Dynamic Ridesharing under Travel Time Uncertainty: Passenger Preference and Optimal Assignment Methods

by

Nathaniel K. Bailey

B.S. Industrial Engineering and Operations Research, University of California, Berkeley (2014)
M.S. Transportation, Massachusetts Institute of Technology (2016)

Submitted to the Department of Civil and Environmental Engineering
in partial fulfillment of the requirements for the degree of

Doctor of Philosophy in Transportation

at the

MASSACHUSETTS INSTITUTE OF TECHNOLOGY

May 2022

Author . . . .
Department of Civil and Environmental Engineering
April 14, 2022

Certified by . . . .
Jinhua Zhao
Associate Professor of Transportation and City Planning
Thesis Supervisor

Accepted by . . . .
Colette L. Heald
Professor of Civil and Environmental Engineering
Chair, Graduate Program Committee

Dynamic Ridesharing under Travel Time Uncertainty: Passenger Preference and Optimal Assignment Methods

by

Nathaniel K. Bailey

Submitted to the Department of Civil and Environmental Engineering
on April 14, 2022, in partial fulfillment of the
requirements for the degree of
Doctor of Philosophy in Transportation

Abstract

The increased prevalence and use of on-demand ridehailing services through the previous decade have had substantial impacts on urban transportation systems. These services offer convenient and flexible door-to-door transportation to users, but have also raised concerns about system equity, environmental sustainability, and efficiency given their high fares and their resulting increase in vehicle-kilometers traveled. Dynamic ridesharing (DRS; or pooled ridehailing) provides a means to mitigate these negative externalities by offering reduced fares for passengers in exchange for increased operational flexibility in pooling multiple trips into the same vehicle concurrently. However, even prior to their suspension during the COVID-19 pandemic, these services struggled to gain the widespread adoption needed to realize many of these benefits. One under-investigated barrier to the adoption of DRS services is travel time uncertainty. Travel times on urban road networks on which DRS services operate are often highly variable, and the potential for vehicle detours due to pooling increases travel time uncertainty for DRS passengers when compared to exclusive ridehailing.

This dissertation investigates the impact of travel time uncertainty, and traveler perceptions thereof, on decisions of whether to use pooling or not in ridehailing services, and formulates new methods to assign vehicles to passengers in DRS operation that improve passenger outcomes in terms of average delay and travel time variability. Using data collected from a survey of 1,600 Singapore residents, we estimate the impact of different presentations of information regarding travel time variability and associated attitudes on respondents' stated preference between exclusive and pooled ridehailing trips. We find that different forms of presenting information on the uncertainty of exclusive ridehailing journey times significantly alter passengers' responses to this uncertainty. However, travelers' decisions to use DRS are driven much more by their attitudes towards time uncertainty than by the magnitude or means of presentation of the uncertainty. We then formulate a two-stage stochastic optimization formulation for the online DRS assignment problem that minimizes average passenger delay in the presence of stochastic travel times. Through simulation experiments on synthetic road networks with stochastic, correlated travel times, we demonstrate the improved performance of this formulation in finding efficient solutions in a stochastic environment and reducing average passenger delay and the variance of passenger arrival times.

Overall, this dissertation finds that passengers' lack of trust in DRS reliability is a significant barrier to greater adoption and use, and also demonstrate the potential for operational methods that account for stochastic travel times to increase DRS reliability. This multidisciplinary exploration of the ramifications of travel time uncertainty on the supply and demand of DRS services provides a foundation for future research that may expand upon these concepts, further develop upon the methods used, and create connections between the supply and demand components.

Thesis Supervisor: Jinhua Zhao
Title: Associate Professor of Transportation and City Planning

Committee Member: Patrick Jaillet
Title: Dugald C. Jackson Professor in Electrical Engineering, Professor of Civil and Environmental Engineering

Committee Member: Paolo Santi
Title: Principal Research Scientist, Senseable City Lab

Acknowledgments

My time at MIT and my journey through the Ph.D. program have been extremely difficult but rewarding. I am thankful to so many people who have supported me during this journey and provided me with the energy I needed to successfully complete this dissertation.

To my thesis supervisor, Professor Jinhua Zhao, thank you for your dedication to your students and passion for research. The compassion you demonstrate for your students as people outside of our research contributions to the lab is inspiring and uplifting. The holistic support you provided as my advisor was instrumental in helping me reach this accomplishment.

To the other members of my dissertation committee, Professor Patrick Jaillet and Dr. Paolo Santi, thank you for the insightful perspectives you brought that helped refine my ideas for the research into a complete project with a clear story. Our discussions provided a wealth of valuable ideas and helped me better identify and articulate the contributions of this dissertation. Thank you as well to my Masters advisor, Professor Carolina Osorio, for helping me find a place at MIT and encouraging my development as a researcher.

To the many faculty, postdocs, and staff of the JTL Transit Lab who provided me with guidance over the years, thank you for establishing and fostering such a positive and supportive research environment that remains grounded in the realities and practicalities of transportation. The perspectives and insights shared by you all helped me to better understand transportation as a discipline and connect my research to the real-world challenges that arise in this field.

Thank you to my friends throughout my time at MIT, both those I met here as colleagues or teammates on the MIT Curling team, as well as those who I am grateful to have stayed in touch with despite the large physical distances separating us. The joyful times we shared helped to make this time of my life greatly rewarding despite many challenges and setbacks.

To my parents, Joe Bailey and Jessica Margolin, thank you for the environment of intellectual curiosity and emotional resilience that you created for me throughout my life. I am very fortunate for the strong foundation that you provided and continue to provide for me.

Finally, to my fiancée, Joanna Moody, you have provided me with so much love and support throughout this process that I can confidently say that I truly could not have done it without you. Thank you for sharing your understanding of the challenges that the Ph.D. presents, for providing me with the support I needed even when I didn't recognize it at the time, and especially for the sacrifices you made during the final push to ensure that I could remain focused on the work. Your kindness and dedication have been and will forever continue to be inspirational.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Background

In recent years, ridehailing (also referred to as ridesourcing or mobility-on-demand) services have become large components of urban transportation systems around the world. Ridehailing services, such as those provided by companies like Uber, Lyft, DiDi, and many other such providers around the world, allow for customers to request point-to-point transportation service on demand provided by a fleet of vehicles, often through the use of a smartphone app. Since the launch of Uber in 2009, ridehailing has rapidly become a significant component of transportation systems around the world, especially in urban areas, with the combined annual share of taxi and ridehailing trips increasing from 1.9 million in 2009 (0.2% of total trips) to 5.1 million in 2017 (0.5%), most of this increase attributable to ridehailing use [1].

These services offer several benefits to travelers, including convenience, accessibility, and flexibility. A traveler using a ridehailing service can order a trip via an internet connected device that arrives within minutes and delivers them within a few blocks at most of their desired location. The ability to order point-to-point trips can facilitate the adoption of car-free or car-light lifestyles among frequent users – for instance, 40% of San Francisco ridehailing users in 2014 who owned cars reported reducing the amount they drove since using these services [2]. These point-to-point trips can also be very beneficial in regions without frequent transit service, enabling increased accessibility for individuals who would otherwise rely on transit. The combination of a convenient booking interface with relatively low fares have enabled ridehailing to quickly acquire much greater mode shares than other forms of transportation which share some of these characteristics, such as taxi, carpooling, or dial-a-ride.

On the other hand, these services come with significant drawbacks, both to individual users and to the transportation systems they operate within. Compared to transit, ridehailing fares are often significantly higher [3], limiting the number of people who can make use of these benefits – for instance, preventing many who depend upon public transit options the ability to take advantage of the resulting increase in accessibility. Meanwhile, the additional deadheading distance traveled before arriving at a customer's location results in significant increases in vehicle miles traveled (VMT) per trip even compared with inefficient modes such as private car use [4, 5]. This increase in VMT can also result in other undesirable externalities such as greater emissions and increased congestion.

1.1.1 Dynamic Ridesharing

The development of dynamic ridesharing (DRS; also referred to as pooled ridehailing or ridesplitting) services offer another avenue with which to mitigate the negative externalities of ridehailing. Dynamic ridesharing allows for multiple trip requests to be pooled into a single vehicle, modifying that vehicle's route to serve all pickup and drop-off locations associated with all assigned requests. This often requires the travelers associated with each request to share the interior of the vehicle with strangers who they had not met prior to the ridehailing trip.

These services offer many theoretical benefits which would appear to improve upon many of the weaknesses noted for exclusive (non-pooled) ridehailing services. With a sufficient amount of demand for pooled trips, dynamic ridesharing could reduce the frequency of vehicles traveling empty and reduce the total deadheading distance traveled by assigning pickups and drop-offs to vehicles already planning to travel nearby. The increased efficiency of the service, capable of serving more passenger requests with a smaller number of vehicles, could also require lower costs on the operator, passing some of those savings on to the end consumer and reducing fares to increase affordability and economic accessibility.

Findings from the literature, both based on empirical data and using simulation, support these intuitions. Empirical data indicates that over 90% of pooled trips reduce total distance traveled compared with equivalent exclusive trips, resulting in a greater than 20% reduction in total travel distance and significant reductions in the resulting emissions [6]. Simulation experiments and other mathematical models indicate that fleets of pooled vehicles could greatly reduce the number of vehicles needed to feasibly serve all taxi trips with limits imposed on the delay each passenger can face, with estimates ranging from 25% to 50% of vehicles needed to serve all trips with a maximum allowable delay of 5 minutes per passenger [7, 8].

However, realizing the full scope of benefits demonstrated by these studies is difficult in practice due mainly to two factors. The first is low demand for these DRS services, which works counter to the network effects needed for large numbers of successful matches. The potential number of matches among trips within a ridehailing system depends on the spatial and temporal density of trip requests that arrive to the system within the coverage area [9]. This means that in areas with low density of DRS trip requests, many of the benefits the service offers compared to exclusive ridehailing services will not arise. However, despite the cost savings presented by pooled services, it is unclear if riders have adopted the service to the extent that is needed to realize these benefits – in 2016, only 20% of trips provided by Uber globally were pooled [10]. With the onset of the coronavirus pandemic leading to their suspension in many regions around the world, it is unclear whether the demand base to support these services at a level which allows them to realize their prospective benefits will exist when they return to service.

The second difficulty is presented by the uncertainty of travel times within the road network that these services operate on. Many of the simulation studies in particular rely on the assumption that travel times are deterministic, which has several major effects on system performance. The methods used to evaluate the shareability of individual trip requests rely largely on deterministic thresholds for the delays that passengers may face in a combined trip. In a stochastic travel time environment, however, the delay is uncertain at the time of assignment. Pooling trips together that are deemed shareable under these frameworks may result in combined trips which violate these deterministic time constraints in practice. Furthermore, the evaluation of these methods largely takes place in simulation environments with static travel times, thus failing to consider what impact stochastic travel times could have on system performance. These two points are also closely related, in the sense that uncertain travel times, the potential delays that arise from them, and passenger perceptions of these factors may reduce the interest in using these services.

1.1.2 Travel Time Uncertainty in Ridehailing

In reality, travel times faced by any vehicles, including ones serving ridehailing trips, traveling through urban road networks are well known to have many sources of uncertainty which would impact heavily any DRS service implemented in the real world. These sources include the degree of congestion on the road, the individual behavior of the driver, and, in urban settings typical of ridehailing services, a large suite of other factors such as traffic signal timings, maneuvers from stopped or stopping vehicles (e.g. buses, delivery vehicles, parked vehicles), or pedestrians and cyclists [11]. While some sources of uncertainty such as congestion may be incorporated into deterministic models of network travel times, many of these sources create unpredictable variations in individual vehicle travel times which cannot feasibly be incorporated into these models yet can result in significant delays.

DRS services additionally create another aspect of travel time uncertainty which impacts riders: the potential for diversions to pool with new requests that arrive to the system during travel. From a traveler perspective, it is unknown whether the trip will divert to pick up any other passengers during their trip, and if so, how many it will detour for. In practice, whether or not the vehicle diverts to serve one passenger location appears to make a large difference in the total trip duration. Previous research has found that each stop made during an UberPool trip adds an average of 5.76 minutes of travel duration [3], and that the average detour penalty (including trips which make no additional stops) is 3.6 minutes, with significant positive skew [12]. Though the empirical results are limited, these studies indicate that the number of diversions made can make a large difference in the timeliness of the resulting trip.

Despite these varied sources of uncertainty, DRS services do not appear to make significant efforts to inform customers regarding the level of travel time uncertainty associated with their trips or to dissuade worries that passengers may have about this unreliability. As of this writing, the booking interfaces that Uber and Lyft use in the United States (depicted in Figure 1-1) display a time range for pooled ridehailing trips, with no additional context indicating the meaning of this time range – for instance, the probability of arriving in this time range. Meanwhile, the exclusive service depicts no time range, instead using a single point value travel time estimate, despite the fact that exclusive trips are clearly subject to many of the same sources of travel time uncertainty that pooled trips are. This presentation may influence potential customers' opinions of each service and their trust in the pooled service's reliability.

Travelers' uncertainty regarding the duration of their trips when using DRS could possibly explain much of their demonstrated aversion to using these services. Meanwhile, the potential benefits of DRS when considering the realities of stochastic travel times on the road networks they operate within are poorly understood due to the assumptions made by many existing simulation studies. This leaves a research gap regarding the effects of travel time uncertainty in many components of DRS services, resulting in a critical lack of knowledge as to the degree to which these services may be able to accomplish their goals of mitigating the negative externalities of exclusive ridehailing services.

1.2 Contributions

The development of reliable and efficient dynamic ridesharing systems is important to attract riders who would otherwise use exclusive ridehailing and thereby reduce the negative impacts that these services have on transportation systems. This dissertation

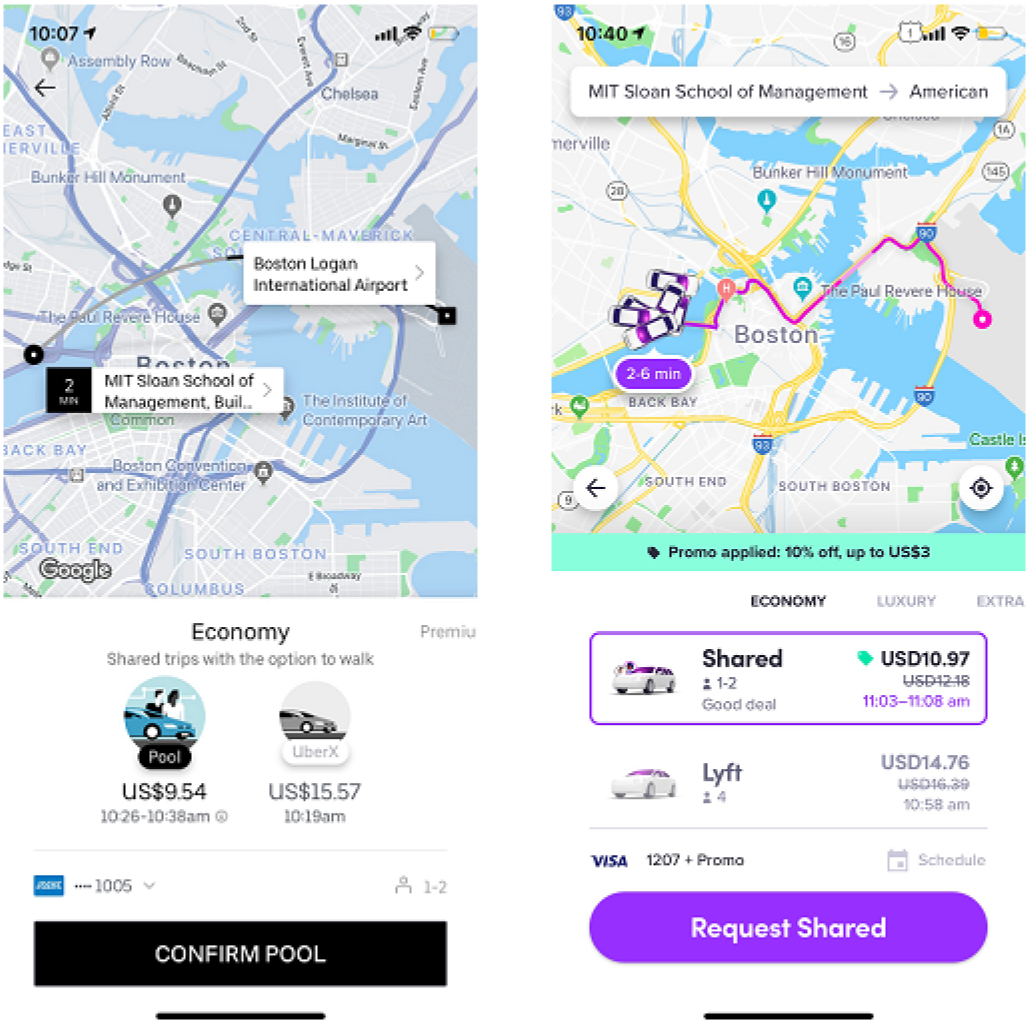


Figure 1-1: Examples of Uber and Lyft interfaces for booking exclusive and pooled ridehailing trips, from [13].

presents the results of multidisciplinary research that aims to answer several important questions of key relevance to the question of how dynamic ridesharing systems are influenced by travel time uncertainty. This research develops further understanding of approaches that ridehailing operators and transportation policy makers could use to improve the attractiveness of these systems to travelers and thereby generate this mode shift.

This broad question is narrowed into two main areas of focus by partitioning the demand and supply sides of dynamic ridesharing service. We investigate the impact of travel time uncertainty on the demand for dynamic ridesharing service within the narrow scope of traveler decision between using pooled and exclusive ridehailing services. By limiting the scope of the research to this decision between these two closely related modes, we are able to present a clear comparison and isolate our findings to the question of mode shift from exclusive to pooled services.

Within this scope of research, we consider how travelers' attitudes regarding travel time uncertainty and reliability differ between the two modes. After identifying these attitudes, we further explore differences between the two modes in how travel time uncertainty affects their decisions when choosing pooled or exclusive ridehailing. The differentiation of these impacts between exclusive and pooled service is novel in the literature, which has previously only considered the impact of travel time uncertainty in passenger preference for individual ridehailing modes [13–15].

On the supply side, we identify the assignment and routing of vehicles to passenger locations as an important operational decision which is impacted by variability of travel times. To date, only a handful of studies have investigated how to formulate ridehailing assignment problems that account for uncertainty [16, 17]; however, these studies place limits on the forms of uncertainty they consider, often using unrealistic distributional forms of travel times. We consider the problem of routing vehicles to pickup and drop-off passengers in an online ridehailing system with any possible travel time distributions describing the stochasticity of network travel times, and aim to minimize average passenger delay.

