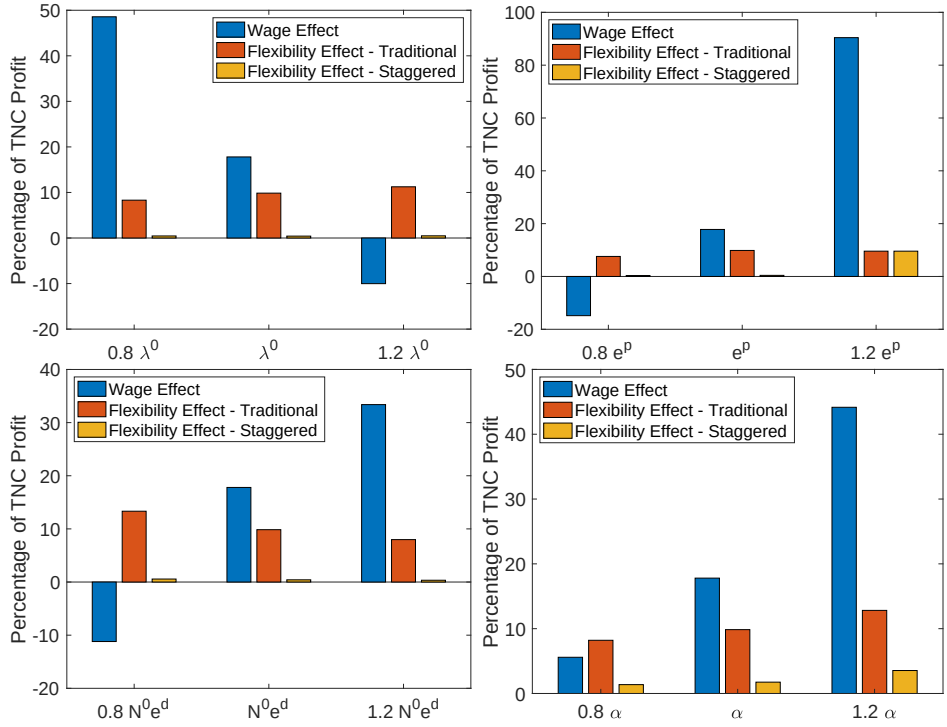


Figure B.1: Sensitivity analysis for the wage effect and flexibility effect under traditional and staggered shifts.

We observe that the wage effect is negative in three cases: $\{1.2\lambda^0, 0.8e^{\mathrm{p}}\}$ (passenger demand higher than estimated), and $\{0.8N^0e^{\mathrm{d}}\}$ (driver supply lower than estimated). Note that this does not mean the platform's profit would be higher than under no regulation case in these cases. In the minimum wage setting (5.5), we assume that the platform does

not face a supply side constraint at wage $w^{\min}$ (as explained in Remark 3). Since prevailing driver wages in our analysis are much lower than $w^{\min}$, this appears to be a valid assumption. However, if there is a substantial increase in passenger demand or decrease in the driver pool (as in the above three cases), platforms would offer wages higher than the minimum wage under no regulation and the wage effect would effectively be zero. As discussed in Section 7.5, such an exogenous shock to the ride-hailing ecosystem was observed as a result of the COVID-19 pandemic. Even so, platforms expect earnings to go back to pre-COVID levels over time [156]. Crucially, the flexibility effect can still be mitigated in all such cases using staggered shifts.

Appendix C

Proofs in Chapter 6

C.1 Proof of Lemma 6.1

In a batch of k orders, the maximum batching time is faced by the first order since it needs to wait for $k-1$ orders to arrive for the batch to be formed. The order arrival rate in a partition is λ/P (since each partition has area A/P). Since orders arrive according to a Poisson process, the order arrival times are exponentially distributed with rate λ/P. Since the first order needs to wait for $k-1$ orders and using the memoryless property of the Poisson process, the maximum batching time turns out to be $(k-1)P/\lambda$.

The maximum dispatch time is faced by the last (k^{th}) order to be delivered in the batch. For this order, the dispatch time is a sum of two terms: (i) time taken to pickup and drop off k orders, and (ii) travel time from store to k^{th} order location. The first term is $kt_{\text{pick}}+kt_{\text{drop}} = kt_{\text{fixed}}$. Notice that the second term denotes the tour distance $D^{\text{sp}}(k,P)$ without the last leg of the tour from the k^{th} order location back to the store. The k-order tour can be viewed as a sum of $k+1$ inter-order distances: {(store $\rightarrow$ order 1), (order 1 $\rightarrow$ order 2), ..., (order k $\rightarrow$ store)}. Since inter-order distances are approximately equal, the travel distance from the store to the k^{th} order is given by $D^{\text{sp}}(k,P)(1-1/(k+1)) = kD^{\text{sp}}(k,P)/(k+1)$.

The queuing time for the $M/G/N$ queue is approximately equal to $(1+\gamma^2)/2$ times the queuing time for the corresponding $M/M/N$ queue (with service times that are exponentially distributed rather than normally distributed as in the case of last-mile delivery), where γ represents the coefficient of variation, i.e., the ratio of the standard deviation of service time to its mean [90]. The variation in service time in our case is explained primarily by travel time variability. Empirical studies suggest that the co-efficient of variability γ is close to 1 [157]. Thus, the queuing time for the $M/G/N$ queue is approximately equal to that of the $M/M/N$ queue, which is given by

$$QT = \frac{\pi_0 \rho (N\rho)^N}{\lambda(1-\rho)^2 N!}, \tag{C.1}$$

where λ is the arrival rate, μ is the inverse of expected service time, $\rho = \lambda/N\mu$ is the

utilization rate, and $\pi_0 = \left(\sum_{m=0}^{N-1} \frac{\lambda^m}{\mu^m m!} + \frac{\lambda^N}{\mu^N N!(1-\frac{\lambda}{N\mu})}\right)^{-1}$. As long as N is sufficiently large and the utilization rate ρ is not close to 1, the queuing time ≈ 0. For instance, even with a utilization rate of 0.9 (much higher than prevailing practice), the queuing time is less than 5 min. The maximum delivery time is then effectively a sum of the maximum batching and the maximum dispatch time, which completes the proof.

C.2 Proof of Lemma 6.2

Since $D^{\text{sp}}(k, P)$ increases sub-linearly in k, $D^{\text{sp}}(k, P)/k$ is decreasing in k. Thus, $\mu^{\text{sp}}(k, P)$ is increasing in k.