We further compare the performance of the new formulation with an existing method from the literature that assumes deterministic travel times. We specify several different network configurations that emulate characteristics of real-world travel time distributions and explore how urban form and the degree of travel time correlation impact performance. The results of simulation experiments on these networks are used to gain insights into the benefits of stochastic-aware assignment methods in an environment with uncertain travel times. These results inform how adopting these methods in the supply of DRS service could affect the characteristics of travel time uncertainty that affect preference to shift travelers to shift to this mode from exclusive ridehailing, as demonstrated by the results of our survey.

By tying together these three research components across a variety of topics relating to travel time uncertainty in DRS, this dissertation provides a strong framework for future research in this area. We investigate key research questions in how demand and supply are impacted by travel time uncertainty, and this investigation highlights the depth of these topics and the variety of interactions between demand and supply as well. The findings of these studies give insight into a variety of further actions that can be taken to increase the attractiveness of DRS service among ridehailing users and mitigate the negative externalities generated by these ridehailing services. These findings are relevant to operators of ridehailing services as well as regulators or transportation policymakers interested in enhancing the sustainability of transportation systems which include these services.

1.3 Dissertation Structure

This dissertation research addresses the many aspects associated with this research gap using a multidisciplinary approach. The research investigates how both the demand and supply sides of pooled ridehailing operation are impacted by travel time uncertainty. On the demand side, we use the results of a stated preference survey of 1,600 Singapore residents to investigate ridehailing users' attitudes regarding travel time uncertainty in both exclusive and pooled ridehailing, and how information presented on trip travel time uncertainty affects preference between using each of these modes. This research is presented in Chapter 2.

On the supply side, we formulate a stochastic optimization dynamic ridesharing assignment problem that can be used to minimize average passenger delay accounting for travel time variability using a distribution-free approach. The details of this formulation as well as experiments demonstrating its computational performance are given in Chapter 3. We further compare the results of this assignment method to a deterministic benchmark method using simulation experiments in Chapter 4. These experiments cover a variety of network configurations and travel time distributions to understand how the incorporation of travel time stochasticity into the ridehailing operation method can impact system performance in different settings.

Chapter 2

Preference

2.1 Introduction

As ridehailing services proliferate in urban areas around the world and make up an increasing portion of mode share, there has come a corresponding increase in congestion and road usage which result in increased emissions, traffic fatalities, and other negative externalities [4, 18]. Pooled ridehailing services, where riders may be grouped with other unknown riders with similar routes, present a potential way to mitigate these issues by increasing vehicle occupancy and decreasing the number of vehicles and amount of road usage needed to serve a specific trip volume. Recent research has found that pooled ridehailing results in decreases of travel distance by 22% compared to exclusive ridehailing, with substantial emissions reductions per pooled trip served [6]. Despite the increasing prevalence of pooled options in many major ridehailing platforms and the potential benefits they present, most ridehailing trips continue to serve single passengers, with recent estimates suggesting an average of 1.4 passengers per ride and distance-weighted vehicle occupancy of 0.8 [4].

Because pooled rides are consistently less expensive than their exclusive counterparts, and information on both is often accessible within the same interface, this low mode share can be attributed to consumer preferences which make pooled rides comparatively unattractive. Pooled trips on average can be expected to have longer journey times than exclusive rides due to the possibility of detours to pick up or drop off other passengers along the route. Beyond these basic utilitarian differences in cost and average travel time, pooled services also feature many other aspects which may influence passengers' decision making. These include the sharing of the vehicle with other passengers, which is associated with different types of social interactions and different levels of safety and security when compared to exclusive rides, as well as a greater degree of uncertainty regarding the wait and travel time, as unforeseen detours may be added to the vehicle's route on its way to pick up or drop off individual passengers.

Travelers' beliefs and attitudes regarding these different aspects of pooled services may also play key roles in their demonstrated reluctance to use pooled ridehailing services by making pooled rides more or less appealing than they would be solely on the basis of cost and average travel time alone in comparison to a competing exclusive option. Examples of these attitudes that might increase willingness to use pooled rides might include excitement about meeting new people or a greater sense of safety with other people in the vehicle besides the driver. Examples that might decrease willingness to pool could include fear of unwanted social interactions with fellow passengers or a belief that the travel time of pooled ridehailing is less reliable than that of its exclusive counterpart.

Many previous studies have investigated the impact of factors such as cost, travel time, wait time, trip purpose, time of day, sociodemographic characteristics, household attributes, and the built environment on decisions to use pooled services [15, 19–21]. Several other studies have investigated the attitudes that lead individuals to consider choosing to use a pooled ridehailing service and those that lead them to avoid using it [22–24]. A handful of studies have additionally focused on travel time uncertainty and quantified the impact of travel time variance on choices between different trips within the same ridehailing service [13, 14, 25]. These works provide a solid body of literature which together provide a holistic overview of how many of these factors influence traveler choice.

However, we identify a few gaps within this literature, specifically focusing on travel time variability. We identify this variability as something that behaves differently between the exclusive and pooled modes, and its importance to ridehailing decision-making is evidenced by the investigation of ridehailing user attitudes [23]. However, the only studies which aim to quantify the impact of this uncertainty investigate its impact solely on one mode or the other. Therefore, the current literature provides no insight into how travelers may place different values on reliability when using pooled ridehailing compared to exclusive ridehailing. We also have poor understanding of the types of attitudes that travelers have regarding the degree of uncertainty in these two modes, and how these attitudes influence decisions between exclusive and pooled ridehailing.

We also note that ridehailing is an interesting context to study the impact of travel time uncertainty, as it is often a quantity explicitly represented in the booking interface when a traveler is presented with a pooled and exclusive option, as in Figure 1-1. The prevalence of real-time information on the degree of uncertainty in arrival time is important, as previous studies in the context of public transit have found that real-time information may increase ridership and satisfaction, but that it can also generate negative experiences when this information is found to be in error [26, 27]. Additionally, the presentation of this information differs from the traditionally recommended format of presenting travel time variability information in stated preference surveys, creating a disconnect between survey methods and realistic situations [28]. To this point, no study has holistically compared how different presentations of variability information affect traveler choice between pooled and exclusive service.

In this chapter, we present research that addresses these questions by use of a stated preference experiment administered via an online survey of 1,600 Singapore residents. Using several varied stated preference question designs, we investigate how passengers respond to wait time and journey time uncertainty in pooled and exclusive modes when presented in different formats. The main research objectives are to estimate the value of time variability in exclusive and pooled ridehailing and to develop an understanding of how travelers' attitudes regarding travel time uncertainty in ridehailing and the presentation of variability information to the traveler impact this value of time variability.

In the following section, we discuss in further detail the relevant works of literature that investigate related issues in pooled ridesharing and travel time uncertainty. We then discuss the methods used for this research, including a description of the data collection process, relevant descriptive statistics of our sample, and the factor analysis and choice modelling process applied to the data. Then, we describe the results of three binomial logit models estimated using the data and describe the significant findings from each. We finally conclude by synthesizing and discussing the results from all three models and identifying limitations of our study and key areas of future research.

2.2 Literature Review

2.2.1 Factors Influencing Ridehailing Adoption and Use

With the increasing prevalence and use of on-demand ridehailing services in the past decade, several studies have modeled travelers' decision-making within these modes as a function of explanatory variables including cost, travel times, socioeconomic factors, and the built environment [15, 19–21, 29]. These studies investigate what prompts travelers to choose to take a pooled ridehailing trip as opposed to an exclusive one using stated or revealed preference data to create quantitative models of these decisions, most often with a multinomial logit choice framework.

Few of these studies differentiate between the value of travel time (whether wait time or in-vehicle travel time) between exclusive and pooled ridehailing, though they generally find that value of time for travelers using ridehailing is comparable to that of car users in the literature. However, pooled services are generally found to provide lower utilities to travelers, with more negative alternative specific constants. [20] estimate that the additional cost of sharing a vehicle with one passenger is equivalent to $0.62 for commute trips, $1.32 for leisure trips, and $1.70 for shopping trips (equivalent to 1.4, 2.9, and 3.7 minutes of travel time, respectively).

Individual characteristics including ethnicity, age, income, and education are explored by many of these studies. White Americans appear to use pooled ridehailing less frequently than other groups, though ethnicity is only explored by [20] out of the reference papers. Age is consistently found to influence both ridehailing use in general and pooling use specifically, with younger people more likely to use ridehailing and more likely among ridehailing users to use pooling. Higher education is also linked to greater use of both exclusive and pooled ridehailing.

While most studies find that greater incomes are associated with greater use of ridehailing, there are mixed results from studies that explore its relation to pooled use. [19] look only at neighborhood-level effects but find that trips originating in or traveling to higher income regions are less likely to be pooled . Meanwhile, [21] find that higher income residents of Singapore are more likely to have used pooling, but use pooling less frequently than low-income travelers. These mixed effects are likely the result of income's association with ridehailing use in general combined with the appeal of discounted fares to lower-income users.

Household characteristics are also influential on these decisions. Houses with lower vehicle ownership are more likely to use ridehailing services and more likely to use pooling for their ridehailing trips. While many studies find no significant effect of total household size on pooled or exclusive ridehailing use, [29] find that individuals who live in households with no children are more likely to use ridehailing in general.

The built environment also appears to be a significant factor. Travelers who live in areas with a greater density of population and/or activity are generally found to engage in greater use of ridehailing services. Likewise, high population density is found to associate strongly with more frequent pooling of ridehailing trips.

Finally, trip purpose appears to play a significant role in travelers' willingness to pool their trips. [19] find that ridehailing trips traveling to or from the airport are significantly less likely to be pooled, with airport-bound trips especially so. As mentioned previously, [20] find that the cost of pooling is highest for shopping trips and lowest for commute trips. These findings generally point to pooling being used less for time-sensitive trips, although it may also be influenced by factors such as the available room in the car for trip purposes such as shopping.

2.2.2 Time Uncertainty and Ridehailing Choice

In addition to the common explanatory variables used in the works above, passengers' uncertainty regarding waiting and travel time when using ridehailing modes may also be a significant factor in explaining their decision-making processes. Many studies have investigated how variability of travel times and service reliability affect travelers' choices in other transportation contexts [30, 31]. Despite the clear relevance of travel time uncertainty to ridehailing, however, only a few studies have used traveler data to quantify the impact that it has on traveler choice.

[25] conducted a discrete choice experiment using a stated preference survey in New York in 2017 that elicited preferences regarding the estimated wait time and delay of a ridehailing service (whether the service was exclusive or pooled does not appear to have been specified). In the survey, passengers were presented with the decision to book a ride with one of two services, one of which had a longer wait time but 100% reliability (a deterministic pickup delay of 0 minutes), and the other of which had a shorter displayed wait time but an average pickup delay between 3 and 13 minutes. Using a binary logit model, they estimate the passengers' likelihood to choose the variable service. For an operator aiming to choose a displayed wait time that maximizes passengers' likelihood of choosing this service, this optimal displayed wait time increased as the variability of the wait time distribution increased. However, the relevance of this insight is unclear due to the lack of a 100% reliable counterpart service in real contexts, and the study also fails to investigate several components of interest such as the impact of total journey time reliability on choice.

A more traditional approach was taken by a series of stated preference experiments conducted in the Netherlands [14]. Respondents to the survey were presented with choices between two different options for a pooled ridehailing ride which additionally required a short walk (1 minute on average) to reach the pickup point, rather than being a door-to-door service. The first phase of experiment showed each option with an associated cost and five equally likely waiting times, while the second phase showed each option with an associated cost and five equally likely in-vehicle travel times. The mean-variance method was used to model the disutility of variability as an explanatory variable of the observed choices. The authors performed a latent class choice model analysis, with respondents separated into different classes representing different market segments by the results of their stated preference choices alone.

The results of this study found a value of time (VOT) for the waiting stage of the ridehailing journey between €9.3-16.5 per hour, and between €7.9-10.8 per hour for the in-vehicle stage, with both depending on trip purpose. They additionally find value of reliability (VOR) between €3.2-5.8 per hour of standard deviation of travel time for the in-vehicle stage, and between €2.2-8.6 per hour of standard deviation for the waiting stage. The reliability ratio (RR), the ratio between the VOR and VOT, were found to be around 0.5 for the in-vehicle stage and between 0.2 and 0.7 for the waiting stage. The latent class analysis additionally finds significant differences in respondents in terms of the disutility experienced by the waiting stage and cost sensitivity, although the RR was relatively stable across market segments. However, as all choices in this study were between two pooled ridehailing options, any differences in VOT or VOR between exclusive and pooled services were not identified.

Finally, [13] conducted a stated preference survey in the United States and used the results to estimate a discrete choice model of consumer preferences regarding the attributes of different ride-hailing services. The choice tasks were modeled based on the interface of Uber and Lyft ridehailing apps as they appeared at the time of the study, with exclusive modes showing a single arrival time estimate and pooled modes displaying a range of possible arrival times (formatted as, e.g., 11:03-11:08am). Respondents were therefore presented with a choice between an exclusive ridehailing option with an associated cost, single valued pickup time, and single valued travel time and a pooled ridehailing option with an associated cost, single valued pickup time, and travel time range. The parameters of the fitted binomial logit model found that the width of the displayed time interval had a significant negative impact on respondents' utility of the pooled option: an additional minute of travel time uncertainty was valued equivalently to 0.4 minutes of additional travel time. However, this study had several limitations in developing a general understanding of the value of uncertainty in ridehailing, as neither variability of wait time nor variability of the exclusive option's travel time was considered.

2.2.3 Attitudes Towards Ridehailing Services

Beyond utilitarian considerations such as cost, travel time, and reliability, travelers' decisions are also shaped by their attitudes towards ridehailing services. This is described in a general decision-making framework by the Theory of Planned Behavior [32], and is widely cited and empirically validated in the transportation context specifically [33]. Studies regarding attitudes in the ridehailing context investigate individuals' motivations for using or avoiding these services, attitudes regarding using shared ridehailing as opposed to the exclusive form of the service, and attitudes regarding the uncertainty of travel time in these modes.

[22] conducted a survey of 997 users of Uber or Lyft who lived in metropolitan regions where UberPool and Lyft Line were operating. The survey investigated the attitudes of both users who had used these pooled services and users who hadn't. The study found that users of these services reported the positive and negative social aspects of the service to be influential in their decision to use pooled services, although not as significantly as utilitarian factors such as speed or cost. Additionally, respondents reported that the risk of negative social interactions was more likely to cause them to not use pooled services than the potential for positive social interactions was to cause them to use them. Finally, prejudice and other harmful attitudes appeared to play significant roles in use, as several riders were found to hold negative attitudes towards other passengers of different social class or race and would prefer to use a pooled service if they had information about their fellow passengers prior to being matched. Additionally, women reported that safety was an important factor, as many reported feeling unsafe or intimidated during rides as reasons they would avoid using a pooled service.

A survey of Mexican users of DiDi's *express* (exclusive) and *comparte* (pooled) services investigated the travel behavior of users of each of these services, and how different factors lead people to decide to use either the exclusive or pooled service

[23]. Notably, this survey was conducted during the summer of 2020 during the COVID-19 pandemic. This study found that among Mexican ridehailing users, the level of safety of the service was the most important factor for considering using a pooled service over an exclusive one. Additionally of note, the trustworthiness of the estimated trip time was ranked above factors such as estimated trip travel time and price in terms of importance, which are traditionally thought to be among the most important drivers of transportation decision-making. Both users and non-users of the comparte service ranked having a better estimation of how long the pooled trip will take as the most important factor that would encourage them to use the pooled service more frequently. Reduced uncertainty was also shown to be an important driver of adoption through better knowledge how many passengers would join during the trip, which was found to be more important to those who solely used the exclusive service than those who had already used the pooled service. These findings differ with the traditional understanding that reliability is less important than cost and travel time in using pooled ridehailing, and are also of interest due to the study's rarity in investigating these issues in a less developed economy.

[24] used a 2019 online survey questionnaire of German customers interested in potentially using pooled ridehailing services to investigate the critical acceptance factors toward use of these services. The structural equation model approach drew upon constructs from the technology acceptance model and the Unified Theory of Acceptance and Use of Technology 2 (UTAUT2). The results of this model were used to develop several recommendation for pooled ridehailing providers. Of particular interest, the study found that intention to use pooled ridehailing was not influenced greatly by discussion of the monetary advantages; rather, the authors recommend that providers should focus on the inherent qualities of the transportation service itself. In addition, the perceived performance of the ridehailing service itself was highly influential on intention to use the service, leading the authors to conclude that these services should allow "quick and reliable transportation" in order to incentivize adoption.

Despite these theoretical considerations of uncertainty in ridehailing, including the description of various attitudes regarding travel time uncertainty and quantifying its importance in decision-making through choice models, our empirical understanding of this uncertainty in real-world operations is limited. However, [3] collected primary data from a small sample of 50 trips made using UberPool in Chicago, Illinois in 2016 at a variety of locations throughout the city and times of day. This research found that the average UberPool trip involved 0.87 intermediate stops, with each adding an average of 5.76 minutes of journey time. This represents a significant source of uncertainty for trips that averaged roughly 35 minutes of travel time. Furthermore, the research found that the average wait time for pickup was 7.3 minutes for UberPool trips, and that over 10% of the trips featured surge pricing with up to 60% increases in fare. Although these data do not specifically address the question of how the actual journey times compare to estimates provided by UberPool upon booking the trip, it provides helpful context in understanding the degrees of uncertainty surrounding wait time, journey time, and cost for real-world trips.

2.3 Methods

2.3.1 Data Collection

We conducted an online survey of Singapore residents in April 2020 to investigate how decisions between using exclusive or pooled service for ridehailing trips are made. Singapore has high rates of ridehailing use and access to cutting-edge transportation, information, and communication technologies, making it an interesting setting for this experiment. The survey was distributed to respondents sampled based on demographic quotas for age, gender, income, and household car ownership matching the general population of Singapore. The survey included questions about respondents' sociodemographic characteristics, travel behavior, attitudes towards and prior use with ridehailing platforms, and behavioral response to the concurrent coronavirus pandemic. It also featured a stated preference questionnaire used to solicit hypothetical travel decisions made between exclusive and pooled ridehailing trips in order to collect data for a discrete choice model, which will be discussed later.

To elicit responses to the stated preference questions that matched as closely as possible to real travel decision-making behavior, we prompted respondents with a specified trip purpose (home to work commute trip, work to home commute trip, home to shopping destination, or social/recreation activities to home) and asked them to recall a recent trip made for this purpose (the “reference trip”). We then asked for detailed information about the reference trip, including the day of week and time of day that it was taken, the primary transportation mode used (defined as the one covering the greatest distance in the case of multimodal trips), and how many other people traveled with them. In addition, we collected the trip origin and destination locations to an accuracy of 200m and estimated the duration of a driving trip between these locations at the appropriate date and time using the Google Directions API. These data were used as inputs to the stated preference questions to create choice scenarios which resembled real-world travel conditions similar to those each respondent encountered when undertaking their reference trip.

2.3.2 Descriptive Statistics

Using the sociodemographic data collected from respondents, we can compare this data to official census data collected by the Singapore Department of Statistics in order to understand the representativeness of our respondents. This information is shown in 2.1. Our sample is significantly younger and more educated than the general Singapore population, as is common for surveys relying on online data collection. Additionally, we find that our sample has somewhat higher household car availability, and is concentrated more in the middle and lower-middle incomes.

The familiarity of respondents with ridehailing services is important in order for them to have meaningful understanding of the scenarios presented to them in the choice experiment. Table 2.2 displays a summary of respondents’ answers to selected questions regarding the travel behavior of respondents. 90% of the respondents to our survey have previously used on-demand ridehailing apps, defined in the survey as any service that allows you to book rides in private cars via a smartphone app

Sociodemographic Category		**Total**	**Sample %**	**Singapore Population %**
Age	18-24	286	17.9%	8.7%
	25-34	358	22.4%	18.4%
	35-44	276	17.3%	20.2%
	45-54	394	24.6%	20.3%
	55-64	204	12.8%	17.5%
	65-74	64	4.0%	9.3%
	75-84	10	0.6%	4.3%
	84+	0	0.0%	1.4%
	Prefer not to say	8	0.5%	-
Monthly Income	<$2,500	232	14.5%	20.4%
	$2,500-$5,999	442	27.6%	19.8%
	$6,000 - $9,999	437	27.3%	21.2%
	$10,000 - $14,999	282	17.6%	17.4%
	≥ $15,000	207	12.9%	21.1%
Education Level	Primary or less	12	0.8%	20.6%
	Secondary	192	12.0%	27.3%
	Post-secondary	97	6.1%	9.1%
	Polytechnic or other diploma	404	25.3%	14.7%
	University	895	55.9%	28.2%
Household Car Availability	0	710	44.4%	57.9%
	1 or more	890	55.6%	42.1%

Table 2.1: Comparison of sociodemographic data recorded from survey respondents with the population statistics of Singapore. The distribution of age groups within the Singaporean population is calculated out of all residents over 18 years of age to reflect our survey frame.

– Grab and Gojek are listed as examples of the service in the Singapore market. Furthermore, over 70% of users who have previously used any of these services report having pooled services, referred to as "ridesharing" in the survey for greater familiarity, with the examples of GrabShare and RydePool.

These findings are much higher than those of comparable studies in the United States, which find ridehailing users to be less than 50% of the population [13, 25], but is more in line with a previous study of Singapore residents which found that 82% of respondents had used ridehailing, though only 55% of those respondents had

Age	**Ridehailing Familiarity**			
	Have used both E and P	Have used E, aware of P but haven't used	Have used E, not aware of P	Have never used E or P
18-24	201	57	5	23
25-34	295	54	4	5
35-44	204	60	4	8
45-54	212	120	5	57
55-64	97	65	2	40
65-74	20	26	0	18
75-84	3	6	0	1
Prefer not to say	3	4	1	0
Total	1035	392	21	152
%	64.7%	24.5%	1.3%	9.5%

Table 2.2: Reported level of familiarity with exclusive (E) and pooled (P) ridehailing services among respondents to the survey, categorized by age group.

used a pooled service, somewhat lower than the 70% of our sample [21]. As greater than 60% of our survey respondents have direct experience with pooled ridehailing services, this high level of familiarity with the context of our choice experiments indicates that their responses may reflect fairly accurately the real-world behavior they would exhibit upon encountering these situations, increasing the power of our findings from this study.

We also note that the level of familiarity with ridehailing varies by age group, with younger respondents more likely to have used both exclusive and pooled services than older ones. Among respondents under the age of 45, over three quarters had previously used a pooled ridehailing service, while just under half of respondents of at least that age had done the same. The youthful bias of the respondents to the survey would therefore appear to result in a greater than average familiarity with these services than would be found in a representative sample of the general public.

We are also interested in attitudes and perceptions among our respondents regarding exclusive and pooled ridehailing. Our survey questions explored many dimensions of these psychological factors: the importance of different factors in respondents' decision-making between pooled and exclusive ridehailing options when presented with a choice, their experiences in previous ridehailing trips, their trust in the information provided by these services, the reasons they choose to use or avoid pooled trips. The answers to these questions provide qualitative information on the thought processes that inform travelers' choices, allowing us to better interpret the results of the choice experiments which provide only quantitative information on the choices themselves.

Table 2.3 displays aggregated statistics summarizing respondents' answers to several 5-point Likert scale questions regarding the importance of various trip features for exclusive and pooled ridehailing in their decision-making behavior. These results are grouped by whether or not the respondent reports any previous use of a pooled ridehailing service. In addition, it displays the mean importance of each factor across each group and across the full sample, obtained by converting the responses into integer values from 1 to 5, with 1 representing "not at all important" and 5 representing "extremely important".

We observe that the most important factors influencing respondents' choices are the cost of the private ridehailing mode, the possibility of arriving late in each mode, and the cost of the pooled ridehailing mode. Meanwhile, the least important factors on average are the possibility of being matched with others and the estimated pickup time for each mode. It is important to note that the possibility of arriving late is ranked similarly to the price of each mode, suggesting that travelers' perceptions of the reliability of each mode will strongly impact their choice between the two.

Moreover, we note an interesting discrepancy between the responses of those who have previously used pooling and those who haven't. For the majority of questions, respondents who have never used pooling report a lower average importance for that factor in their decision-making than respondents who have. This makes intuitive sense, as respondents who have previously used pooled services may be more open to either option, and therefore when making a decision between the two will consider more aspects of the choice. Meanwhile, respondents who have never used pooling may be more predisposed to using the exclusive ride option, therefore placing less

		Have used pooled ridehailing						Have used exclusive ridehailing but not pooled						Overall
		E (5)	V (4)	M (3)	S (2)	N (1)	Average	E (5)	V (4)	M (3)	S (2)	N (1)	Average	Average
Exclusive	Cost	30.7%	36.4%	21.6%	8.8%	2.4%	3.84	24.5%	31.7%	25.2%	12.1%	6.5%	3.55	3.76
	Estimated journey time	19.7%	34.4%	29.4%	13.4%	3.1%	3.54	17.9%	31.7%	31.5%	12.8%	6.1%	3.43	3.51
	Estimated pickup time	16.1%	31.8%	31.1%	17.3%	3.7%	3.39	15.0%	27.1%	30.8%	20.3%	6.8%	3.23	3.35
	Possibility of arriving late	26.1%	33.8%	23.4%	12.9%	3.8%	3.66	22.0%	33.4%	26.2%	13.1%	5.3%	3.54	3.62
	Uncertainty of journey time	21.0%	32.4%	28.9%	13.1%	4.6%	3.52	18.9%	29.5%	32.7%	13.8%	5.1%	3.43	3.49
Pooled	Cost	25.6%	37.5%	21.5%	10.8%	4.5%	3.69	21.1%	32.4%	21.8%	10.9%	13.8%	3.36	3.59
	Estimated journey time	17.3%	31.7%	31.8%	14.9%	4.3%	3.43	16.7%	33.2%	22.8%	10.9%	10.4%	3.35	3.40
	Estimated pickup time	14.3%	28.5%	31.6%	18.7%	6.9%	3.25	13.3%	25.7%	30.0%	19.1%	11.9%	3.09	3.20
	Possibility of arriving late	23.2%	31.5%	29.2%	12.5%	3.7%	3.58	27.4%	34.4%	23.5%	8.2%	6.5%	3.68	3.61
	Uncertainty of journey time	17.6%	32.4%	31.7%	13.3%	5.0%	3.44	23.7%	34.9%	23.5%	10.4%	7.5%	3.57	3.48
	Possibility of being matched with strangers	14.7%	22.0%	27.1%	20.3%	15.9%	2.99	24.7%	24.5%	26.4%	12.8%	11.6%	3.38	3.10

Table 2.3: Summary of data describing how important different components of exclusive and pooled ridehailing trips to travelers making a decision between two competing trips, categorized by whether or not the respondent had previous experience using pooled ridehailing trips or not. The columns indicate what proportion of respondents rated each factor as extremely important (E), very important (V), moderately important (M), slightly important (S), or not at all important (N).

importance on the specific options presented to them. However, the factors for which this does not hold true are the possibility of arriving late for the pooled mode, the uncertainty of journey time in the pooled mode, and the possibility of being matched with others. The higher mean importance placed on these factors by those who have never used pooling may indicate a strong relationship between concerns over the reliability of the pooled service and passengers' willingness to use it, although we cannot ascribe causality to this relationship from this data.

Related to this insight, the survey also included questions regarding the aspects of pooling that may attract or repel respondents from using it. These questions used a 3-point Likert scale format, with the response options being "agree", "neither agree nor disagree", and "disagree" to each statement regarding why the respondent might choose to use pooled ridehailing or might avoid to use it. The survey results for these questions are presented in Tables 2.4 and 2.5.

From these results, it appears that the strongest reasons that travelers choose to pool Is that it is more comfortable than other affordable transportation options such as transit, biking, or walking, and that it is beneficial for the environment when compared to exclusive rides. Social benefits appear to be less influential in promoting pooling, with respondents erring towards disagreement with statements that safety or meeting people are significant reasons they might choose to use a pooled service. However, we note that agreement with these statements was much more common among those who have used pooling in the past, suggesting that these social benefits may be enjoyed upon discovering them when using the service, but that they are not influential in getting people to try the service for the first time.

Meanwhile, we find that social factors are much more commonly cited as reasons why travelers might choose not to use pooling. Respondents' levels of agreement with statements regarding preferring privacy in the backseat, being afraid of unpleasant social interactions, and being uncomfortable with a lack of norms on social interactions with fellow passengers are comparably high to the most commonly agreed upon reasons to use pooling discussed previously.

Finally, given the state of the coronavirus pandemic at the time that the survey was

Reason	**Have used pooled ridehailing**				**Have used exclusive ridehailing but not pooled ridehailing**			
	Agree	Neither	Disagree	Avg.	Agree	Neither	Disagree	Avg.
A car is more comfortable than transit, biking, or walking.	75.4%	19.5%	5.1%	2.70	68.0%	24.9%	7.0%	2.61
Sharing rides is better for the environment.	59.8%	29.0%	11.2%	2.49	47.2%	39.7%	13.1%	2.34
I feel safer having another person in the car other than the driver.	22.4%	39.4%	38.2%	1.84	12.6%	40.0%	47.5%	1.65
I want to meet people heading to or coming from the same event as me.	30.4%	35.5%	34.1%	1.96	17.4%	38.5%	44.1%	1.73
I enjoy meeting and making small talk with people from different social circles.	31.2%	34.3%	34.5%	1.97	16.9%	35.8%	47.2%	1.70

Table 2.4: Respondent agreement with statements that they might choose to book an on-demand shared ride instead of other modes because of each of the listed reasons. Responses are categorized by whether the respondent had or hadn't used shared ridehailing previously.

being conducted, we additionally asked specifically about infection-related reasons to use or avoid pooling. Table 2.6 displays respondents' answers to questions on the topic of the pandemic. Respondents were ambivalent regarding the statement that less risk of infection on pooled ridehailing compared to transit was a reason they might choose to use pooling. However, they agreed with statements regarding both the risk of infection due to a stranger's car and that due to sharing the car with fellow passengers being reasons they might avoid using pooled ridehailing. These attitudes would appear to indicate that ridehailing use would be reduced if the population has concerns about the pandemic or other risks of infection. However, the similar level of agreement with statements regarding both the driver's car and the fellow passengers as potential sources of risk could indicate that travelers' attitude toward pooling vs. exclusive ridehailing may not be large. We note that these results are from very early during the timeline in the pandemic, and that updated results which may reflect the

Reason	Have used pooled ridehailing				Have used exclusive ridehailing but not pooled ridehailing			
	Agree	Neither	Disagree	Avg.	Agree	Neither	Disagree	Avg.
I prefer privacy in the backseat of the car.	68.9%	24.2%	7.0%	2.62	75.5%	18.9%	5.6%	2.70
I am afraid of unpleasant social interactions with other passengers.	53.9%	32.0%	14.1%	2.40	64.6%	28.6%	6.8%	2.58
There are no clear norms of interaction.	42.8%	47.5%	9.7%	2.33	45.5%	47.0%	4.5%	2.63

Table 2.5: Respondent agreement with statements that they might choose not to use an on-demand shared ride instead of other modes because of each of the listed reasons. Responses are categorized by whether the respondent had or hadn't used shared ridehailing previously.

greater level of understanding many people have regarding the pandemic and the risks of infection would be helpful in understanding how these attitudes shape ridehailing decision-making now and in the future.

The final attitudes explored directly in the survey regard travelers' perceptions of service reliability for both pooled and exclusive services. Respondents were asked to think about their prior experiences when using each form of ridehailing that they had first-hand experience with and to estimate the frequency with which a ride using that service would arrive within the estimated drop-off time window provided at the start of the trip. They were additionally asked to report their trust in these estimated time windows for both exclusive and pooled trips as long as they had used either service in the past. Figure 2 displays the results of these questions.

Respondents' perceptions of the on-time performance of pooled and exclusive ridehailing appear to vary significantly. While 58.5% of respondents who have used exclusive ridehailing report arriving within the displayed time range over half of the time for exclusive rides, only 26% of respondents who have experience with pooled ridehailing say the same for that service. Taking the midpoint of each provided confidence range (e.g. 70% for the 50-90% range) and averaging across responses, we find that the average reported on-time arrival frequency for the exclusive mode is 62.6%, while

Reason	**Have used pooled ridehailing**				**Have used exclusive ridehailing but not pooled ridehailing**			
	Agree	Neither	Disagree	Avg.	Agree	Neither	Disagree	Avg.
There is less risk of infectious diseases compared to public transit.	39.0%	32.4%	28.6%	2.10	30.5%	38.5%	31.0%	2.00
Sharing the ride with other passengers increases my risk of contracting infectious diseases.	68.5%	24.3%	7.1%	2.61	68.0%	27.1%	4.8%	2.63
Riding in a stranger's car increases my risk of contracting infectious diseases.	58.4%	32.3%	9.3%	2.49	59.8%	34.3%	5.9%	2.54

Table 2.6: Respondent agreement with statements that they might choose to use or not use shared on-demand rides because of several reasons relating to the ongoing COVID-19 pandemic at the time of the survey in April 2020.

for pooled ridehailing respondents on average believe they arrive within the shown time range just less than half the time at 49.5%. This widespread belief that shared ridehailing exhibits worse reliability than exclusive ridehailing could be due to several reasons. First, it could reflect real-world behavior of these systems that lead shared ridehailing services to suffer from worse reliability than exclusive ones; the lack of empirical data makes this claim hard to verify or refute. Alternatively or in addition, some psychological process could result in passengers' perceptions exaggerating this difference – for instance, late arrivals when using pooled ridehailing could be more salient in travelers' memories and shape their attitudes more strongly.

Meanwhile, respondents' reported levels of trust in the time estimates shown in ridehailing booking interfaces prior to requesting a ride also differ greatly between the two modes. While nearly three quarters of respondents report some degree of agreement with the statement "I trust the arrival times for private rides," only half of respondents agree with the identical statement for pooled rides. The average agreement, as measured by a 7-point Likert scale, differs by nearly 1 point (exclusive: 5.19; pooled: 4.27). Combined, the reported attitudes for these two questions indicate a large gap in both the perception and experience of reliability between pooled and

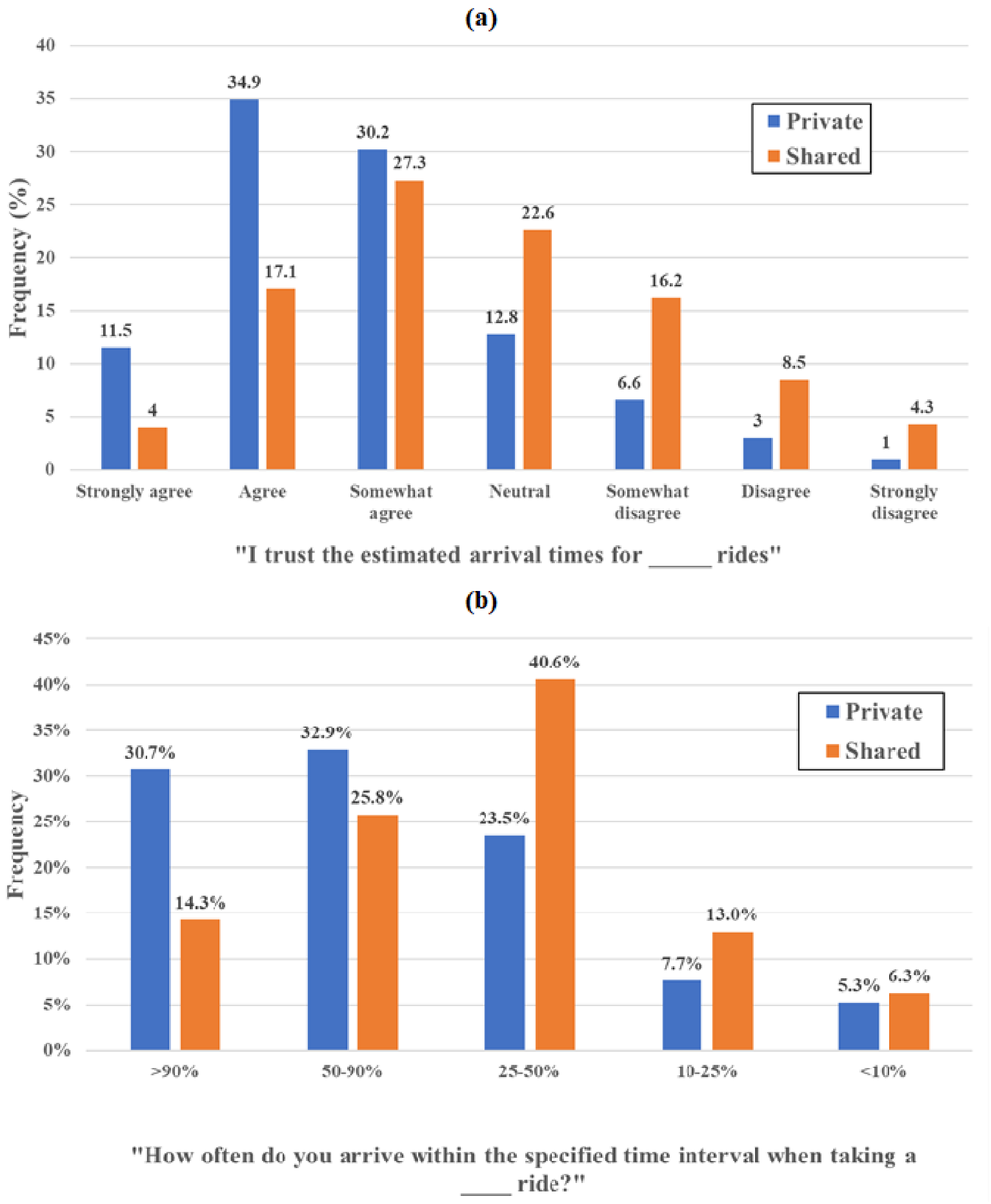


Figure 2-1: Selected statistics on respondents' attitudes towards reliability in pooled and exclusive ridehailing services. (a) Agreement with the statement "I trust the estimated arrival times for private (blue)/shared (orange) rides." (b) Respondents' estimated frequency of on-time arrival in prior ridehailing trips.

exclusive service.