The maximum delivery time $T^{\text{sp}}(k, P)$ is a sum of three terms:

$$T^{\text{sp}}(k, P) = \frac{(k-1)P}{\lambda} + \frac{kD^{\text{sp}}(k, P)}{(k+1)v} + kt_{\text{fixed}}. \tag{C.2}$$

It is clear that the first and third terms are linearly increasing in k. For the second term, notice that both $\frac{k}{k+1}$ and $D^{\text{sp}}(k, P)$ are increasing in k. Thus, the second term is increasing in k since it is a product of two increasing functions in k. This completes the proof.

Appendix D

Extended Analysis of Occlusions in Chapter 8

D.1 Left Turn Occlusion - Analysis of AV's actions

In this section, we demonstrate why the AV cannot do much to ensure its safety once it decides to make an unprotected left turn.

According to traffic laws in most jurisdictions, a left turning vehicle must enter the intersection only if it is sure that it can clear it, i.e., it is illegal to enter the intersection with an intention to stop so as to get a better view of through moving vehicles (TMVs). AVs have been observed to be much more conservative compared to human drivers when it comes to obeying traffic rules [111, 112, 113]. We assume that the left-turning AV in our example is programmed to obey traffic rules and hence, will attempt to stop while making the left turn only if it detects a TMV in its field of view.

Let v_{l} denote the AV's left turn speed (starting from rest, it accelerates to v_{l} and then completes the turn at this speed). To ensure that the AV has the best possible view of TMVs, we assume that there is almost no gap between the AV and the left edge of its lane, whereas all other vehicles are at the center of their lanes. Due to the intersection geometry in this example, the AV traverses a quarter of a circle of radius $R = 9$ m. We use $\theta \in [0, \pi/2]$ to parametrize the point on this circular trajectory at which the TMV comes into the field of view of the AV. At this point, the TMV is at a distance $d_{\text{CZ}}(\theta)$ from the conflict zone. Let us first derive how $d_{\text{CZ}}(\cdot)$ varies as a function of θ. We use the same intersection geometry and vehicle parameters as in Section 8.3:

- Lane width $l_{\text{w}} - 4$ m,
- AV's length $l_{\text{AV}} - 4$ m,
- Maximum acceleration rate $a_{\text{acc}} = 3$ m/s^2,
- Maximum deceleration rate $a_{\text{dec}} = 4$ m/s^2.

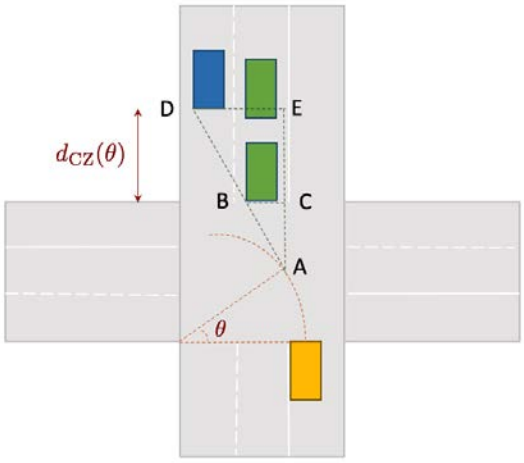


Figure D.1: A detailed analysis of the AV's actions to ensure safety during the unprotected left turn.

Thus, $l(\text{BC}) = 9\cos\theta - 5$, $l(\text{DE}) = 9\cos\theta - 1$, and $l(\text{AC}) = 12 - 9\sin\theta$. As $\triangle(\text{ABC})$ and $\triangle(\text{ADE})$ are similar, we have

$$\frac{l(\text{BC})}{l(\text{DE})} = \frac{l(\text{AC})}{l(\text{AE})}. \tag{D.1}$$

Therefore, $l(\text{AE}) = \frac{(9\cos\theta - 1)(12 - 9\sin\theta)}{9\cos\theta - 5}$. Thus, we have

$$\begin{aligned} d_{\text{CZ}}(\theta) &= l(\text{CE}), && \text{(D.2)} \\ &= l(\text{AE}) - l(\text{AC}), && \text{(D.3)} \\ &= \frac{4(12 - 9\sin\theta)}{9\cos\theta - 5}. && \text{(D.4)} \end{aligned}$$

Recall that the minimum distance $d_{\text{CZ}}^{\min}$ such that TMVs moving at 25 mph can stop before reaching the conflict zone is equal to 23.45 m. This corresponds to $\theta_{\max} = 0.86$ on the AV's trajectory. If the AV and TMV first come into each other's field of view at any $\theta > \theta_{\max}$, then a crash would be averted as the TMV would have enough time to brake to a stop before the conflict zone. Thus, we can focus on $\theta \leq 0.86$.

The AV is making the left-turn with velocity $v_{\text{l}} = 4.5$ m/s (10 mph - much lower than the 15 mph speed limit). Let ρ_{AV} denote the AV's reaction time. Once the TMV comes into its field of view, it takes ρ_{AV} seconds to take either of the two actions - (i) decelerate to a stop before the conflict zone, and (ii) accelerate to clear the conflict zone before the TMV's arrival. We'll now argue why either of these actions will not ensure the AV's safety.

Suppose the AV tries to brake to a stop. Let θ_{conf} denote the angle corresponding to the AV entering the conflict zone. Then, the AV's distance to the conflict zone when it first sees the TMV is equal to $R(\theta_{\text{conf}} - \theta)$. It travels a distance $v_{\text{l}}\rho_{\text{AV}}$ before it starts decelerating. Thus, it can stop before reaching the conflict zone when

$$\frac{v_{\text{l}}^2}{2a_{\text{dec}}} \leq R(\theta_{\text{conf}} - \theta) - v_{\text{l}}\rho_{\text{AV}}. \tag{D.5}$$

Thus, the AV can be safe in this case only as long as

$$\theta \leq \theta_{\text{conf}} - \frac{v_{\text{l}}^2}{2a_{\text{dec}}R} - \frac{v_{\text{l}}\rho_{\text{AV}}}{R} \tag{D.6}$$

On substituting the required parameters, we see that the AV can decelerate to a stop before the conflict zone only when $\theta \leq 0.48$. On the other hand, if the AV decides to accelerate, it must clear the conflict zone before the TMV's arrival. The TMV is at distance $d_{\text{CZ}}(\theta)$ from the conflict zone when the AV comes into its field of view. It travels a distance ρv_{th} before it starts decelerating. Thus, the time t_{dec} it takes to reach the conflict zone after it starts decelerating satisfies

$$d_{\text{CZ}}(\theta) = v_{\text{th}}t_{\text{dec}} - \frac{1}{2}a_{\text{dec}}t_{\text{dec}}^2. \tag{D.7}$$