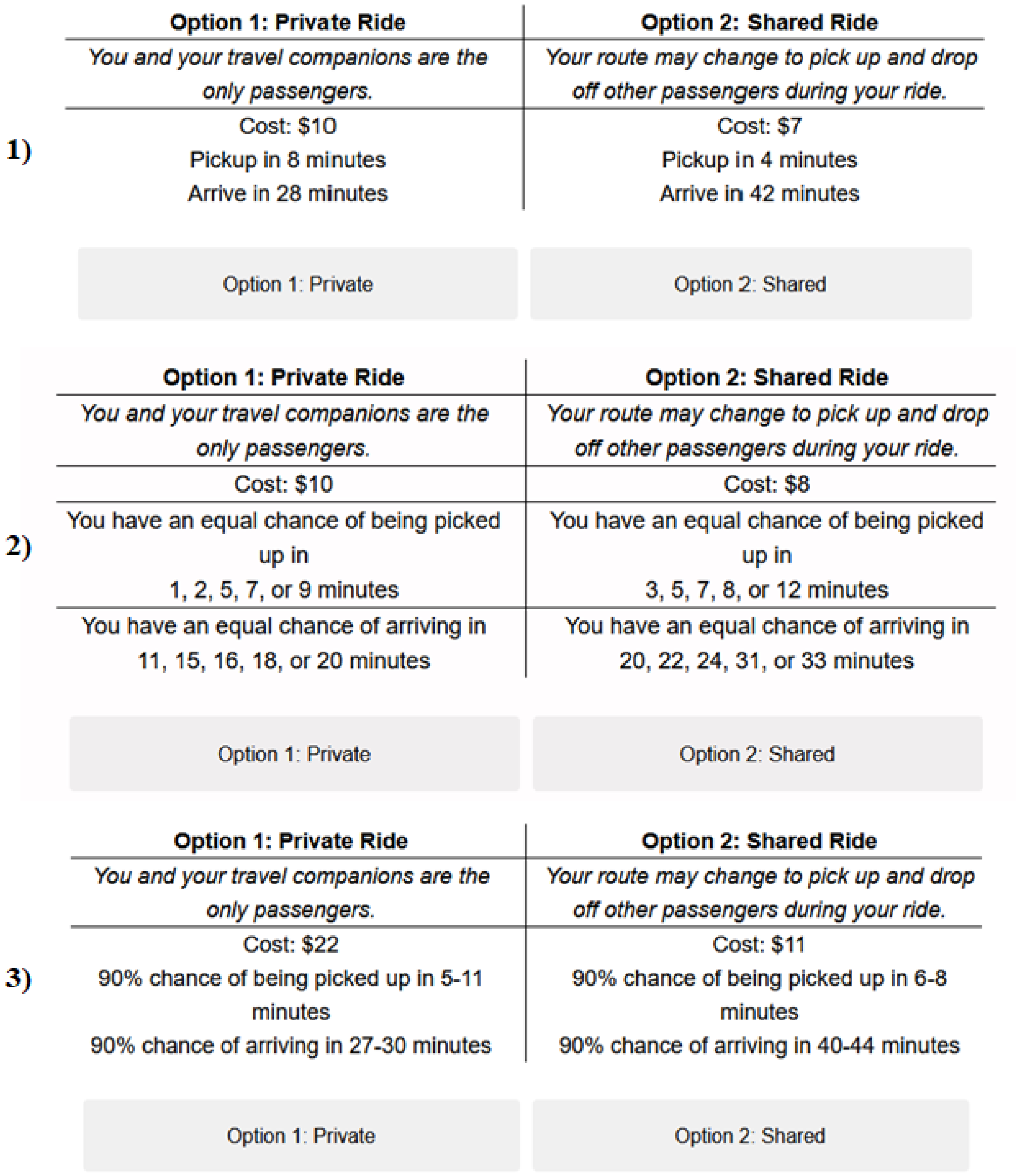
2.3.3 Choice Experiment Design

After collecting information about the reference trip, each respondent was asked to imagine a scenario in which they were conducting the reference trip using a ridehailing service with both exclusive and pooled options. These options were described in the survey as “private” and “shared” for the sake of comprehension among a general audience who may be unfamiliar with more technical terms. Respondents were shown relevant information for each option and asked which they would take for their journey.

The first block shown to each respondent provided information on the cost, estimated waiting time, and estimated total journey time (including wait) for each option (question format 1 in Figure 2-2). Subsequently, they were shown a second block that included information on the variability of the waiting time and of the total journey time as well. One of two possible formats was used for each respondent. With 25% probability, they would see questions where five possible realizations of the waiting and journey time, each occurring with equal probability, were shown (question format 2 in Figure 2-2). Otherwise, they would see a block of questions where the waiting and journey times were presented as time ranges (question format 3 in Figure 2-2). For respondents given the time range questions, one-third were told that the time ranges were correct 75% of the time, one-third were told they were correct 90% of the time, and one-third were not told any explicit information regarding the associated probability.

The three levels of each parameter are displayed in Table 2.7. The cost and journey time parameters for the exclusive trip option were derived from the estimated driving time collected using the Directions API. The same parameters for the pooled mode were subsequently derived from the exclusive mode values for each question (using the mean or midpoint of the value in the second and third question formats, respectively). A partial factorial design was created for these questions such that each respondent was shown six questions, with all three levels of each parameter equally distributed among these six questions.

We choose to include two question formats including time variability information



1)

Option 1: Private Ride	Option 2: Shared Ride
You and your travel companions are the only passengers.	*Your route may change to pick up and drop off other passengers during your ride.*
Cost: $10 Pickup in 8 minutes Arrive in 28 minutes	Cost: $7 Pickup in 4 minutes Arrive in 42 minutes

Option 1: Private

Option 2: Shared

2)

Option 1: Private Ride	Option 2: Shared Ride
You and your travel companions are the only passengers.	*Your route may change to pick up and drop off other passengers during your ride.*
Cost: $10	Cost: $8
You have an equal chance of being picked up in 1, 2, 5, 7, or 9 minutes	You have an equal chance of being picked up in 3, 5, 7, 8, or 12 minutes
You have an equal chance of arriving in 11, 15, 16, 18, or 20 minutes	You have an equal chance of arriving in 20, 22, 24, 31, or 33 minutes

Option 1: Private

Option 2: Shared

3)

Option 1: Private Ride	Option 2: Shared Ride
You and your travel companions are the only passengers.	*Your route may change to pick up and drop off other passengers during your ride.*
Cost: $22 90% chance of being picked up in 5-11 minutes 90% chance of arriving in 27-30 minutes	Cost: $11 90% chance of being picked up in 6-8 minutes 90% chance of arriving in 40-44 minutes

Option 1: Private

Option 2: Shared

Figure 2-2: Example questions of each format used in the stated preference survey. 1) No explicit information on variability, shown to all respondents. 2) Five possible pickup and arrival times, shown to 25% of respondents. 3) Time ranges, shown to 75% of respondents.

in the experimental design because we observe an inherent tradeoff across the formats between realism and mathematical precision. While many studies on the value of travel time variability that use stated preference data employ these types of ques-

Attribute	**Levels**	**Applicable Formats**
Waiting time for vehicle	Exclusive – 3 min, 5 min, 8 min Pooled – 4 min, 7 min, 9 min	1, 2, 3
Total journey time	Exclusive – 80% of estimated reference trip travel time, 100%, 140% Pooled – 125% of estimated reference trip travel time, 150%, 160%	1, 2, 3
Fare	Exclusive – S$2.50 + S$0.45/mi; S$2.50 + S$0.60/mi; S$2.50 + $0.90/mi Pooled – 50% of exclusive fare, 60%, 70%	1, 2, 3
Waiting time variance	1 minute, 2 minute, 3 minutes	2
Journey time variance	Exclusive – 3 min, 4 min, 6 min Pooled – 3 min, 5 min, 8 min	2
Waiting time range width	±1 minute, ±2 minutes, ±3 minutes	3
Journey time range width	Exclusive – ±1 min, ±2 min, ± 5 min Pooled – ±2 min, ±4 min, ±6 min	3

Table 2.7: Attributes and levels for the stated preference choice experiments, and which question formats they applied to. Formats 1, 2, and 3 are the same as in Figure 2-2.

tions [14, 25, 30] and they perform well in comparison to other methods that have been used to convey variability information [28], we note that they do not resemble the information provided to travelers in a real world context. Meanwhile, the time range information displayed within many ridehailing booking interfaces cannot be used to generate moments of the associated distribution of wait times or journey times, meaning that the mean-variance approach cannot be used to estimate the value of travel time variability. In our study, we sought to investigate the ways that the format of the variability information may impact choice behavior. Therefore, we include both the format employed in prior studies and the one more similar to real-world implementations and compare the results of choice models calibrated on the two resulting data sets in our analysis.

A time range without an explicit associated confidence level, such as those present in ridehailing interfaces, provides little information about the underlying variability. In such cases, it is left to the traveler to interpret how likely a trip is to arrive within the estimated time range. Even when a confidence level is provided, the distribution of travel times falling within that range is subject to interpretation. To address the first point, we provide explicit confidence levels for two-thirds of our choice respondents to understand how, if at all, the confidence level displayed to travelers may influence their decisions, in addition to collecting the respondents' own interpretations of the reliability of the estimated time intervals. For the second point, although it is outside the scope of this study to investigate the details of how travelers may implicitly perceive time ranges as probability distributions, the results of the discrete choice models can help us understand this effect at a high level. The size and magnitude of the estimated coefficient of the time range widths gives some insight into how travelers respond to a wider or tighter range of possible times.

Factor Analysis

To investigate any underlying psychological constructs that may influence travelers' perceptions regarding the time uncertainty in exclusive and pooled ridehailing, we use exploratory factor analysis (EFA) to reduce several potential indicators of such attitudes into a smaller number of factors. We include 5 7-point and 12 5-point Likert scale statements in the analysis, with the responses to the 5-point questions adjusted to a 7-point scale so that all question types and the resulting factor loadings use the same scale and are comparable to one another. We find the best fit with five factors corresponding to underlying latent attitudes. The indicators included in the survey and their loadings with these five factors in the EFA model are shown in Table 7. We use a threshold of 0.45 as the threshold for defining strong factor loadings, which is a commonly cited empirical threshold [34].

The first and second factors grouped statements relating to the importance of different time attributes relating to pooled and exclusive ridehailing, respectively, in making decisions between the two services. These statements correspond to the data shown in 2.8. The specific statements used in these factors regarded the estimated pickup and journey time for the exclusive ride option as well as the uncertainty of the journey time and the possibility of arriving late when using the private option (Factor 2); and the same statements for the shared ride option, along with the possibility of being matched with strangers when using the shared option (Factor 1). Note that these factors do not include the related statement for each mode about the importance of cost when using that mode. Using confirmatory factor analysis (CFA), we tested the inclusion of these indicators with the relevant factors and found reduced model fit. Therefore, we interpret these factors as attitudes towards time importance for pooled and exclusive ridehailing. We speculate that the indicators regarding cost do not load to the same factors because they relate to different attitudes regarding price distinct from conceptions of the journey time.

The third factor was associated with four statements regarding the estimated arrival times and lateness of the exclusive mode, though several are comparative between the exclusive and pooled mode. The variety of these indicators makes it difficult to ascribe them to a single underlying attitude or psychological construct. As the fifth factor associates only with agreement of the single statement "I trust the estimated arrival time for shared rides", we choose to identify the corresponding statement "I trust the estimated arrival time for exclusive rides" as the indicator of most interest among the four with significant loadings on the third factor, and use this sole indicator for the remainder of the research while ignoring the estimated factor.

The fourth factor grouped responses to the two questions regarding the estimated frequency with which prior exclusive and pooled ridehailing trips arrived within the time ranges displayed prior to the trip. While the identification of these two attitudes as a single factor seems reasonable as it is likely that respondents may answer these questions similarly due to their similar topic and wording, we choose to consider them separately in the subsequent discrete choice models rather than combine them into a single factor. This choice allows us to investigate how respondents' beliefs in the reliability of exclusive and pooled services separately influence their decision-making

Statement	**Factor Loadings**				
	1	**2**	**3**	**4**	**5**
When booking a private (unshared) ride via smartphone app, roughly how often do you think you arrive by the estimated arrival time, or within the estimated arrival time interval?	-0.02	-0.02	-0.13	**0.76**	0.03
When using ridesharing, roughly how often do you think you arrive by the estimated arrival time, or within the estimated arrival time interval?	-0.04	0.07	0.11	**0.57**	-0.24
When using ridesharing, how often are you paired with at least one other passenger?	-0.06	-0.03	-0.19	0.35	0.04
The estimated arrival times provided are important in deciding whether to book any kind of app-based taxi/minicab ride for my trip	-0.18	0.24	**-0.61**	0.19	-0.03
The estimated arrival times provided are important in deciding between booking a private ride option and a shared ride option	-0.22	0.24	**-0.59**	0.09	-0.14
I trust the estimated arrival times for private rides	-0.03	0.13	**-0.48**	0.14	-0.23
I trust the estimated arrival times for shared rides	-0.09	0.02	-0.15	0.09	**-0.86**
I prefer to book private rides rather than rideshare because I am less likely to arrive late	-0.12	0.16	**-0.54**	-0.06	0.08
When booking a ridehailing trip via a smartphone app, how important are each of the following to you in choosing between a private ride and a shared ride?					
The estimated journey time of the private ride option	-0.2	**0.63**	-0.29	0.05	0.04
(Importance of) The estimated pickup time of the private ride option	-0.25	**0.67**	-0.15	-0.04	-0.06
(Importance of)The estimated journey time of the shared ride option	**-0.51**	0.31	-0.16	0.06	-0.01
(Importance of) The estimated pickup time of the shared ride option	**-0.56**	0.44	-0.02	-0.01	-0.15
(Importance of) The possibility of arriving late with the private ride option	-0.25	**0.55**	-0.27	0.06	-0.01
(Importance of) The uncertainty of the journey time with the private ride option	-0.34	**0.56**	-0.21	0	-0.04
(Importance of) The possibility of arriving late with the shared ride option	**-0.73**	0.16	-0.23	0.08	0.04
(Importance of) The uncertainty of the journey time with the shared ride option	**-0.71**	0.17	-0.22	0.06	0.03
(Importance of) The possibility of being matched with strangers	**-0.51**	0.2	0	-0.07	-0.12

Table 2.8: Results of the exploratory factor analysis (EFA) on latent factors relating to attitudes regarding time uncertainty in ridehailing. Strong factor loadings, defined to be greater in magnitude than 0.45, are listed in bold.

processes.

After making the decision to proceed with only two factors (Factor 1 and Factor 2) that group items relating to the importance of displayed time attributes for each form of ridehailing when deciding between booking an exclusive and pooled trip, we then define these factors in terms of the corresponding statements described previously. We label these two factors as time importance in exclusive or pooled ridehailing. We then confirm that these factors accurately describe the collected data by running a CFA model to evaluate the model fit and demonstrate convergent validity of the identified factors. The CFA model fit (CFI = 0.951, TLI = 0.923, RMSEA = 0.083, SRMR = 0.043) meets established benchmarks [34], indicating that the items identified for each individual construct are highly related to one another.

Discrete Choice Modeling

Having identified latent attitudes that relate to specific statements, we then estimated a binomial logit (BNL) model for the trip-level choices between exclusive and pooled ridehailing for different question types. We model the utility for each mode as a function of the trip-activity context (such as trip purpose, urgency of arriving on time, and number of travel companions), the service attributes of that mode, individual sociodemographics, and latent attitudes. In addition, we incorporate inertia from prior travel behavior that may result from an individual's actual travel mode used for the reference trip. We define inertia variables for the shared and exclusive ridehailing modes that equal 1 if that mode was the respondent's reported primary mode for their reference trip, an approach common in existing literature [35]. We additionally hypothesize that respondents who took public transit for their reference trip may prefer to take pooled ridehailing due to its similarity as a less expensive travel mode shared with strangers, and define an inertia variable that equals 1 if public transit was the primary mode for the reference trip.

Formally, we model the expression for the utility of an individual respondent given a specific mode and choice situation is as follows:

$$V_{mnc} = \beta_m^0 + \beta_m^S S_{mnc} + \beta_m^X X_{mnc} + \beta_m^L L_n + \beta_m^I I_{mn} + \epsilon_{mnc} \tag{2.1}$$

Where V_{mnc} is the utility of mode m for individual n and choice context c, S_{mnc} is a vector of the service attributes of mode m, X_{mnc} is a vector of sociodemographic characteristics and other attributes describing the choice context, L_n is a vector containing the factor scores of individual n for factors 1 and 2 found in the prior section (computed sequentially rather than simultaneously due to computational limitations), and I_{mn} is a vector of the inertia variables for mode m and individual n. β_m^0 is the alternative specific constant, and β_m^S, β_m^X, β_m^L, and β_m^I are vectors of mode-specific coefficients; these are unknown and to be estimated by the model. ϵ_{mnc} is a random disturbance term relating to the mode, individual, and choice context. In order to estimate a panel effect relating different responses by the same respondent, we specify this term as follows:

$$\epsilon_{mnc} = \eta_{mn} + \xi_{mnc} \tag{2.2}$$

Where η_{mn} is a random Gaussian variable with zero mean and standard deviation σ_{η_m} that captures the variation of preferences across individuals. It varies across individuals but is constant across different choice contexts for each individual, more accurately capturing the connections between respondents' choices in multiple contexts. ξ_{mnc} is a Gumbel distributed disturbance term.