Thus, the total time the AV has in order to clear the conflict zone starting from θ is given by $t_{\text{TMV}} = \rho + t_{\text{dec}}$. The AV needs to cover the distance $R(\theta_{\text{conf}} - \theta) + l_{\text{w}} + l_{\text{AV}}$ (distance to conflict zone + lane width + AV's length). It covers a distance $v_{\text{l}}\rho_{\text{AV}}$ before it starts accelerating. It can clear the conflict zone within t_{TMV} seconds only as long as

$$v_{\text{l}}(t_{\text{TMV}} - \rho_{\text{AV}}) + \frac{1}{2}a_{\text{acc}}(t_{\text{TMV}} - \rho_{\text{AV}})^2 \geq R(\theta_{\text{conf}} - \theta) + l_{\text{w}} + l_{\text{AV}} - v_{\text{l}}\rho_{\text{AV}}. \tag{D.8}$$

Plugging in the required parameters in (D.7) and (D.8), we find that the AV will be safe using this approach only when $\theta \geq 0.81$. Combining our findings for both approaches, we can conclude that neither accelerating nor decelerating can ensure that a crash is averted when $\theta \in [0.48, 0.81]$. Thus, the AV cannot be worst-case safe while making a left turn in this scenario even if it attempts an evasive maneuver.

In Section 8.3, we did not explicitly account for the AV's evasive maneuvers. This corresponds to a conflict occurring when $\theta \in [0, 0.86]$. Thus, the range of unsafe θ in the case without evasive maneuvers is $0.86/(0.81 - 0.48) \approx 2.6$ times that in the case with evasive maneuvers. Recall our result from Section 8.3.2 that the AV needed to wait 443 seconds to be confident enough to make the turn. Assuming the AV tolerates the risk of evasive maneuvers as described above, it would need to wait $443/2.6 \approx 170$ seconds (about 3 min). In other words, it would still have to wait multiple traffic cycles in order to be confident enough to turn.

D.2 Detailed Explanation of the Occluded Pedestrians Scenario

In this section, we provide a more detailed explanation for the analysis presented in Section 8.3.3.

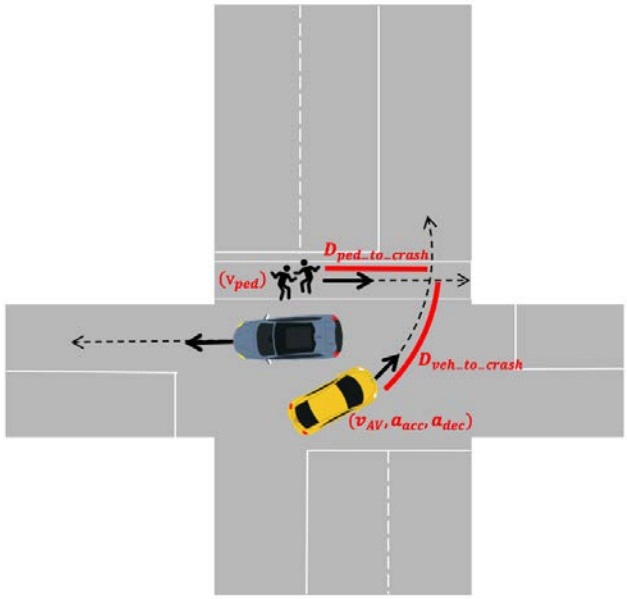


Figure D.2: Calculating conflict probability for Pedestrian/Maneuver Type 2.

We reproduce here the notation used in Section 8.3.3:

- λ_{ped}: arrival rate (ped/s) of pedestrians,
- v_{ped}: pedestrian speed (m/s) when crossing on yellow/flashing phase/at the end of pedestrian phase,
- $D_{\text{ped\_to\_crash}}$: distance (m) from pedestrian's position at $t = 0$ to the center of pedestrian conflict zone,
- v_{AV}: speed of the AV (m/s),
- $D_{\text{veh\_to\_crash}}$: distance (m) from vehicle's position at $t = 0$ to the pedestrian conflict zone,
- a_{acc}: maximum AV acceleration rate (m/s^2),
- a_{dec}: maximum AV deceleration rate (m/s^2),
- w_{AV}: width of AV (m).

As discussed in Section 8.3.3, the safety challenge in this scenario is that the AV's field of view of pedestrians on the crosswalk is obstructed by a neighboring vehicle. Once the AV detects the pedestrian, it can take two possible actions to avert a crash:

1. Accelerate so that it reaches the conflict zone before the pedestrian enters it,
2. Decelerate so that it reaches the conflict zone after the pedestrian crosses it.

We consider $t = 0$ to be the time at which the AV first detects the pedestrian. Let δ denote the time required by the pedestrian to cross the conflict zone, i.e., $\delta = w_{\text{AV}}/v_{\text{ped}}$. Let v_{AV} denote the AV's initial speed and let t_{acc} denote the time taken by the AV to reach the conflict zone while it accelerates. As the distance between the AV and the conflict zone is $D_{\text{veh\_to\_crash}}$ and it is accelerating at rate a_{acc}, we have

$$D_{\text{veh\_to\_crash}} = v_{\text{AV}} t_{\text{acc}} + \frac{1}{2} a_{\text{acc}} t_{\text{acc}}^2. \tag{D.9}$$

Taking the positive root of this quadratic equation gives

$$t_{\text{acc}} = \frac{\sqrt{2a_{\text{acc}} D_{\text{veh\_to\_crash}} + v_{\text{AV}}^2} - v_{\text{AV}}}{a_{\text{acc}}}, \tag{D.10}$$

At t $= t_{\text{acc}}$, a conflict occurs if the pedestrian is already in the conflict zone. Notice that this corresponds to the pedestrian being at the center of the conflict zone in the interval $\mathcal{T}_{\text{acc}} = [t_{\text{acc}} - \delta/2, t_{\text{acc}} + \delta/2]$. For instance, if the pedestrian is at the center of the conflict zone at $t = t_{\text{acc}} - \delta/2$, then it implies that at $t = t_{\text{acc}}$ (the time at which AV enters the conflict zone), the pedestrian is at the right edge of the conflict zone. Similarly, $t_{\text{acc}} + \delta/2$ corresponds to the pedestrian being at the left edge of the conflict zone.