We additionally model the alternative specific constant of the pooled mode and the coefficient of travel time as normally distributed random variables – the model estimates the mean and standard deviation of the distributions. While simultaneous estimation of the respondents' factor scores, L_n, as part of a single solution procedure to this discrete choice problem is theoretically possible, due to computational limitations we compute them sequentially. Figure 2-3 shows a graphical representation of our framework, including the relationships between the covariates and the utility of each travel mode. We implement three choice models, one for each different question format implemented in the choice experiment:

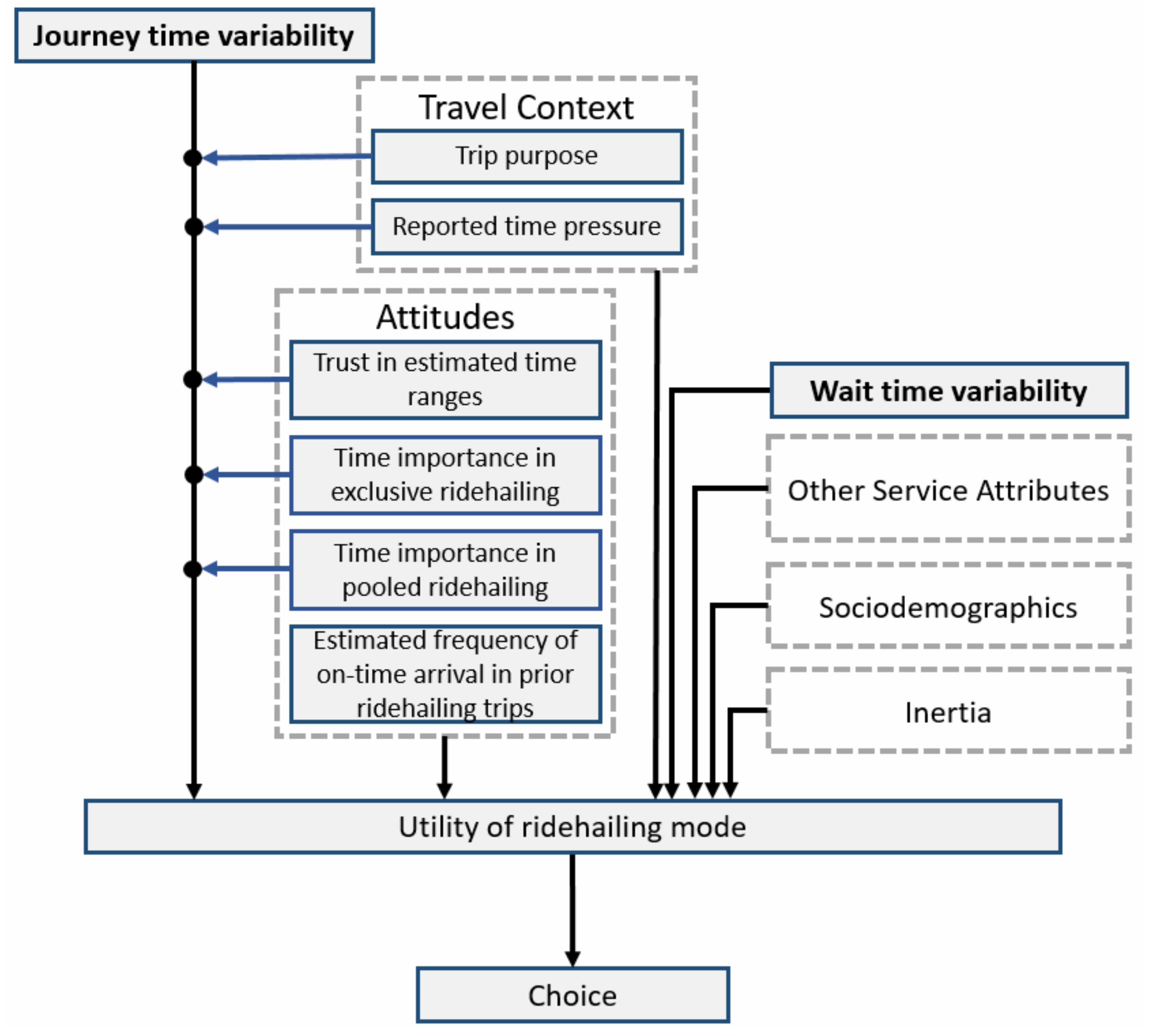

Journey time variability
Travel Context
Trip purpose
Reported time pressure
Attitudes
Trust in estimated time ranges
Time importance in exclusive ridehailing
Time importance in pooled ridehailing
Estimated frequency of on-time arrival in prior ridehailing trips
Wait time variability
Other Service Attributes
Sociodemographics
Inertia
Utility of ridehailing mode
Choice

Figure 2-3: Framework of the binomial logit choice model. The blue arrows indicate moderating effects that were tested through the introduction of interaction terms.

- A base model, using only the first question block, which shows no explicit information on the variability of waiting or journey time.

- A mean-variance model, using the questions showing five possible realizations and using the mean and variance of the displayed wait and journey times as the parameters relating to variability.

- A model using the time range questions, using the midpoint and width of the time range as the parameters relating to variability.

To investigate the influence of travel context and attitudes on the perception and valuation of uncertainty, we included interaction terms between each travel context and attitude variable and the wait and journey time variability variables in the second and third models tested. These interaction terms represent moderating effects of these factors on journey time variability. However, very few were significant in any of the models tested – for the sake of easier comprehension, we present in this paper only the interaction terms found to be significant in any model. This has slight impacts on the significance of some other variables in the model, but the main conclusions we draw are unaffected.

2.4 Choice Experiment Results

2.4.1 Base Model without Uncertainty

The estimation results for the base model are shown in Table 2.9. The model is calibrated on the responses of 951 individuals, with 5,706 observations. The log likelihood of the estimated model is -3,032, with $\bar{\rho}^2 = 0.225$.

Because there is no explicit information regarding the variance of wait time or total journey time in the questions used for this model, we can see that the estimated total journey time is equal to the wait time plus the in-vehicle travel time. Because the coefficient for waiting time for the exclusive mode is not significantly different from zero, the value of waiting time and the value of in-vehicle travel time for this mode are similar. However, because the corresponding coefficient for the pooled mode is significantly negative, the value of waiting time for the pooled mode is significantly greater in magnitude than the value of in-vehicle travel time for either mode. This finding could be due to perceived unreliability in waiting times for the pooled mode, as vehicles may be detoured en route as new requests are assigned to the vehicle. An increase in estimated waiting time shown may lead respondents to worry that there is more time for other trips to be assigned to their vehicle, further delaying their pickup.

All sociodemographic characteristics included in the model have significant impacts on the willingness to use pooled ridehailing. Respondents over the age of 35

Attribute		Estimate		Std. Error	t-stat
Alternative specific constant, mean^{P}		0.336		0.236	1.42
Alternative specific constant, std. dev.^{P}		1.82	***	0.103	17.7
Service Attributes	Cost^{E,P}	-0.311	***	0.016	-19.1
	Estimated wait time^{E}	-0.0201		0.018	-1.11
	Estimated wait time^{P}	-0.127	***	0.018	-6.94
	Estimated journey time, mean^{E,P}	-0.156	***	0.011	-14.9
	Estimated journey time, std. dev.^{E,P}	0.0897	***	0.013	6.72
Socio-demographics	Age ($\geq$35 years old)^{P}	-0.197		0.157	-1.25
	Education – Bachelor's degree^{P}	0.173		0.182	0.95
	Education – Advanced degree^{P}	-0.762	***	0.238	-3.2
	Gender – Female^{P}	0.292	*	0.159	0.184
	Income – 1st quintile (<\$2,500)^{P}	0.545	**	0.247	2.21
	Income – 2nd quintile (\$2,500-5,999)^{P}	-0.375	*	0.206	-1.82
	Income – 4th quintile (\$10,000-14,999)^{P}	0.408	***	0.15	2.72
	Income – 5th quintile ($\geq$\$14,999)^{P}	-0.072		0.182	-0.40
Attitudes	Trust in arrival time estimates for exclusive mode^{E}	0.201	***	0.068	2.96
	Trust in arrival time estimates for pooled mode^{P}	0.179	***	0.055	3.22
	Time importance in exclusive mode^{E}	0.225		0.184	1.23
	Time importance in pooled mode^{P}	0.153		0.187	0.81
	Estimated frequency of on-time arrival in prior exclusive trips^{E}	0.021		0.057	0.36
	Estimated frequency of on-time arrival in prior pooled trips^{P}	0.163	***	0.060	2.71
Trip Context	Not at all important to arrive on time^{P}	-0.131		0.269	-0.49
	Extremely important to arrive on time^{P}	-0.097		0.203	-0.48
	Home to work/school^{P}	0.096		0.31	0.31
	Work/school to home^{P}	-0.36		0.236	-1.53
	Social/recreation activity to home^{P}	-0.213		0.212	-1.0
Interaction Terms	Not at all important to arrive on time × home to work/school^{P}	-1.4	**	0.703	-2.0
	Extremely important to arrive on time × home to work/school^{P}	-0.785	**	0.388	-2.02
	Not at all important to arrive on time × work/school to home^{P}	0.585		0.516	1.13
Inertia	Exclusive ridehailing^{E}	0.143		0.209	0.684
	Pooled ridehailing^{P}	1.41	***	0.269	5.24
	Public transit^{P}	0.488	**	0.201	2.43

Table 2.9: Estimation results of discrete choice model for questions with no explicit information on time uncertainty. Attributes marked with ^{E} are coefficients for the exclusive ridehailing mode, and those marked with ^{P} are coefficients for the pooled ridehailing mode. Significance: * = 0.1, ** = 0.05, *** = 0.01.

and those holding graduate degrees are significantly less likely to choose pooling, while women and respondents in the 1st and 4th income quintiles are more likely. Furthermore, many of the attitudes we measure in this study significantly influence traveler choice between exclusive and pooled ridehailing, namely trust in the provided arrival time estimates for both modes, as well as respondents' estimated frequency of arriving on-time in previous pooled ridehailing trips. An increase in an individual's reported trust in the time estimates for a mode yields a significant increase in their likelihood of selecting that mode in the choice experiment, as does an increase in their experiential estimate of arriving on-time when using pooled services. Because of the gaps in reported trust and in beliefs regarding on-time arrival found in our sample between exclusive and pooled ridehailing, this finding indicates that travelers' lack of trust in the on-time performance of pooled ridehailing may be a significant contributor to its low ridership.

Additionally, the inertia of the reference trip is significant in predicting respondents' choices in our experiment in some cases. Respondents who reported taking a pooled ridehailing service as the primary mode of their reference trip and, to a lesser extent, those who reported taking public transit were more likely to select the pooled ridehailing option. However, those who reported taking an exclusive ridehailing service as the primary mode were not significantly more likely to choose the exclusive option. This would indicate that although these respondents were willing to take ridehailing and also often willing to choose a comparable pooled option in our study, they took exclusive trips in real life rather than pooled ones. This could occur if the pooled option was not available in real life, forcing the use of an exclusive service, a likely possibility given the recent shutdown of pooled services in Singapore at the time of this survey.

2.4.2 Mean-Variance Model

We next investigate how the traditional approach in stated preference experiments, using five possible displayed wait or journey times, influences respondents' choices. We use the mean-variance model of value of time variability, taking the mean and standard deviation of the five displayed values as the covariates for the choice model. The estimation results of the choice model are displayed in Table 2.10. This model uses the data from 266 responses, with 1,596 total observations. The resulting log likelihood of the model is -823 and $\bar{\rho}^2 = 0.208$.

Comparing to the base model, many attributes are statistically significant in both models, including all service attributes, the 1st income quintile, trust in pooled and exclusive arrival time estimates, and inertia from pooled ridehailing or public transit mode use in the reference trip. However, we also find some attributes that are present in both models, but significant in only one. The coefficient for the mean wait time for exclusive ridehailing is significantly negative only in this model.

Several sociodemographic attributes – education, gender, age, and the 4th income quintile – were significant only in the base model. The 5th income quintile and the inertia of exclusive ridehailing mode use in the reference trip are both significant and negative in this model, while not significant in the base model. While the latent attitude of time importance in the pooled mode is significant and positive in this model, the estimated frequency of on-time arrival in previously taken pooled trips is not; the opposite case is true in the base model.

The coefficients for the variance of the displayed times indicate respondents' value of travel time variability for each mode. The coefficients of the standard deviation of wait time are not significantly different than zero for either the exclusive or the pooled mode, suggesting information on wait time variability does not affect traveler decision-making behavior. However, the coefficients for the standard deviation of the displayed journey times are negative, indicating that greater variability of journey time does significantly reduce willingness to use either exclusive or pooled service, though it is statistically significant only for exclusive service. For exclusive service, an increase in standard deviation of total journey time by 1 minute is equivalent to a \$0.45 increase in fare, or to 0.88 minutes of mean journey time on average. The value of travel time variability appears to be greater when using exclusive ridehailing than pooled ridehailing, which could be because the information displayed is not trusted as much for pooled rides and thus is less impactful in decision-making.

Attribute		Estimate		Std. Error	t-stat
Alternative specific constant, mean^{P}		-1.12	*	0.586	-1.91
Alternative specific constant, std. dev.^{P}		1.67	***	0.244	6.86
Service Attributes	Cost^{E,P}	-0.354	***	0.036	-9.92
	Mean of estimated wait times^{E}	-0.102	***	0.035	-2.92
	Mean of estimated wait times^{P}	-0.0823	**	0.035	-2.34
	Std. dev. of estimated wait times^{E}	-0.0512		0.091	-0.56
	Std. dev. of estimated wait times^{P}	-0.0669		0.092	-0.73
	Mean of estimated journey times, mean^{E,P}	-0.176	***	0.023	-7.69
	Mean of estimated journey times, std. dev.^{E,P}	0.138	***	0.028	4.99
	Std. dev. of estimated journey times^{E}	-0.153	***	0.058	-2.63
	Std. dev. of estimated journey times^{P}	-0.096		0.060	-1.59
Socio-demographics	Age (≥35 years old)^{P}	0.147		0.333	0.44
	Education – Bachelor's degree^{P}	0.003		0.363	0.01
	Education – Advanced degree^{P}	0.365		0.506	0.72
	Gender – Female^{P}	0.21		0.313	0.67
	Income – 1st quintile (<$2,500)^{P}	1.09	**	0.486	2.25
	Income – 2nd quintile ($2,500-5,999)^{P}	0.02		0.392	0.05
	Income – 4th quintile ($10,000-14,999)^{P}	-0.171		0.356	-0.48
	Income – 5th quintile (≥$14,999)^{P}	-0.947	**	0.432	-2.19
Attitudes	Trust in arrival time estimates for exclusive mode^{E}	0.308	**	0.138	2.24
	Trust in arrival time estimates for pooled mode^{P}	0.323	***	0.117	2.75
	Time importance in exclusive mode^{E}	0.266		0.364	0.73
	Time importance in pooled mode^{P}	0.698	*	0.369	1.89
	Estimated frequency of on-time arrival in prior exclusive trips^{E}	-0.107		0.118	-0.91
	Estimated frequency of on-time arrival in prior pooled trips^{P}	-0.117		0.119	-0.98
Trip Context	Not at all important to arrive on time^{P}	0.831		0.569	-1.46
	Extremely important to arrive on time^{P}	-0.482		0.394	-1.22
	Home to work/school^{P}	-0.069		0.585	-0.12
	Work/school to home^{P}	-0.497		0.463	1.07
	Social/recreation activity to home^{P}	0.323		0.422	0.77
Interaction Terms	Time importance in pooled mode × std. dev. of journey times^{P}	-0.049		0.095	-0.51
	Home to work/school × std. dev. of journey times^{E}	-0.248	*	0.144	-1.72
	Not at all important to arrive on time × home to work/school^{P}	3.35	**	1.45	2.32
	Extremely important to arrive on time × home to work/school^{P}	-0.059		0.782	-0.08
	Not at all important to arrive on time × work/school to home^{P}	2.93	**	1.08	2.73
Inertia	Exclusive ridehailing^{E}	-0.879	*	0.444	-1.98
	Pooled ridehailing^{P}	1.33	**	0.554	2.40
	Public transit^{P}	1.44	***	0.421	3.42

Table 2.10: Estimation results of discrete choice model with questions showing 5 possibilities of wait and journey times. ^{E}, ^{P}, *, **, and *** are defined as in Table 2.9.

We include interaction terms to investigate the presence of any moderating effects that travel context or individual attitudes may have on the value of time variability, as well as particular combinations of trip contexts. Home-to-work and home-to-school commute trips result in significantly greater value of travel time variability for exclusive trips: travelers engaging in such a trip have an additional disutility from 1 additional minute of standard deviation of journey time equivalent to \$1.40, or 2.8 minutes of mean travel time on average. However, pooled trips did not see a corresponding increase in value of travel time variability. We also find that none of the attitudes measured in our study had a significant moderating effect on the variability of travel times. For trip context combinations, we find that commute trips where the respondent reported no time pressure were significantly more likely to be taken using pooled ridehailing.

2.4.3 Time Range Model

Finally, we investigate the set of choice experiment questions with waiting time and total journey time displayed as time ranges. Because a time range has no specific underlying probability distribution and we did not want to impose any suppositions on how respondents may perceive such a distribution if at all, we use only the midpoint, representing a center of probability mass, and the width, representing the variability, of the time range. We fit this model on the data collected from 685 respondents, with 4,110 total observations. The final log likelihood of the model is -2,131, and $\bar{\rho}^2 = 0.233$. The results of the estimation model are shown in Table 2.11.