On the other hand, let t_{dec} denote the time taken by the AV to reach the conflict zone while it decelerates at rate a_{dec}. Let v_{AV}^{f} denote the the AV's speed when it enters the conflict zone. Then, the following equations hold:

$$v_{\text{AV}}^{\text{f}} = v_{\text{AV}} - a_{\text{dec}} t_{\text{dec}}, \tag{D.11}$$

$$(v_{\text{AV}}^{\text{f}})^2 = v_{\text{AV}}^2 - 2a_{\text{dec}} D_{\text{veh\_to\_crash}}. \tag{D.12}$$

Solving for t_{dec}, we have

$$t_{\text{dec}} = \frac{v_{\text{AV}} - \sqrt{v_{\text{AV}}^2 - 2a_{\text{dec}} D_{\text{veh\_to\_crash}}}}{a_{\text{dec}}}, \tag{D.13}$$

and a conflict occurs if the pedestrian arrives at the center of the pedestrian conflict zone in the time interval $\mathcal{T}_{\text{dec}} = [t_{\text{dec}} - \delta/2, t_{\text{dec}} + \delta/2]$. Thus, $\mathcal{T}_{\text{acc}} \cap \mathcal{T}_{\text{dec}}$ represents the time interval in which no matter what action the AV chooses (accelerate or decelerate), a conflict is unavoidable. Since $t_{\text{acc}} < t_{\text{dec}}$, $\mathcal{T}_{\text{acc}} \cap \mathcal{T}_{\text{dec}} = [t_{\text{dec}} - \delta/2, t_{\text{acc}} + \delta/2]$.

The probability of this event can be interpreted as the probability of arrival of at least one pedestrian in a time interval of length $(t_{\text{acc}} - t_{\text{dec}} + \delta)$. As the pedestrian arrival process is Poisson with rate λ_{ped}, we have

$$P_{\text{scenario2}}(\text{Conflict} \mid \text{Conditions 1 and 2 hold}) = 1 - e^{-\lambda_{\text{ped}}(t_{\text{acc}} - t_{\text{dec}} + \delta)}. \tag{D.14}$$

Recall that $D_{\text{ped\_to\_crash}}$ denotes the distance from the center of the conflict zone at $t = 0$. When the pedestrian is at the center of the conflict zone at $t = t_{\text{dec}} - \delta/2$ (lower limit of $\mathcal{T}_{\text{acc}} \cap \mathcal{T}_{\text{dec}}$), it implies that the pedestrian was at a distance $(t_{\text{dec}} - \delta/2)v_{\text{ped}}$ at $t = 0$. Similarly, the pedestrian being at the center of the conflict zone at $t = t_{\text{dec}} + \delta/2$ (upper limit of $\mathcal{T}_{\text{acc}} \cap \mathcal{T}_{\text{dec}}$) implies that the pedestrian was at a distance $(t_{\text{dec}} + \delta/2)v_{\text{ped}}$ at $t = 0$. Thus, we can conclude that

$$D_{\text{ped\_to\_crash}} \in [(t_{\text{dec}} - \delta/2)v_{\text{ped}}),\ (t_{\text{acc}} + \delta/2)v_{\text{ped}}]. \tag{D.15}$$

The same analysis can be suitably modified for the other two occluded pedestrian scenarios, as suggested in Section 8.3.3.

Bibliography

[1] Jean-Paul Rodrigue. *The Geography of Transport Systems*. Routledge, 2020.

[2] Tom Standage. *A Brief History of Motion: From the Wheel, to the Car, to what Comes Next*. Bloomsbury Publishing, 2021.

[3] Eric Morris. "From Horse Power to Horsepower". In: *Access Magazine* 1.30 (2007), pp. 2–10.

[4] World Health Organization. *Global Status Report on Road Safety 2018*. Tech. rep. The World Health Organization, 2018.

[5] National Center for Statistics and Analysis. *2018 Fatal Motor Vehicle Crashes: Overview. (Traffic Safety Facts Research Note. Report No. DOT HS 812 826)*. Tech. rep. United States National Highway Traffic Safety Administration, 2019.

[6] Inrix. *Congestion Costs Each American 97 hours, $1,348 A Year*. https://inrix.com/press-releases/scorecard-2018-us/. 2019.

[7] Environmental Protection Agency. "Fast Facts on Transportation Greenhouse Gas Emissions". In: (2018). URL: https://www.epa.gov/greenvehicles/fast-facts-transportation-greenhouse-gas-emissions.

[8] Imran Cronk. "The Transportation Barrier". In: *The Atlantic* (2015). URL: https://www.theatlantic.com/health/archive/2015/08/the-transportation-barrier/399728/.

[9] Matt Alderton. "Nearly 30 years after the ADA, the nation's transit agencies report successes and shortfalls". In: *The Washington Post* (2020). URL: https://www.washingtonpost.com/local/trafficandcommuting/nearly-30-years-after-ada-nations-transit-agencies-report-successes-and-shortfalls/2020/06/25/76e102d8-af22-11ea-8758-bfd1d045525a_story.html.

[10] Roger Lanctot. "Accelerating the future: The economic impact of the emerging passenger economy". In: *Strategy Analytics* 5 (2017).

[11] Peter Cohen et al. *Using Big Data to Estimate Consumer Surplus: The Case of Uber*. Tech. rep. National Bureau of Economic Research, 2016.

[12] Uber. *New Survey: Drivers Choose Uber for its Flexibility and Convenience*. https://www.uber.com/i partner-survey. 2015.

[13] Mark Gurman. "Waymo CEO says self-driving cars won't be ubiquitious for decades". In: *Bloomberg News* (2018). URL: https://www.bloomberg.com/news/articles/2018-11-13/waymo-ceo-says-self-driving-cars-won-t-be-ubiqitious-for-decades.

[14] Fehr and Peers. *What are TNCs Share of VMT?* https://www.fehrandpeers.com/what-are-tncs-share-of-vmt/. 2019.

[15] J. A. Parrott and M. Reich. *An earning standard for New York City app-based drivers: Economic analysis and policy assessment*. The New School, Center for New York City Affairs. 2018.

[16] J. A. Parrott and M. Reich. *A minimum compensation standard for Seattle TNC drivers*. The New School, Center for New York City Affairs. 2020.

[17] Akhil Shetty et al. "An Analysis of Labor Regulations for Transportation Network Companies". In: *Economics of Transportation* 32 (2022), p. 100284.

[18] Akhil Shetty et al. "Impact of Driver Classification Regulations on Transportation Network Companies". In: *2020 IEEE 23rd International Conference on Intelligent Transportation Systems (ITSC)*. IEEE. 2020, pp. 1–6.

[19] Akhil Shetty et al. "The Value of Pooling in Last-Mile Delivery". In: *2022 IEEE 61st Conference on Decision and Control (CDC)*. IEEE. 2022, pp. 531–538.

[20] Akhil Shetty et al. "Safety Challenges for Autonomous Vehicles in the Absence of Connectivity". In: *Transportation Research Part C: Emerging Technologies* 128 (2021), pp. 103–133.

[21] Akhil Shetty et al. "Risk Assessment of Autonomous Vehicles across Diverse Driving Contexts". In: *2021 IEEE International Intelligent Transportation Systems Conference (ITSC)*. IEEE. 2021, pp. 712–719.

[22] Akhil Shetty et al. "Automated Vehicle Safety and Deployment: Lessons from Human Crashes". In: *2022 International Conference on Connected Vehicle and Expo (ICCVE)*. IEEE. 2022, pp. 1–6.

[23] DMR. *110 amazing Uber stats and facts by the numbers*. https://expandedramblings.com/index.php/uber-statistics/. 2019.

[24] M. Iqbal. *Uber Revenue and Usage Statistics*. https://www.businessofapps.com/data/uber-statistics/. 2020.

[25] Bill Aull et al. "Making online grocery a winning proposition". In: *McKinsey & Company* (2022).

[26] Janine Perri. "Which company is winning the restaurant food delivery war?" In: *Bloomberg Second Measure* (2022).

[27] ResearchandMarkets.com. "Global Last-mile Food and Grocery Delivery Growth Opportunities". In: (August 2021).

[28] Mercatus. *eGrocery Transformed: 2021 market projections and insight into online grocery's elevated future*. October 2021.

[29] Anne Brown and Whitney LaValle. "Hailing a change: Comparing taxi and ridehail service quality in Los Angeles". In: *Transportation* 48.2 (2021), pp. 1007–1031.

[30] Thais Rangel et al. "Exploring ride-hailing fares: an empirical analysis of the case of Madrid". In: *Transportation* 49.2 (2022), pp. 373–393.

[31] B. Schaller. *The New Automobility: Lyft, Uber and the Future of American Cities*. Schaller Consulting. July 2018.

[32] J. Castiglione et al. *TNCs today: a profile of San Francisco transportation network company activity*. San Francisco County Transportation Authority. 2016.

[33] J. Castiglione et al. *TNCs & Congestion*. San Francisco County Transportation Authority. 2018.

[34] Don Anair. *Ride-Hailing's Climate Risks: Steering a Growing Industry toward a Clen Transporatation Future*. Union of Concerned Scientists., 2020.

[35] NYCTLC (New York City Taxi and Limousine Commission)). *Notice of promulgation: statement of basis and purpose of rules*. https://www1.nyc.gov/assets/tlc/downloads/pdf/driver_income_rules_12_04_2018.pdf. 2018.

[36] Office of Labor Standards. *Transportation Network Company Minimum Compensation Ordinance*. https://www.seattle.gov/laborstandards/ordinances/tnc-legislation/minimum-compensation-ordinance. 2020.

[37] S. Li et al. "Regulating TNCs: should Uber and Lyft set their own rules?" In: *Transportation Research Part B: Methodological* 129 (2019), pp. 193–225.

[38] Jay Peters and Andrew Hawkins. "Uber and Lyft ordered by California judge to classify drivers as employees". In: *The Verge* (2020). URL: https://www.theverge.com/2020/8/10/21362460/uber-lyft-drivers-employees-california-court-ruling.

[39] SF SPUR. *SF Prop D Ride-hailing Tax*. https://www.spur.org/voter-guide/2019-11/sf-prop-d-ride-hailing-tax. 2019.

[40] City of Chicago. *City of Chicago Congestion Pricing*. https://www.chicago.gov/city/en/depts/bacp/supp_info/city_of_chicago_congestion_pricing.html. 2019.

[41] California Air Resources Board. *California requires zero-emissions vehicle use for ridesharing services, another step toward achieving the state's climate goals*. URL: %7Bhttps://ww2.arb.ca.gov/news/california-requires-zero-emissions-vehicle-use-ridesharing-services-another-step-toward%7D.

[42] Uber. *Uber Announces Results for Fourth Quarter and Full Year 2022*. https://s23.q4cdn.com/407969754/files/doc_financials/2022/q4/Uber-Q4-22-Earnings-Press-Release.pdf. 2022.

[43] Preetika Rana. "DoorDash's 2021 Revenue Tops Pandemic Record". In: *The Wall Street Journal* (2022).

[44] Kellen Browning. "Instacart Cuts Its Valuation by 38 Percent, Citing 'Turbulence'". In: *The New York Times* (2022).

[45] Richard Arnott. "Taxi travel should be subsidized". In: *Journal of Urban Economics* 40.3 (1996), pp. 316–333.

[46] Nikita Korolko et al. "Dynamic pricing and matching in ride-hailing platforms". In: *Available at SSRN* (2018).

[47] Lyft. *San Francisco Bay Area Ride Costs*. https://www.lyft.com/pricing/sfo. 2021.

[48] Michael Reich et al. *Pay, Passengers and Profits: Effects of Employee Status for California TNC Drivers*. Tech. rep. Institute of Industrial Relations, UC Berkeley, 2020.

[49] Louis Hyman et al. "Platform Driving In Seattle". In: (2020).

[50] Alison Stein. *Unpacking Pay Standards: a response to the UC Berkeley Labor Center*. Uber Under the Hood. https://medium.com/uber-under-the-hood/unpacking-pay-standards-a-response-to-the-uc-berkeley-labor-center-aa5549871040. 2019.

[51] Libby Mishkin. *Which drivers do the most trips?* Uber Under the Hood. https://medium.com/uber-under-the-hood/which-drivers-do-the-most-trips-9c475e99e071. 2020.

[52] Mathieu Rosemain and Dominique Vidalon. "Top French court deals blow to Uber by giving driver 'employee' status". In: *Reuters* (2020). URL: https://www.reuters.com/article/us-uber-court/top-french-court-deals-blow-to-uber-by-giving-driver-employee-status-idUSKBN20R23F.

[53] Adam Satariano. "In a First, Uber Agrees to Classify British Drivers as 'Workers'". In: *The New York Times* (2021). URL: https://www.nytimes.com/2021/03/16/technology/uber-uk-drivers-worker-status.html.

[54] David Meyer. "Gig economy: Europe tells companies to negotiate with workers or face new laws". In: *Fortune* (2021). URL: https://fortune.com/2021/02/24/gig-economy-europe-worker-rights-eu/.

[55] Nandita Bose. "U.S. Labor Secretary supports classifying gig workers as employees". In: *Reuters* (2021). URL: https://www.reuters.com/world/us/exclusive-us-labor-secretary-says-most-gig-workers-should-be-classified-2021-04-29/.