Again, the coefficients of some attributes differ in significance level from previous models that included them. In the time range model, the significance of all service attributes not related to time variability are consistent with the five possibilities model. The sociodemographic attributes are split, with the 1st income quintile significant across all three models, gender and education consistent between the time range model and the base model, and age and 4th income quintile consistent with the five possibilities model. The attitudinal attributes are consistent with the base model as well, with experienced on-time arrival performance of pooled ridehailing

	Attribute	Estimate		Std. Error	t-stat
	Alternative specific constant, mean^{P}	-0.291		0.318	-0.92
	Alternative specific constant, std. dev.^{P}	1.94	***	0.125	15.6
Service Attributes	Cost^{E,P}	-0.376	***	0.021	-17.6
	Midpoint of wait time range^{E}	-0.098	***	0.022	-4.53
	Midpoint of wait time range^{P}	-0.088	***	0.022	-4.03
	Width of wait time range^{E}	-0.012		0.027	-0.46
	Width of wait time range^{P}	-0.025		0.027	-0.93
	Midpoint of journey time range, mean^{E,P}	-0.162	***	0.013	-12.8
	Midpoint of journey time range, std. dev.^{E,P}	-0.080	***	0.02	-3.99
	Width of journey time range^{E}	0.052	***	0.013	3.96
	Width of journey time range^{P}	0.018		0.014	1.27
Socio-demographics	Age (≥35 years old)^{P}	0.025		0.193	1.06
	Education – Bachelor's degree^{P}	0.204		0.223	0.91
	Education – Advanced degree^{P}	-0.644	**	0.288	-2.24
	Gender – Female^{P}	0.452	**	0.199	2.27
	Income – 1st quintile (<$2,500)^{P}	0.622	**	0.305	2.04
	Income – 2nd quintile ($2,500-5,999)^{P}	-0.191		0.251	-0.76
	Income – 4th quintile ($10,000-14,999)^{P}	0.028		0.192	0.15
	Income – 5th quintile (≥$14,999)^{P}	-0.319		0.237	-1.35
Attitudes	Trust in arrival time estimates for exclusive mode^{E}	0.218	**	0.086	2.54
	Trust in arrival time estimates for pooled mode^{P}	0.17	**	0.067	2.53
	Time importance in exclusive mode^{E}	0.012		0.231	0.05
	Time importance in pooled mode^{P}	0.007		0.235	0.03
	Estimated frequency of on-time arrival in prior exclusive trips^{E}	0.017		0.071	0.24
	Estimated frequency of on-time arrival in prior pooled trips^{P}	0.266	***	0.074	3.61
Trip Context	Not at all important to arrive on time^{P}	-0.098		0.284	-0.35
	Extremely important to arrive on time^{P}	-0.54	**	0.215	-2.52
	Home to work/school^{P}	-0.362		0.277	-1.31
	Work/school to home^{P}	-0.003		0.268	-0.01
	Social/recreation activity to home^{P}	-0.022		0.266	-0.08
Interaction Terms	Time importance in pooled mode × width of journey time range^{P}	0.049	**	0.022	2.23
	Home to work/school × width of journey time range^{E}	-0.037		0.031	-1.21
	Not at all important to arrive on time × home to work/school^{P}	0.43		0.311	1.38
	Extremely important to arrive on time × home to work/school^{P}	0.423		0.457	0.93
	Not at all important to arrive on time × work/school to home^{P}	0.43		0.311	1.38
Inertia	Exclusive ridehailing^{E}	0.289		0.256	1.13
	Pooled ridehailing^{P}	1.38	***	0.328	4.20
	Public transit^{P}	0.348		0.25	1.39

Table 2.11: Estimation results of discrete choice model with questions showing time ranges for wait and journey times. ^{E}, ^{P}, *, **, and *** are defined as in Table 2.9.

significant but not either of the time importance attitudes. The coefficient for exclusive ridehailing inertia is significant in this model as well, but it surprisingly has a negative coefficient, meaning it leads to greater likelihood to choose pooling in the choice experiment.

We investigate whether time variability represented by time ranges, as is frequently the case in real-world ridehailing booking interfaces, impacts respondents' likelihood to use either mode. For this question format, we find that the variability of the wait times does not significantly impact decision-making, a similar finding as from the five possibilities model. However, the variability of total journey time has a very different impact in this model than in the previous one. The coefficient for the width of the journey time range for the pooled mode is not significantly different than zero, indicating that respondents were no more or less likely to choose a trip with a wider time range than one with a narrower range, all else being equal. However, the corresponding coefficient for the exclusive mode is positive, indicating that respondents actually preferred wider time ranges to narrower ones for exclusive trips.

This result, though surprising, could be explained by a risk-seeking or optimistic bias of the respondents. This optimistic bias may lead respondents to weigh the lower end of the range more heavily, either in perceived likelihood, importance, or both. Another contributing factor could be attribute non-attendance, where respondents ignore the upper bound of the time range and consider only the lower bound in their response. We also note that the coefficient of journey time range width is much smaller in magnitude than that of the midpoint of the journey time. If we take a time range width midpoint m, lower bound l, and width $2 * (m - l)$, comparing it to a time range with the same lower bound and midpoint $m + 1$ results in a utility loss of:

$$\begin{aligned} \Delta V &= -0.168 * (m + 1 - m) + 0.0517 * (2(m - l) - 2(m + 1 - l)) \\ &= -0.168 + 2 * 0.0517 \\ &= -0.0646 \end{aligned} \tag{2.3}$$

Therefore, despite the positive utility from time range width, our model suggests a rational response to a theoretically dominated trip with equivalent minimum journey time and worse maximum journey time. All else equal, increasing the time range width by 1 minute on each side (for a total increase of 2 minutes of time range width) is equivalent to a fare reduction of $0.26, or a savings of 0.62 minutes of total journey time.

We again include interaction terms to identify any potential moderating effects on the value of travel time variability. The only significant finding is that the latent attitude we identified regarding the importance of time attributes in pooled ridehailing corresponds with a significant increase in the coefficient for the journey time range width. Although they do not specifically moderate the influence of journey time variability, trips made from home to work/school and those made under high degrees of time pressure are less likely to be taken using pooled ridehailing.

2.5 Conclusions

In this study, we used the results of a stated preference choice experiment conducted by an online survey of Singapore residents to model travelers' decisions within ridehailing services to take exclusive or pooled trips. Specifically, we sought to understand how information and attitudes regarding the time uncertainty of these services impacted these decisions.

Using multiple question formats with different presentations of variability information, we found that the measured value of travel time variability was highly dependent on the presentation format. While a traditional approach found a negative value of travel time variability, as observed in prior studies [14], a realistic approach using time ranges found the opposite response to variability. Travelers were significantly more likely to choose exclusive ridehailing as the width of the journey time range increased, while for pooled service the relationship was in the same direction though not statistically significant.

Although this method is uncommon in existing literature, a study which used a similar format found a decrease in utility for wider time ranges for pooled ridehailing service in the United States [13]. The disparity between our findings and those of the extisting study could be the result of differences in either survey frames or question formats. Our study presented time ranges for wait and total journey time in both the exclusive and pooled options, while the previous study used them only for in-vehicle travel time for the pooled option.

Besides investigating the impact of information, we included several attitudes towards time variability in our models to understand the impacts of these on willingness to use pooled ridehailing and sensitivity to travel time variability in both ridehailing modes. Travelers' decisions to use pooling in on-demand ridehailing services appear to be influenced greatly by their experiences with reliability in these modes. Their previous experiences riding these services influence their trust in the information provided to them about these services, as well as their beliefs regarding the frequency of on-time arrivals, and greater trust or perceptions of reliability for pooling lead to greater willingness to use this service in our choice experiments. However, our survey results also indicate that for many riders, these previous experiences are generally poor – 60% of travelers report beliefs that fewer than half of their pooled trips arrive within the time interval specified at the beginning of the journey. Practically, this indicates that the usage of pooled services is hampered by their poor reliability, or at least perceptions of poor reliability.

Furthermore, while the information provided in our experiments regarding the variability of arrival time for the exclusive mode was influential in that mode, we found that the same information for pooled trips had negligible impact on the decision to use pooled or exclusive. In combination with our other findings, this could suggest that travelers' negative attitudes and perceptions of pooled ridehailing shape how they process information provided by the service itself.

Overall, these findings highlight that passengers' lack of trust in the reliability of DRS services is a major reason why ridehailing users choose to take exclusive trips instead of pooled ones. In order to shift riders from exclusive to pooled trips, a key step towards addressing the sustainability issues present in ridehailing as it currently stands, recognizing and addressing this barrier should be a priority for operators of DRS services. Despite identifying the presence and importance of these attitudes, the findings of this research alone do not indicate the best approach for DRS operators or policymakers to achieve these goals. However, we note several important theoretical considerations extending upon our research that are relevant to its practical applications.

Both the accuracy and precision of the time estimates made by the service operator are influential in travelers' willingness to use the mode. Here, we use the term "precision" to refer to the level of uncertainty captured within the estimate. A narrow time range may be precise, but more likely to be inaccurate and lead to a true arrival time that falls outside of the estimate.

Decreasing the precision of the time estimates communicated to travelers is likely to make these estimates more accurate, and should theoretically increase the trust that the traveler places in the service meeting this promise. In practice, however, this will likely discourage use of the service by highlighting its unreliability. Meanwhile, providing time estimates that are overly precise risks more frequently failing to deliver on these promises. In this case, though the information on the trip's travel time variability may seem appealing, travelers are unlikely to trust it - this dynamic appears similar to the results we observe in this research.

The trust gap we find between exclusive and pooled modes may reflect a greater sense of worry among riders about the potential to arrive late when using DRS services. These worries may arise as much or more from intuitions and emotional responses to the mode as they do from quantitative understanding of the underlying probabilities, which is likely to be rare among the general population of travelers using these services. While research focus is often placed on quantifiable measures of this variability, an understanding of the psychological processes underpinning travelers' intuitions of uncertainty and risk are likely to be much more helpful in addressing the attitudes that we demonstrate play a large role in at least this one aspect of mode choice.

Finding ways to increase travelers' perceptions of the empirical accuracy of the quantitative probabilistic forecasts that DRS operators make when estimating the travel time ranges for potential riders would certainly help to increase ridership. However, given the challenges involved with demonstrating and communicating this accuracy to travelers, it is also important to pursue methods of building trust that address the psychological roots of this distrust. For instance, communicating to travelers that meeting the time estimates shown when booking trips is a priority to the DRS operator, and committing to that priority publicly, may also be able to make an impact on travelers' trust without having to demonstrate through data that performance is actually improved.

Limitations

We acknowledge several limitations of the research presented in this chapter. The survey was conducted during the national lockdown of Singapore in response to the COVID-19 pandemic in the spring of 2020. Because of these unprecedented circumstances, respondents' real-world travel behavior and/or their stated preferences in the choice experiment sections may reflect idiosyncratic responses to the pandemic, hindering the generalizability of our conclusions to a broader transportation context. Limiting our survey frame to residents of Singapore also hinders generalizability due to Singapore's many uncommon features in a transportation context, including high costs of car ownership and use and a large and well-used transit network.

Although we include several attitudes regarding time uncertainty due to its particular focus in our study, we acknowledge that several other attitudes are used in the literature to model willingness to use pooled ridehailing that were not included in our survey, such as pro-car and environmental attitudes [36]. Additionally, we focus only on the decision between pooled and exclusive ridehailing, as someone who has already decided to use ridehailing and opened a ridehailing app would do. However, the inclusion of other mode alternatives could affect the results that we find, as some research indicates that pooled ridehailing trips are often substitutes for public transit, walking, or biking trips rather than for exclusive ridehailing trips [21].

Future Research

By exploring how travel time variability impacts passenger decisions between using exclusive and pooled ridehailing, as well as capturing some of the attitudes that travelers hold regarding travel time uncertainty and ridehailing, this research opens up many interesting areas of future research. Further exploration of these topics will be helpful in exploring the variety of ways in which travel time uncertainty affects the demand for DRS.

We identify attitudes towards travel time uncertainty and trust in information provided by ridehailing platforms that play a significant role in travelers' willingness to pool rides. In order to understand these dynamics further, future research should be undertaken to develop and validate robust psychological scales to measure these and other relevant attitudes. Research on the formation of these attitudes in response to trip experiences, and how these experiences relate to information shown prior to riding, would also be useful. This research could shed light on the most effective measures to increase consumer confidence in pooled ridehailing over the long term and to display real-time information in a way that is clear and accurate.

While the respondents to our survey report very poor on-time performance for prior pooled ridehailing trips, data regarding arrival time estimates provided by these services as well as the resulting on-time performance is extremely rare. Experiments that collect data on these estimates and the reliability with which they are met would be helpful to understand whether there is a mismatch between reality and the perceptions reported by our survey respondents. Policymakers may also be able to support in this by requiring that this data is collected and released publicly by operators, and operators can support it by sharing this data with researchers without the requirement of any public sector regulations.

In much of the research literature which models the potential impacts of large-scale pooled ridehailing services on transportation systems, travel time variability is generally considered only insofar as an assumption that any individual rider's expected arrival time at their destination under a given assignment is not allowed to deviate from its earliest possible expected arrival time by more than a specified amount. This neglects the impacts of travel time uncertainty, the resulting variability in ridehailing system performance, and the related long-term influence on traveler attitudes, perception, and behavior. When these factors are considered, the system-wide benefits and the costs on individual users and the system operator are likely to change significantly.

To address these issues, future research is needed that models the performance of these pooled ridehailing systems when subjected to travel time uncertainty. Additionally, research in this area that models these systems should report information when available on aspects such as on-time performance, variability of arrival time, and other such factors related to passengers' trust in the system. Future research should also directly address this uncertainty by incorporating more accurate models of travel time variability and create decision-making processes and algorithms which explicitly consider the costs of late arrivals and high variability in arrival time on passengers' willingness to use or return to the system. In chapters 3 and 4 of this dissertation, we begin to address these areas of research by formulating a stochastic optimization problem for dynamic ridesharing under uncertainty and simulating it under a variety of realistic stochastic conditions.

Some of our results are surprising or counterintuitive, and difficult to develop a robust explanation using only the data available to us. Among these are the presence of an optimistic bias present in travelers' valuation of travel time variability when using the time range method of presenting this information. Optimistic bias regarding travel time is comparatively rare in the literature. Further research could explore the degree to which this is caused by respondents' anchoring onto the lower value at the start of the travel time range by testing different presentations of time ranges (using a graphic rather than writing it out, for instance).

Additionally, the contrasting findings between the two presentations of time variability information warrant further study. Our findings call into question whether the traditional approach to including time variability information in stated preference questionnaires is suitable for answering research questions relating to traveler response to real-time information. This is especially relevant for the case of ridehailing as in this paper, where real-time information is displayed to every rider before they book their trip, but can also be applied to other modes as new services, platforms, and technologies make real-time information a more salient aspect of the travel decision-making process.

Finally, despite respondents' trust in the arrival time estimates of both exclusive and pooled ridehailing having significant impact on their decision to choose each of those respective modes, we find that variability information shown to respondents still has a significant impact on willingness to choose to use the exclusive trip, but not for the pooled trip. Because of the trust gap between the two modes, we conjecture that this is the result of a moderation effect, where individuals' degree of trust in the system moderates the influence of the variability of time estimates on their choice. Future research using structural equation modeling to explore these types of more complicated relationships between the attitudes and choices made would be useful to explore these more complex relationships.

Chapter 3

Optimization

3.1 Introduction

In the context of ridehailing systems, dynamic online routing is of key importance. This is because, with limited exceptions, passenger trip requests are not known in advance. Rather, travelers using these services give their trip information to the operator roughly around the time that they wish to undertake this trip, and the operator will be penalized if a vehicle assignment cannot be arranged, or if a vehicle cannot arrive at the passenger's destination, in a short amount of time. At the same time, our findings regarding travelers' willingness to use pooled rides demonstrate that ridesharing is likely to lose passengers over the long term if pickup and drop-off times are highly variable or if the travelers do not trust the ability of the system to meet their estimated travel times.

In the online vehicle-passenger assignment problem, an operator with a set of vehicles distributed throughout a road network faces an incoming set of trip requests, each with pickup and drop-off locations, and aims to generate a matching of requests to vehicles, along with a route for each vehicle to take through the road network to each request's pickup and dropoff locations. The trip requests are often associated with a time constraint or time window, and arriving outside of that window is either penalized or treated as infeasible. This problem is related to the Vehicle Routing Problem with Time Windows (VRP-TW) under stochastic travel times, but differs in two key aspects. First, the set of locations must be visited in a specific order – a passenger's pickup must be visited before their drop-off, and both have to be visited by the same vehicle. Also, this problem is often incorporated into an online system where new requests arrive into the system and must be assigned to vehicles iteratively, as opposed to in a VRP where the set of locations that need to be visited are known in advance.

In practical applications, this problem is made more difficult by a variety of sources of uncertainty that arise when planning passenger-vehicle assignments in ridesharing systems. The first is travel time variability on the road network, which may result both from unforeseen changes in network conditions at a macroscopic scale (emergent traffic, weather, collisions or other incidents, etc.) or from variability in individual vehicle movements or signal phase timings at a microscopic scale. The stochasticity inherent in vehicle travel times means that any estimate of pickup or drop-off time made by a dynamic ridesharing operator is subject to variance, putting operators at risk of passenger delay beyond those estimates and of the passengers losing trust in their ability to serve trips in a timely manner.

Another aspect of uncertainty arising in dynamic ridesharing systems is the variability of future demand. Knowing when and where future trips will be requested from can have a large impact on the performance of different assignment strategies by increasing the system's capacity to handle future arrivals into the system, either by strategically positioning vehicles in areas where requests are likely to arrive or by choosing assignments of vehicles to passengers that allow for more potential to share future trips within the same vehicle without delaying the presently known passengers.

The final source of uncertainty we will touch upon in these applications is that of traveler choice, which is highlighted by the findings of the research on ridesharing demand in the previous chapter. Dynamic ridesharing systems provide information to travelers prior to their decision to request a ride from the service with estimates of the wait time, travel time, and price associated with the trip. Services which provide pooled trips alongside exclusive ones often show one set of estimates for each type of trip. The estimates provided will influence the traveler's decision to use either a pooled or exclusive trip or to balk entirely and choose another mode of transportation or to not make a trip at all. The method in which these estimates are generated and their capacity to influence the traveler's decisions is an underexplored area of the literature.

Meanwhile, the traveler's decision of which, if any, of the available trip offers to take is yet another source of uncertainty that relates to the questions of future demand and what estimates to provide to other riders. For instance, if several travelers in the same area receive the same trip offers and all choose to book an exclusive trip, the operator may find themselves in a situation analogous to an airline with an overfull flight because of overestimates for how many cancellations or no-shows would occur. While we do not consider these aspects of uncertainty in this chapter, we highlight this area as one where future research could be helpful to clarify and elaborate upon these ideas, as only one study considering this type of uncertainty has been conducted thus far [37].