[56] Uber. "A better deal: partnering to improve platform work for all". In: (2021). URL: https://www.uber.com/global/en/about/reports/a-better-deal/.

[57] Alison Stein. *Independent couriers' reaction to employee reclassification: learnings from Geneva*. Uber Under the Hood. https://medium.com/uber-under-the-hood/independent-couriers-reaction-to-employee-reclassification-learnings-from-geneva-e3885db12ea3.

[58] Ken Jacobs and Michael Reich. "The Effects of Proposition 22 on Driver Earnings: Response to a Lyft-Funded Report by Dr. Christopher Thornberg". In: ().

[59] Colin Camerer et al. "Labor supply of New York City cabdrivers: One day at a time". In: *The Quarterly Journal of Economics* 112.2 (1997), pp. 407–441.

[60] Michiel Evers, Ruud De Mooij, and Daniel Van Vuuren. "The wage elasticity of labour supply: a synthesis of empirical estimates". In: *De Economist* 156.1 (2008), pp. 25–43.

[61] Andreas Lichter, Andreas Peichl, and Sebastian Siegloch. "The own-wage elasticity of labor demand: A meta-regression analysis". In: *European Economic Review* 80 (2015), pp. 94–119.

[62] Nicholas Buchholz, Matthew Shum, and Haiqing Xu. "Flexible Work Hours and Dynamic Labor Supply: Evidence from Taxi Drivers". In: *Available at SSRN 2748697* (2018).

[63] Michael Reich et al. "San Francisco's Proposed City Minimum Wage Law: A Prospective Impact Study". In: *Institute for Research on Labor and Employment University of California, Berkeley. Retrieved from http://irle. berkeley. edu/cwed/briefs/2014-04. pdf* (2014).

[64] Sergio Avedian. "Uber & Lyft Are Locking Drivers Out In New York City". In: *The Rideshare Guy* (2020). URL: https://therideshareguy.com/nyc-uber-lyft-driver-lockout/.

[65] Shan Jiang et al. "On ridesharing competition and accessibility: Evidence from uber, lyft, and taxi". In: *Proceedings of the 2018 World Wide Web Conference*. 2018, pp. 863–872.

[66] M Keith Chen et al. *The value of flexible work: Evidence from uber drivers*. Tech. rep. National Bureau of Economic Research, 2017.

[67] Aarian Marshall. "The Secret Uber Data That Could Fix Your Commute". In: *Wired* (2017). URL: https://www.wired.com/2017/02/ubers-coughing-data-nyc-fix-commute/.

[68] Elizabeth Creely. "City report says Uber and Lyft are hoarding vital transit data". In: *Mission Local* (2018). URL: https://missionlocal.org/2018/07/city-report-says-uber-and-lyft-are-hoarding-vital-transit-data/.

[69] Akshat Pandey and Aylin Caliskan. "Disparate Impact of Artificial Intelligence Bias in Ridehailing Economy's Price Discrimination Algorithms". In: *Association for the Advancement of Artificial Intelligence* (2021).

[70] Hongtai Yang, Yuan Liang, and Linchuan Yang. "Equitable? Exploring ridesourcing waiting time and its determinants". In: *Transportation Research Part D: Transport and Environment* 93 (2021), p. 102774.

[71] S. Li et al. "Spatial Pricing in Ride-Sourcing Markets under a Congestion Charge". In: *arXiv preprint arXiv:2010.09260* (2020).

[72] Graham Rapier. *Lyft and Juno are suing New York City over new minimum wage requirements for drivers*. Business Insider. https://www.businessinsider.com/lyft-juno-sue-new-york-city-minimum-wage-law-for-drivers-2019-1. Jan. 2019.

[73] Faiz Siddiqui. *Where have all the Uber drivers gone?* The Washington Post. https://www.washingtonpost.com/technology/2021/05/07/uber-lyft-drivers. May 2021.

[74] Marc-André Kamel et al. "How to Ramp Up Online Grocery without Breaking the Bank". In: *Bain & Company* (2020).

[75] Kees Jacobs et al. "The last-mile delivery challenge". In: *Capgemini Research Institute* (2019).

[76] DoorDash. *DoorDash Releases Fourth Quarter 2022 Financial Results*. https://ir.doordash.com/financials/quarterly-results/. 2022.

[77] Martin Joerss, Florian Neuhaus, and Jurgen Schroder. "How customer demands are reshaping last-mile delivery". In: *McKinsey & Company* (2016).

[78] Laetitia Dablanc et al. "The rise of on-demand 'Instant Deliveries' in European cities". In: *Supply Chain Forum: An International Journal*. Vol. 18. 4. Taylor & Francis. 2017, pp. 203–217.

[79] Margot Boyer-Dry. "Will Rapid Grocery Delivery Change N.Y.C.? Look to Berlin." In: *The New York Times* (2022).

[80] Iris I Lin and Hani S Mahmassani. "Can online grocers deliver?: Some logistics considerations". In: *Transportation Research Record* 1817.1 (2002), pp. 17–24.

[81] Kenneth K Boyer, Andrea M Prud'homme, and Wenming Chung. "The last mile challenge: evaluating the effects of customer density and delivery window patterns". In: *Journal of Business Logistics* 30.1 (2009), pp. 185–201.

[82] Praveen Adhi et al. "Reimagining the role of physical stores in an omnichannel distribution network". In: *McKinsey & Company* (2021).

[83] Richard M Karp. *Reducibility among Combinatorial Problems*. Springer, 2010.

[84] Jillian Beardwood, John H Halton, and John Michael Hammersley. "The shortest path through many points". In: *Mathematical Proceedings of the Cambridge Philosophical Society*. Vol. 55. 4. Cambridge University Press. 1959, pp. 299–327.

[85] Carlos F Daganzo. "The length of tours in zones of different shapes". In: *Transportation Research Part B: Methodological* 18.2 (1984), pp. 135–145.

[86] Elena Belavina. "Grocery store density and food waste". In: *Manufacturing & Service Operations Management* 23.1 (2021), pp. 1–18.

[87] Anthony Wren and Alan Holliday. "Computer scheduling of vehicles from one or more depots to a number of delivery points". In: *Journal of the Operational Research Society* 23.3 (1972), pp. 333–344.

[88] Billy E Gillett and Leland R Miller. "A heuristic algorithm for the vehicle-dispatch problem". In: *Operations Research* 22.2 (1974), pp. 340–349.