While recent works have undertaken the topic of demand uncertainty in the context of online dynamic ridesharing applications including both passenger-vehicle assignment [38–40] and pricing [41], there are few papers that incorporate the existence of travel time uncertainty into decision-making for DRS system operators. In particular, vehicle-passenger assignment methods found in the literature to this point assume that travel times are static and known in advance of dispatching vehicles to serve passenger pickups and drop-offs [7, 42–45]. These known, static travel times allow these formulations to treat a passenger arriving past their latest acceptable arrival time as a binary outcome which can be fully controlled by the operator, and either incorporated as a fixed cost into the objective function or treated as a violation of a constraint leading to an infeasible assignment.

Meanwhile, there has been a comparatively large quantity of research on the related but distinct VRP-TW problem under travel time uncertainty [46–49]. The diverse formulations introduced by these papers make varying decisions and assumptions about the state of knowledge of this travel time uncertainty, the distributional forms underlying it, the operator's decision space, and how to optimally respond with this uncertainty. In addition, a few works investigate the offline ride-sharing or dial-a-ride problem under time uncertainty [50–52], which similarly assume that the operator has access to all trip requests for the entire planning horizon in advance but captures the problem-specific constraints on location order in vehicle routes as described before. Both categories of existing works are inapplicable to the specific context of the dynamic ridesharing assignment problem.

In this chapter, we seek to add to the sparse set of literature that bridges these two groups of existing research by proposing a formulation to the online dynamic ridesharing assignment problem under travel time uncertainty. This mixed integer stochastic programming problem is a novel contribution to the literature, and we additionally discuss the distribution-free sample average approximation method that can be used to solve this problem in a wide variety of contexts and states of knowledge regarding the structure of travel time uncertainty in the network of operation. we additionally discuss the computational challenges of solving this problem in a real-world online context and discuss several modifications to the formulation which can be used to improve the computational performance, along with experimental results demonstrating the effectiveness of these interventions across a variety of initial conditions, optimization parameters, and problem scales.

This formulation serves as a first step into the important space of online dynamic ridesharing operations that incorporate travel time uncertainty into the decision-making process. Further exploring this area of research and building upon this formulation will be crucial for the operation of pooled ridehailing services that reduce the variability of travel time, enhance customer trust in the service. These improvements should thereby attract more travelers to use this less space- and emissions-intensive form of on-demand transportation.

3.2 Literature Review

3.2.1 Dial-a-Ride Problem with Stochastic Travel Times

The first area of the literature which is of relevance to the problem of interest is work on the dial-a-ride problem with stochastic travel times. Building off of the literature on the vehicle routing problem with time windows, in the dial-a-ride problem a service operator must find an optimal assignment of vehicles to trip requests in the network. The full itinerary of trip requests for the planning horizon is known at the start of the problem. This literature draws on existing works on the Vehicle Routing Problem with Time Windows and Travel Time Uncertainty, but takes many different approaches in how to represent travel time uncertainty within the problem, how the problem is formulated, what objective function is used, and what solution method is used to find optimal or near-optimal assignments.

A 2002 paper introduces the consideration of travel time stochasticity to this problem space [50]. Assuming a normal distribution for travel times, with the parameters of the distribution varying over time, this problem optimizes vehicle routings to minimize a weighted combination of cost to the operator and cost to the passengers of the system. The operator's cost is represented by the expectation of total travel time for an assignment, and the passenger's cost is represented by deviation between the actual and desired times for both passenger pickup and drop-off. To solve this problem at large scales, the parallel insertion algorithm is extended to solve the problem given the changes to constraints and objective formulation given the travel time stochasticity. This algorithm begins with the earliest scheduled trips and iteratively inserts trips into vehicles' schedules according to the minimal cost, adjusting the costs as needed when the feasible pickup or drop-off time windows of two requests overlap.

The performance of the method is evaluated by simulating on a 20 km by 20 km synthetic grid network with uncorrelated, though time-varying, travel times and uniformly distributed demand. As this research predates much of the modern interest in this problem driven by on-demand ridesharing services, a very generous 30 minute maximum deviation from desired arrival time is used, and the maximum allowable ride time for any individual passenger is equal to 20 minutes plus twice the expected direct travel time from that passenger's origin to their destination. 120 vehicles with 8 seats per vehicle are employed to serve 2800 trips that arrive over a 10 hour simulation period. The insertion heuristics take upwards of 30 minutes to complete using a Pentium II 300 processor on a Windows NT computer with 256 MB RAM.

[51] implement stochastic travel times in an offline peer-to-peer ridesharing framework, where a subset of the requests provide their own vehicle and are allocated to either pickup and drop-off other requests within the system or to be picked up and dropped off by another driver. The cost function used to evaluate each trip is a weighted combination of the distance-based driving costs of the driver, the expected in-vehicle travel time costs (in terms of utility) for both the driver and passenger, and the expected penalties for both the driver and passenger for arriving to each of their destinations outside of their respective desired arrival time windows. The trip costs between any feasible pair of requests are evaluated using a Monte Carlo method with 10,000 simulations, which are then used as the input for a maximum weight bipartite matching problem that minimizes the total trip cost across the entire set of requests.

This approach is evaluated using the Chicago sketch network with 933 vertices and 2950 edges with 1000 to 10000 participants, with the travel times assumed to be distributed as shifted Gamma random variables. Compared to a ridesharing matching model that does not capture uncertainty, the authors find that their matching formulation is solvable in less computational time, though this would appear to be largely due to the consequence of fewer feasible matches once uncertainty is factored into the decision-making. This comparison also does not appear to include the preprocessing time spent to evaluate the costs of each pairing.

[53] develop a model for optimal offline assignment in the context of pooled taxi services with requests known in advance with a subset of the taxis designated as serving only women (although women can decide whether to request either a general taxi or a female-only taxi). They aim to minimize the system cost, which includes operation cost for the vehicle fleet, passenger travel time costs, and penalties for unanticipated delays. The authors formulate a discretized taxi flow time-space net-

work graph, separating the planning horizon into discrete intervals (the authors of this study use 10 minute intervals) and creating a vertex at each passenger origin or destination location for each time interval. Edges between vertices in the graph represent possible travel between two locations in the network in a certain number of discretized timesteps, with each physical flow having a discrete distribution of travel times. The costs of each edge represent the expected penalty if the taxi arrives at that location with that travel time, weighted by the probability of that travel time occurring. Using these graphs, the problem is formulated as an integer multiple-commodity network flow problem.

The size of the resulting graphs and the number of constraints and variables in the resulting formulation is very large, rendering traditional mathematical programming solvers unviable. To solve this problem, the authors use a heuristic method where a subset of the graph is initialized and used to formulate a reduced model, which is solved using CPLEX. Then, a local search algorithm tries to improve upon the resulting assignment using the full network. This method was tested using synthetic data on the Taipei road network using cost parameters obtained from travel surveys in the local region. For a scenario with 25 vehicles and 150 passengers over a 4 hour time horizon, the method yielded solutions within 12 to 60 minutes that were within a 4% optimality gap, and demonstrated significantly improved performance compared to a deterministic routing method. However, for problems that require a finer discretization than 10 minute time intervals, the corresponding problem size and therefore solution time would increase quickly.

[52] consider a case of the offline dial-a-ride problem where the operating fleet handles requests from both passengers and parcels (for our purposes, the formulation is equivalent to using two different classes of requests with different parameters). They formulate a two-stage stochastic optimization problem, with the upper-stage problem handling the logic of feasible routing and the lower-stage problems evaluating the routing of the upper-stage problem under either travel time or request location uncertainty. The authors implement several different solution methods for this problem: adaptive large neighborhood search (ALNS), sample average approximation, and a

sequential sampling procedure. Each of these methods are agnostic to the underlying distribution of travel times, instead sampling from whatever data may be available when solving – historical data, an estimated distribution, a possibility space, etc. The objective formulation is a weighted combination of the operator profit from satisfying trip requests, with a flat fee plus expected distance-based revenue as well as fuel or other operating costs that depend the actual distance traveled, combined with penalties for violating the requests' desired time windows and ride time constraints. The first component of the objective function is evaluated in the upper-level problem, while the second component is evaluated in the lower-level problem and added to the upper-level problem using optimality cuts in a Benders Decomposition framework.

This method is evaluated for a small problem with 30-75 requests using vehicles with capacity 5. The authors find that the performance of solutions to the problem is more sensitive to uncertainty in travel time than it is to uncertainty in the delivery locations of requests. Using a variety of sample sizes for the sampling-based solution methods, they found that a sample size of 200 was sufficient for adequate performance. Each of the three solution methods used performed similarly, though the sample average approximation method was more variable in performance. The solutions obtained using the stochastic methods are also compared to solutions from a deterministic solution, and the results indicate a large value of stochastic solution for this problem, which increases as the number of requests in the network grows.

Lastly, [54] investigate a dial-a-ride problem with some requests arriving dynamically during the planning horizon in addition to some requested in advance of operation. Travel times are modeled as deterministic, with randomly occurring traffic incidents that generate congestion (and therefore increased travel times) that spreads radially from the site of the incident throughout the network for a period of time, then subside. These events are highly impactful when they do occur and so it may be useful to find robust solutions that can handle these disruptions, but they are exceedingly rare compared to other sources of travel time variability; for instance, congestion arising without a traffic incident, traffic signal timings, individual vehicle speed fluctuations. The formulation proposed by this work is heavily dependent upon these specific assumptions, and so the relevance of the paper to other contexts with more widely applicable sources of uncertainty is limited.

A multi-tier objective is used, with assignments first being evaluated in terms of total violation of desired pickup or drop-off time windows, then proceeding to the number of vehicles used to serve the requests if the total time violations are equivalent, and finally using the total expected route duration if both of the first two objectives are equivalent. A block scheduling algorithm is used to generate the assignments, which combines insertion heuristics, used to assign the pre-scheduled requests to vehicles, with a dynamic stochastic variable neighborhood search (DS-VNS), which modifies the planned routes upon the entry of an online request to the system during operation. The procedure is evaluated using the Vienna road network over a 10-hour period, with vehicles that can serve up to 3 passengers at a time handling between 215 and 762 requests. Results indicate that the DS-VNS approach improves performance significantly when 40% or more of the requests in the system are arriving online, which highlights the importance of solution methods being able to find good solutions to these problems in an online context.

As can be seen, the literature in the space of the offline ridesharing or dial-a-ride problem with travel time uncertainty has little consensus in terms of problem context or solution approaches. The assumptions regarding travel times vary significantly, with two of the works considered assuming specific distributional forms, one assuming a 10-minute discretization of travel time distributions, and the last two using sampling-based approaches which are less dependent on specific problem context. In contrast to the insertion heuristic approaches that make use of assumptions regarding specific distributional forms, these sampling-based approaches find optimal solutions through integer programming formulations, although the bipartite matching formulation proposed by [51] relies on the offline context to guarantee optimality and would be unsuitable for handling requests that arrive online during operation. However, the findings of these studies build a consensus that ridesharing assignment methods that incorporate travel time stochasticity are able to generate much better solutions than those which do not.

3.2.2 Online DRS Assignment with Deterministic Travel Times

While several studies have considered offline ridesharing assignment problems with travel time stochasticity, most of the literature on online dynamic ridesharing assignment assume that travel times are deterministic. These works aim to efficiently solve the online problem iteratively, often in a reoptimization context where incoming requests over a given time period are batched together and given to the assignment problem at once. The assignment problem then must generate an optimal assignment and communicate it to the requests and drivers before the next batching period finishes. The performance and profitability of these systems can be highly dependent on the efficiency of the solution methods used, meaning that much of the work has pushed the computational edge of these problems by decomposing the problem into several sub-stages and/or parallelizing the workload across the system; however, the deterministic travel time assumption at the heart of these methods plays a large role in this speed.

[7] were one of the first to develop efficient methods to solve this problem optimally at large scales. As opposed to prior mixed integer programming formulations which can be computationally intensive and intractable at scale, the authors decomposed the solution method into computationally simpler phases such that a straightforward network flow formulation can be used to generate assignments, for which efficient and scalable solution methods exist. The first stage of this assignment method is to find pairwise matches of requests and vehicles which can be served feasibly. Next, a flow graph is constructed with edges traveling from nodes representing the vehicles to nodes representing feasible combinations of requests, and finally to nodes representing individual requests. The constraints on feasible matching of requests and vehicles include a maximum allowable travel delay relative to the shortest path travel time, and a latest acceptable drop-off time. Each edge is assigned a cost equal to the total delay beyond the shortest path travel time experienced by all requests in the trip. Finally, a minimum cost flow assignment integer program is solved, with an additional penalty for any requests which are unserved, and vehicles are reallocated through the

network based on any unmet demand.

Using the Manhattan road network and demand and speed data gathered from data provided by the New York City Taxi and Limousine Commission, one week of operations was simulated with batches of requests arriving every 30 seconds. Between 380,000 and 460,000 requests arrived in the system each day, and up to 2,000 requests were active within the system at any given time. Using a fleet of 10-seat vehicles, 99% of requests were able to be served with average waiting time and delay of 2.5 minutes using only 25% as many vehicles as the active taxi fleet of New York City.

Building off of this approach, [42] slightly modify the problem definition and use a federated architecture to distribute and parallelize the computation of routing costs across the fleet of operating vehicles, while a centralized server stores the incoming requests and solves the resulting assignment problem with the resulting costs. The problem formulation used in this research assumes a time window with an earliest allowable arrival time in addition to a latest, imposes a maximum travel distance constraint on each request, and assumes that only 1 request can be added to any vehicle's itinerary in each iteration, eliminating the need for a pairwise matching step. Upon receiving a batch of requests, the centralized server allocates each vehicle a subset of the requests based on geographic proximity, and each vehicle solves a single-vehicle dial-a-ride problem that minimizes total travel time. The resulting objective cost of vehicle/request pair is used as the edge costs in a matching problem formulation which is computed on the centralized server to generate the final assignment in each time period. This approach is tested using the same data as the previous study, and finds comparable performance with vastly reduced computational time.

Another notable approach considers a related problem, but only for the single occupancy case. [43] consider a taxi routing problem that attempts to maximize operator profit rather than minimize passenger delay. They formulate a problem for an online batching reoptimization framework to optimize the operation of single-occupancy vehicles considering the matching of subsequent requests in high-demand environments, where vehicles can be assigned a second passenger to serve after the drop-off of their immediately assigned passenger. A graph is constructed by connecting all requests, with known fixed pickup time windows and deterministic travel times, with edges that indicate the travel time from serving one request and driving to the pickup of the second as well as the profit from this sequence (the fare of the second trip minus the cost of the driving in between them). Conditions are put in place to assure an acyclic graph, i.e. the order of two requests being served cannot be reversed while still fulfilling all time window constraints.

In order to solve this problem efficiently for large networks, several strategies are developed to generate a sparser graph, upon which the optimal flow problem can be solved to yield a near-optimal result when considering the full graph. The authors define a "backbone" graph algorithm which samples random pickup times within each request's acceptable pickup time constraint and generates a maximum flow heuristic solution by using the simplex algorithm to solve the integral version of the assignment problem with these fixed pickup times, iteratively building up a sparse graph using only the edges included in these optimal solutions. In the online reoptimization context, the solution to each previous iteration is used as a warm start to the next iteration as many of the requests remain the same for short time intervals. Using the same Manhattan taxi dataset, the new approaches devised by this paper demonstrate improvements of profits between 1% and 3.5%, which represent hundreds of passengers' worth of additional capacity for the system.

A few other approaches have been tried which incorporate other considerations into the basic online ridesharing problem framework. [44] investigates how to optimally route a fleet of shared autonomous vehicles considering the resulting traffic congestion through a network, assuming all traffic in the network is being served by this fleet. They incorporate a link transmission model, which models how vehicles flow from one link in the network to the next given link capacities and intersection flow capacities, into an assignment framework. This problem minimizes the total system travel time, computed by multiplying each link's occupancy in each timestep by the duration of that timestep. This formulation was evaluated with tests on a small 2 mi × 2 mi network with 24 total links, with demand arriving for 10 30-second timesteps. Computation times were reported to take between 35 and 45 minutes per scenario, demonstrating the computational difficulties that can arise when modifying the basic assumption of deterministic link travel times.

Finally, [45] investigate a problem where passengers have the capability to reject a match made by the operator if they deem it unsatisfactory. They devise a centralized system that receives vehicle information and requests and solves a mixed integer optimal matching problem for each request that enters the system given the current state (as opposed to the batching method used in other approaches). This approach is only tested on small networks with few users in the context of the Genoa network, though it does demonstrate a significant reduction in match refusals compared to a method which only serves requests along the drivers' fixed paths through the network.

This body of literature demonstrates methods that can be used to efficiently solve the deterministic online dynamic ridesharing problem, which can be used to generate assignments in real-time operations even for extremely large fleets and high demand volumes. However, the performance of these methods in environments with uncertain travel times remains unexamined by the literature. Because of the strong assumptions made by these efficient solution methods that requests and vehicles can be matched as long as the expected travel time falls within a defined time bound, solutions could be sensitive to violating these time bounds in real-world contexts where a deterministic travel time does not exist. For methods that do incorporate other aspects of the problem, such as link travel times, computational issues again re-emerge, demonstrating the simplifications of the problem that are needed for efficiency.

3.2.3 Online Vehicle Routing with Stochastic Travel Times

In recent years, there has been a small but growing body of literature that incorporates stochastic travel times into online vehicle routing methods. In the early stages of work in this field, many different approaches have been taken to incorporate concepts of travel time uncertainty into this problem and deal with the computational challenges that arise from it.

Although developments for high capacity online dynamic ridesharing fleets that incorporate stochastic travel times have been introduced only recently, the first related work was undertaken by [17], who investigated the dynamic dial-a-ride problem under stochastic environments. The authors consider a wide variety of stochastic events which may arise during operation, including new service requests at random locations, stochastic travel times, stochastic service times at pickup and drop-off locations, customer absences, cancellations, traffic jams, and vehicle breakdowns. The method implemented to deal with these stochastic events relies on local search, with schedules generated by inserting requests into a vehicle's existing route or swapping requests between vehicles iteratively until a good solution is found. Diversification strategies are also employed to better avoid local optima. The different events that may occur are enumerated, and each follows a specified procedure for adjusting the system state and recomputing the heuristic optimal assignment.

This method is simulated for a 10-hour day, with maximum allowable wait time of 20 minutes and maximum allowable extra travelling time of 45 minutes and demand rates between 0.1 and 1 request per hour. Responding to new events in the system takes up to 1 to 2 minutes in 5% of cases, and provides an improvement of between 4% and 20% in terms of system cost compared to the deterministic methods, with a greater benefit at lower rates of incoming demand.