[89] Gabriel R Bitran and Devanath Tirupati. "Approximations for product departures from a single-server station with batch processing in multi-product queues". In: *Management Science* 35.7 (1989), pp. 851–878.

[90] Wallace J Hopp and Mark L Spearman. *Factory Physics*. Waveland Press, 2011.

[91] Carlos F Daganzo. "The distance traveled to visit N points with a maximum of C stops per vehicle: An analytic model and an application". In: *Transportation Science* 18.4 (1984), pp. 331–350.

[92] INRIX. *2021 INRIX Global Traffic Scorecard*. December 2021. URL: https://inrix.com/scorecard/.

[93] United States Census Bureau. *Quick Facts, Los Angeles City, California*. October 2021. URL: https://www.census.gov/quickfacts/losangelescitycalifornia.

[94] Centers for Disease Control and Prevention. "Web-based Injury Statistics Query and Reporting System (WISQARS)[Online].(2020)". In: *National Center for Injury Prevention and Control, Centers for Disease Control and Prevention (producer). Available from: URL: www. cdc. gov/ncipc/wisqars* (2020).

[95] Lawrence Blincoe et al. *The economic and social impact of motor vehicle crashes, 2010 (Revised) (Report No. DOT HS 812 013)*. Tech. rep. United States National Highway Traffic Safety Administration, 2015.

[96] Cameron F. Kerry and Jack Karsten. *Gauging investments in self-driving cars*. https://www.brookings.edu/research/gauging-investment-in-self-driving-cars. 2017.

[97] Alex Davies. "GM's Cruise rolls back its target for self-driving cars". In: *Wired* (2019). https://www.wired.com/story/gms-cruise-rolls-back-target-self-driving-cars/.

[98] Yoko Kubota. "Toyota Aims to Make Self-Driving Cars by 2020". In: *The Wall Street Journal* (2015). URL: https://www.wsj.com/articles/toyota-aims-to-make-self-driving-cars-by-2020-1444136396.

[99] Alexis Madrigal. "The Most Important Self-Driving Car Announcement Yet". In: *The Atlantic* (2018). URL: https://www.theatlantic.com/technology/archive/2018/03/the-most-important-self-driving-car-announcement-yet/556712/.

[100] J. Niccolai. *Self-driving cars a reality for 'ordinary people' within 5 years says Google's Sergey Brin*. Computer World. September 25, 2012.

[101] Raaga Kannan and Ronald C Lasky. "Autonomous Vehicles Still Decades Away: 2019". In: *2020 Pan Pacific Microelectronics Symposium (Pan Pacific)*. IEEE. 2020, pp. 1–6.

[102] John M. Scanlon et al. "Waymo Simulated Driving Behavior in Reconstructed Fatal Crashes within an Autonomous Vehicle Operating Domain". In: *Accident Analysis & Prevention* 163 (2021), p. 106454.

[103] Andrew Hawkins. "Serious safety lapses led to Uber's fatal self-driving crash, new documents suggest". In: *The Verge* (2019).

[104] Tom Krisher and Olga R. Rodriguez. "NTSB releases details in 2 crashes involving Tesla Autopilot". In: *Associated Press* (2020).

[105] National Highway Traffic Safety Administration. *National Motor Vehicle Crash Causation Survey: Report to Congress*. 2008.

[106] Elizabeth D Swanson et al. *Statistics of Light-Vehicle Pre-Crash Scenarios Based on 2011–2015 National Crash Data*. Tech. rep. Washington, DC: National Highway Traffic Safety, 2019.

[107] Waymo. "Waymo Safety Report". In: (2020). https://storage.googleapis.com/sdc-prod/v1/safety-report/2020-09-waymo-safety-report.pdf.

[108] Phillip Koopman. "A Reality Check on the 94 Percent Human Error Statistic for Self-Driving Cars". In: *Safe Autonomy Blog* (2018). http://safeautonomy.blogspot.com/2018/06/a-reality-check-on-94-percent-human.html.

[109] L Lubbe et al. "Predicted road traffic fatalities in Germany: The potential and limitations of vehicle safety technologies from passive safety to highly automated driving". In: *Proceedings of IRCOBI conference. Athena, Greece*. 2018.

[110] Casualty Actuarial Society Automated Vehicles Task Force. "Restating the National Highway Transportation Safety Administration's National Motor Vehicle Crash Causation Survey for Automated Vehicles". In: *Casualty Actuarial Society E-Forum, Fall 2014-Volume 1* (2014).

[111] Hannah Boland. "Waymo's self-driving cars 'struggle to turn left and don't understand basic road features'". In: *The Telegraph* (2018). URL: https://www.telegraph.co.uk/technology/2018/08/29/waymos-self-driving-cars-struggle-turn-left-dont-understand/.

[112] Amir Efrati. "Waymo's Foes: Left Turns and the Mean Streets of Phoenix". In: *The Telegraph* (2017). URL: https://www.telegraph.co.uk/technology/2018/08/29/waymos-self-driving-cars-struggle-turn-left-dont-understand/.

[113] Andrew Hawkins. "Waymo's driverless cars hit a new milestone: 10 million miles on public roads". In: *The Verge* (2018). URL: https://www.theverge.com/2018/10/10/17958276/waymo-self-driving-cars-10-million-miles-challenges.

[114] "Collision between vehicle controlled by developmental automated driving system and pedestrian". In: *National Transportation Safety Board, Washington, DC, USA, Tech. Rep. HAR-19-03* (2019).

[115] National Highway Traffic Safety Administration. *Crash Factors in Intersection-Related Crashes: An On-Scene Perspective*. 2010.

[116] Matthew Schwall et al. "Waymo Public Road Safety Performance Data". In: *arXiv preprint arXiv:2011.00038* (2020).

[117] The Washington Post. *Waymo to launch fully driverless service to the public — a first just in time for the pandemic*. https://www.washingtonpost.com/technology/2020/10/08/waymo-driverless-rides/. Oct. 2020.

[118] Andrew J Leslie. *Analysis of the Field Effectiveness of General Motors Production Active Safety and Advanced Headlighting Systems*. Tech. rep. University of Michigan, Ann Arbor, Transportation Research Institute, 2019.

[119] Mike Isaac. "Uber Suspends Tests of Self-Driving Vehicles After Arizona Crash". In: *The New York Times* (2017).

[120] SAE On-Road Automated Vehicle Standards Committee. "Taxonomy and definitions for terms related to on-road motor vehicle automated driving systems". In: *SAE Standard J* 3016 (2014), pp. 1–16.

[121] Shai Shalev-Shwartz, Shaked Shammah, and Amnon Shashua. "On a formal model of safe and scalable self-driving cars". In: *arXiv preprint arXiv:1708.06374* (2017).

[122] Offer Grembek et al. "Making intersections safer with I2V communication". In: *Transportation Research Part C: Emerging Technologies* 102 (2019), pp. 396–410.

[123] James A Misener. *Cooperative intersection collision avoidance system (CICAS): Signalized left turn assist and traffic signal adaptation*. Tech. rep. 2010.

[124] SAE International. *J2735 Standard*. https://www.sae.org/standardsdev/dsrc.

[125] Ann Hsu et al. "Protocol Design for an Automated Highway System". In: *Discrete Even Dynamic Systems: Theory and Applications* 2 (1993), pp. 183–206.

[126] Yugong Luo et al. "A dynamic automated lane change maneuver based on vehicle-to-vehicle communication". In: *Transportation Research Part C: Emerging Technologies* 62 (2016), pp. 87–102.

[127] Xue Yang et al. "A vehicle-to-vehicle communication protocol for cooperative collision warning". In: *The First Annual International Conference on Mobile and Ubiquitous Systems: Networking and Services, 2004. MOBIQUITOUS 2004.* IEEE. 2004, pp. 114–123.

[128] Jin Cui et al. "A review on safety failures, security attacks, and available countermeasures for autonomous vehicles". In: *Ad Hoc Networks* 90 (2019), p. 101823.

[129] Ben Mccluskey. "Connected cars–the security challenge". In: *Engineering & Technology* 12.2 (2017), pp. 54–57.

[130] Chuck Holland et al. "UPS optimizes delivery routes". In: *Interfaces* 47.1 (2017), pp. 8–23.

[131] Marc Green. ""How long does it take to stop?" Methodological analysis of driver perception-brake times". In: *Transportation human factors* 2.3 (2000), pp. 195–216.

[132] Stuart R Perkins and Joseph L Harris. "Traffic conflict characteristics-accident potential at intersections". In: *Highway Research Record* 225 (1968).

[133] William D Glauz, KM Bauer, and Donald J Migletz. "Expected traffic conflict rates and their use in predicting accidents". In: *Transportation Research Record* 1026 (1985), pp. 1–12.

[134] Nathan H Gartner, Carrol Jl Messer, and Ajay Rathi. "Traffic flow theory-A state-of-the-art report: revised monograph on traffic flow theory". In: (2002).

[135] Transportation Research Board. "Highway Capacity Manual, Chapter 9 (Signalized Intersections)". In: *Transportation Research Board* (1993).

[136] Robert W Keener. *Theoretical Statistics: Topics for a Core Course*. Springer, 2010.

[137] Safe Transportation Research and Education Center. *Transportation Injury Mapping System (TIMS)*. 2020. URL: https://tims.berkeley.edu/.

[138] A. Muralidharan et al. "Management of intersections with multi-modal high-resolution data". In: *Transportation Research, Part C* 68 (2016), pp. 101–112.

[139] Dorsa Sadigh et al. "Planning for autonomous cars that leverage effects on human actions." In: *Robotics: Science and Systems*. Vol. 2. Ann Arbor, MI, USA. 2016.

[140] Wilko Schwarting et al. "Social behavior for autonomous vehicles". In: *Proceedings of the National Academy of Sciences* 116.50 (2019), pp. 24972–24978.

[141] Aarian Marshall. "How Robocars Handle the Frustratingly Human Act of Merging". In: *Wired* (2018). https://www.wired.com/story/self-driving-cars-merging-highways/.

[142] James Colyar and John Halkias. "US Highway 101 Dataset". In: *Federal Highway Administration (FHWA), Tech. Rep. FHWA-HRT-07-030* (2007).

[143] George T Taoka. "Brake reaction times of unalerted drivers". In: *ITE journal* 59.3 (1989), pp. 19–21.

[144] Vinayak V Dixit, Sai Chand, and Divya J Nair. "Autonomous vehicles: disengagements, accidents and reaction times". In: *PLoS one* 11.12 (2016), e0168054.

[145] "Intersection Safety". In: *Federal Highway Administration* (2020). URL: https://highways.dot.gov/research/research-programs/safety/intersection-safety.

[146] Akhil Shetty et al. "Risk Assessment of Autonomous Vehicles across Diverse Driving Contexts". In: *2021 IEEE International Intelligent Transportation Systems Conference (ITSC)*. IEEE. 2021, pp. 712–719.

[147] Michelle Grinnell. "Michigan, Cavnue Creating Road of Future Between Ann Arbor and Detroit". In: *Michigan Economic Development Corporation* (2020). https://www.michiganbusiness.org/press-releases/2020/08/michigan-cavnue-creating-road-of-future-between-ann-arbor-and-detroit/.

[148] Nidhi Kalra and Susan M Paddock. "Driving to safety: How many miles of driving would it take to demonstrate autonomous vehicle reliability?" In: *Transportation Research Part A: Policy and Practice* (2016).

[149] Philip Koopman and Michael Wagner. "Challenges in autonomous vehicle testing and validation". In: *SAE International Journal of Transportation Safety* 4.1 (2016), pp. 15–24.

[150] Maricopa Association of Governments. *MAG Fast Facts - Transportation*. https://azmag.gov/About-Us/Divisions/Transportation-Division/MAG-Fast-Facts-Transportation. 2018.

[151] Robert G Newcombe. "Two-sided confidence intervals for the single proportion: comparison of seven methods". In: *Statistics in medicine* 17.8 (1998), pp. 857–872.

[152] Nidhi Kalra and David G Groves. *The enemy of good: Estimating the cost of waiting for nearly perfect automated vehicles*. RAND Corporation, 2017.

[153] Andrew Hawkins. "Everyone hates California's self-driving car reports". In: *The Verge* (2020).

[154] Kyle Vogt. *The Disengagement Myth*. https://medium.com/cruise/the-disengagement-myth-1b5cbdf8e239. 2017.

[155] Philip Koopman and Beth Osyk. "Safety argument considerations for public road testing of autonomous vehicles". In: *SAE International Journal of Advances and Current Practices in Mobility* 1.2019-01-0123 (2019), pp. 512–523.

[156] Dennis Cinelli. *Getting Drivers Back on the Road*. Uber Newsroom. https://www.uber.com/newsroom/getting-drivers-back-on-the-road/. 2021.

[157] Hani S Mahmassani, Tian Hou, and Jing Dong. "Characterizing travel time variability in vehicular traffic networks: deriving a robust relation for reliability analysis". In: *Transportation Research Record* 2315.1 (2012), pp. 141–152.