[55] also examine online approaches to responding to delays that arise in dial-a-ride problem contexts and develop optimal stopping approaches that can be used in situations when the realization of stochastic travel times delays a vehicle to the point where it will arrive late and there are backup services which can be used. The optimal stopping policies define under what conditions in the problem state will an irreversible recourse action be taken, i.e. calling a taxi to serve a passenger that is going to be delayed. The authors formulate the optimal policy as a stochastic dynamic program, and find approximate solutions using a binomial lattice method. They additionally provide a two-stage stochastic programming formulation that could be used to integrate dynamic recourse actions into offline scheduling and routing for the dial-a-ride problem where backup services are available.

Finally, only one paper found in the literature examines the online dynamic ridesharing problem under stochastic travel time uncertainty focused on large-scale applications rather than dial-a-ride services. [16] extend the deterministic formulation proposed by [7] by using a stochastic storing step in the request-vehicle matching flow graph process to include reliability information within the assignment formulation. The authors aim to optimize the service rate and the overall reliability of the assignment in each dispatch iteration, which is obtained by maximizing the total expected probability of each vehicle succeeding in picking up and dropping off their requests within the allowable time windows, minus a penalty for each request left unserved. To obtain these probabilities, their solution method enumerates the α-shortest paths between each pair of locations in a given vehicle-trip assignment and uses a recursive algorithm to synthesize these distinct paths into an overall route with a maximum α value for the final destination. Using these reliability estimates, the problem can then be formulated as an integer linear program.

The authors assume that travel times are Gaussian distributed and independent across links, and cite a paper by [56] to argue that it is a realistic assumption. However, those properties are an assumption made by the authors for their models: their own comparison of the empirical dataset used with Gaussian random variables finds that this assumption does not hold on the tails of the distributions, with the lowest and highest 5% of travel time outcomes both being underestimated, and they provide no evidence for or against independence of these distributions. In contrast, transportation literature focused on investigating the distributional form of travel time uncertainty using empirical travel time data finds that the distributions show significant skew [57, 58] and that link travel time covariance is nonzero [59, 60]. The efficiency of this reliability-aware assignment method depends on these assumptions as the computational complexity of finding α-reliable paths is greater for networks with more complex distributional forms and correlated link travel times [61].

Using the same New York taxi dataset as the deterministic baseline method, the authors evaluate the method, with the mean and standard deviation of each link's Gaussian distributed travel time equal to the observed mean and standard deviations derived from processed taxi data. The maximum allowable travel time delay is equal to twice the maximum allowable wait time delay, which varies between 3 and 7 minutes. The algorithm is run every 30 seconds for a one-hour simulation period. It demonstrates significant improvements in performance relative to the baseline, with the proportion of trips which experience constraint violation decreasing by up to 7.3 percentage points. Computational performance, as well as the average passenger delay, is not reported.

As made clear by the current literature, there are a large amount of considerations to make in formulating a method for DRS assignment in a stochastic travel time environment. A formulation must decide how to handle the many facets of ridesharing operations, the breadth of ways to consider and address travel time uncertainty, and the interactions between these two elements. In the next section, we propose a two-stage stochastic optimization formulation which uses a sampling-based approach to gather information about the travel times, requiring no assumptions about the distributional form of the travel time uncertainty. This formulation extends the work of [52] into an online reoptimization context, and aims to minimize the total expected traveler delay. This combination of approaches to the problem makes it a novel contribution to the literature.

3.3 Formulation

3.3.1 Definition

This problem takes a batching approach where a bundle of incoming trip requests are grouped together and given to the assignment algorithm at once, at which point an assignment is generated and vehicles are routed accordingly. The system operator collects the unserved requests and passes them, along with the set of requests that have not yet been fulfilled by dropping off the associated passengers at their destination, into this assignment problem at a time interval of ΔT. These N trip requests $\mathcal{R} = \{r_1, ..., r_N\}$ represent the set of customers reserving rides in the system. Each request $r \in \mathcal{R}$ is represented by a tuple $\left(o_r, d_r, \ell_r^o, \ell_r^d, q_r\right)$, where o_r is the origin (pickup location), d_r is the destination (drop-off location), ℓ_r^o and ℓ_r^d are soft upper time bounds for the origin and destination, respectively, and q_r is the number of passengers associated with that request (a family or group of friends traveling together may have $q_r > 1$). We assume that these time bounds are fixed, reflecting the agreement already made between the ridehailing operator and the passenger at the time the ride was booked.

The operator has perfect knowledge of the states of the K vehicles it operates in the network, represented by the set $\mathcal{K} = \{v_1, ..., v_K\}$. Each vehicle $k \in \mathcal{K}$ has a current location s_k and a capacity C_k which represents how many passengers it can hold at once.

We use the information from the set of requests and vehicles to generate a directed flow graph, which is defined as:

$$\mathcal{G} = (\mathcal{V}, \mathcal{E})$$
$$\mathcal{V} = \mathcal{V}^s \cup \mathcal{V}^p \cup \mathcal{V}^d \cup \mathcal{V}^0$$

Where $\mathcal{V}^s = \{s_{v_1}, ..., s_{v_K}\}$ is the set of starting locations of the vehicles in the current system state, $\mathcal{V}^p = \{o_{r_1}, ..., o_{r_N}\}$ is the set of pickup locations of all requests, $\mathcal{V}^d = \{d_{r_1}, ..., d_{r_N}\}$ is the set of drop-off locations of all requests, and $\mathcal{V}^0$ is a dummy sink vertex representing the completion of each vehicle's route. Each pickup location vertex $v_r \in \mathcal{V}^p$ is associated with the time limit ℓ_r^o and passenger quantity q_r of its corresponding request, and each drop-off location similarly has time limit ℓ_r^d and quantity $-q_r$.

We borrow and extend the indexing of vertices used in [52] to easily reference the index of any node in the formulation. With K vehicles in the system and N requests, the vertices are indexed according to the following scheme:

$$
\begin{aligned}
&\mathcal{V}^s := \{1, 2, ..., K\}\\
&\mathcal{V}^p := \{K+1, K+2, ..., K+N\}\\
&\mathcal{V}^d := \{K+N+1, K+N+2, ..., K+2N\}\\
&\mathcal{V}^0 := K+2N+1
\end{aligned}
$$

These vertices are connected by a set of M directed edges:

$$\mathcal{E} = \{e_1, ...e_M\} \subset \{(i,j)\,|i \in \mathcal{V}, j \in \mathcal{V}, i \neq j\}$$

These flow edges are an abstraction from the road network and instead correspond to a vehicle movement through the road network starting at the location represented by the upstream node and ending at the location represented by the downstream node. Each edge e is associated with a distance d_e and distribution of travel times t_e. In order to generate the distance and travel time distribution for each edge, the underlying road network will generally need to be used to find a route between the upstream and downstream location. For this work, we find the shortest path routing for each edge using the median travel time along each edge as the edge weights.

It is important to note that this approach treats the routing as exogenous to the assignment problem and independent of the combination of requests being served by the vehicle and their associated time constraints. This is not necessarily the case, however, as an operator could choose different routes between the same two locations depending on the relative risk profile of violating the associated constraints of that journey, as in [16]. This limitation could be mitigated by allowing multiple edges between any pair of nodes with different distances and travel time distributions reflecting different possibilities for the underlying path, but this direction of investigation is not pursued as part of the research presented here.

An edge is added connecting any two vertices in the network as long as a movement between those two vertices corresponds to a feasible movement. An edge travels from each vehicle's starting location to each pickup location, but not to any drop-off locations as a vehicle cannot drop-off a passenger as the first action on its route. Each pickup location has edges flowing to each other pickup location and each drop-off location. Each drop-off location has edges flowing to each other drop-off location and each pickup location besides the one associated with its own request, as a movement from a request's destination to its origin doesn't make physical sense.

We also define some special cases in the initialization of the network. Each vehicle's starting location $v_s \in \mathcal{V}^s$ is connected to the dummy terminal vertex by an edge $(v_s, \mathcal{V}^0)$, and each dropoff location $v_d \in \mathcal{V}^d$ is also connected to the dummy terminal vertex by an edge $(v_d, \mathcal{V}^0)$. If a vehicle has already visited the pickup locations of one or more passengers but not yet dropped them off, they are converted to dummy nodes and connected serially to the vehicle's starting location node. All other edges leaving the vehicle's starting location node are removed, including the edge connecting it to the terminal vertex, as the vehicle must visit at least the dropoff locations of its current passengers on its route and therefore a flow directly from its initial location to the sink node (representing no movement) is not allowed.

Figure 3-1 shows the an example of this process. The initial system state is depicted in Figure 3-1a. 2 vehicles, V_1 and V_2, are being assigned to serve 3 requests in the system. Request 1 has already been picked up by vehicle 1, so a dummy origin location O_1 is used for it and created in the same location as V_1, while D_1 designates its dropoff location. Neither request 2 nor request 3 has been picked up yet, so O_2 and O_3 represent their pickup locations and D_2 and D_3 their dropoff locations. Figure 3-1b depicts the flow graph realization of this system state, with a dummy terminal node $\mathcal{V}^0$ added and edges drawn between the nodes corresponding to the possible feasible movements of each vehicle.

Figure 3-1c shows a hypothetical routing solution to this problem, with vehicle 1 assigned to pick up request 3, then drop off request 3, then drop off request 3, while vehicle 2 is assigned to pick up and then drop off request 2. Figure 3-1d isolates the flow edges corresponding to this routing.

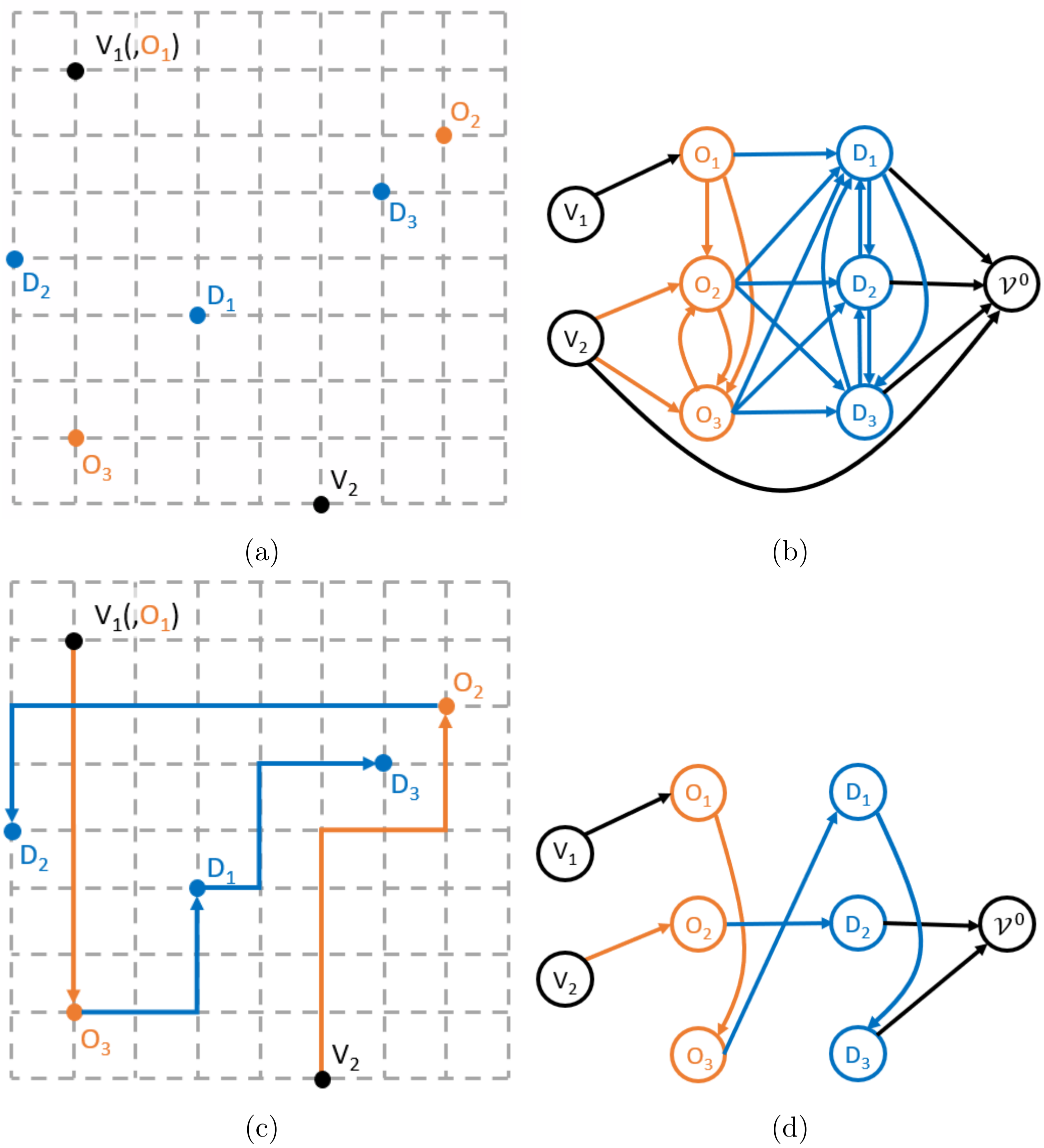


Figure 3-1: Example of network graph conversion into flow graph for stochastic assignment formulation. Subfigure (b) shows the complete flow graph constructed from the scenario depicted in subfigure (a), with 2 vehicles and 3 requests and request 1 having already been picked up by vehicle 1. Subfigure (d) shows the flow along edges corresponding to the routing in subfigure (c).

For each vertex $v \in \mathcal{V}$ in the network, we further define the following sets of edges associated with that vertex representing its outbound and inbound edges for convenience in the formulation:

$$\mathcal{E}^+ = \{(i,j) \in \mathcal{E} | i = v\}$$
$$\mathcal{E}_v^- = \{(i,j) \in \mathcal{E} | j = v\}$$

Finally, we define a binary variable $\omega_i k$ which indicates the status of any assignments made in earlier iterations of the assignment algorithm. $\omega_i k$ takes a value of 1 if and only if the request associated with vertex $i \in \mathcal{V}^p$ was last assigned to vehicle $k \in \mathcal{V}^s$ in the most recent passenger-vehicle assignment, and is 0 otherwise.

3.3.2 Two-Stage Stochastic Programming Formulation

Once the flow graph has been initialized, we define the following decision variables:

x_{ek} = 1 if vehicle k travels on flow edge

Δ_{ik} = 1 if vehicle k is newly assigned to request i (i.e. pickup node $K+i$) and was not assigned to that request in the latest assignment iteration. 0 otherwise.

n_{ik} = the number of passengers in vehicle k after it serves location i

P_{ik} = the order of location $i \in \mathcal{V}$ in vehicle k's route, in descending order. The sink node is fixed to have $P_{\mathcal{V}^0,k} = 0 \forall k \in \mathcal{V}^s$. Then, the final location that vehicle k serves before the sink node has $P_{ik} = 1$, and so on.

η_i = the time that a vehicle arrives at location i. If $i \in \mathcal{V}^p$, then η_i is the pickup time of the associated request. If $i \in \mathcal{V}^d$, then η_i is the drop-off time of the associated request.

Then, we solve a two-stage stochastic optimization problem to identify the optimal assignment accounting for travel time uncertainty. The upper stage is as follows:

$$\min_{x, P, n, \Delta} \quad \gamma_1 \sum_{e \in \mathcal{E}} \sum_{k \in \mathcal{V}^s} d_e x_{ek} + \gamma_2 \sum_{i \in \mathcal{V}^p} \sum_{k \in \mathcal{V}^s} \Delta_{ik} + \gamma_3 \sum_{i \in \mathcal{V}^p} \mathbb{E}[\eta_i] + \gamma_4 \sum_{i \in \mathcal{V}^d} \mathbb{E}[\eta_i] \tag{3.1a}$$

$$\text{s.t.} \qquad \sum_{e \in \mathcal{E}_k^+} x_{ek} = 1, \qquad \forall k \in \mathcal{V}^s, \tag{3.1b}$$

$$\sum_{e \in \mathcal{E}_{\mathcal{V}0}^-} x_{ek} = 1, \qquad \forall k \in \mathcal{V}^s, \tag{3.1c}$$

$$\sum_{e \in \mathcal{E}_i^-} x_{ek} - \sum_{e \in \mathcal{E}_i^+} x_{ek} = 0, \qquad \forall i \in \mathcal{V}^p \cup \mathcal{V}^d, \forall k \in \mathcal{V}^s, \tag{3.1d}$$

$$\sum_{e \in \mathcal{E}_i^-} x_{ek} - \sum_{e \in \mathcal{E}_{i+R}^+} x_{ek} = 0, \qquad \forall i \in \mathcal{V}^p, \forall k \in \mathcal{V}^s, \tag{3.1e}$$

$$P_{ik} - P_{jk} \leq 1 - M\left(x_{(i,j)k} - 1\right), \forall (i,j) \in \mathcal{E}, \forall k \in \mathcal{V}^s, \tag{3.1f}$$

$$P_{ik} - P_{jk} \geq 1 - M\left(1 - x_{(i,j)k}\right), \forall (i,j) \in \mathcal{E}, \forall k \in \mathcal{V}^s, \tag{3.1g}$$

$$P_{ik} - P_{i+R,k} \geq \sum_{e \in \mathcal{E}_i^+} x_{ek}, \qquad \forall i \in \mathcal{V}^p, \forall k \in \mathcal{V}^s, \tag{3.1h}$$

$$n_{jk} - n_{ik} \geq q_j - M\left(1 - x_{(i,j)k}\right), \forall (i,j) \in \mathcal{E}, \forall k \in \mathcal{V}^s, \tag{3.1i}$$

$$n_{ik} \leq C_k, \qquad \forall i \in \mathcal{V}^p \cup \mathcal{V}^d, \forall k \in \mathcal{V}^s, \tag{3.1j}$$

$$\Delta_{ik} + \omega_{ik} \geq \sum_{e \in \mathcal{E}_i^+} x_{ek}, \qquad \forall i \in \mathcal{V}^p, \forall k \in \mathcal{V}^s, \tag{3.1k}$$

$$\sum_{e \in \mathcal{E}_i^-} \sum_{k \in \mathcal{V}^s} x_{ek} = 1, \qquad \forall i \in \mathcal{V}^p, \forall k \in \mathcal{V}^s, \tag{3.1l}$$

$$P_{\mathcal{V}^0,k} = 0, \qquad \forall k \in \mathcal{V}^s, \tag{3.1m}$$

$$\Delta_{ik}, P_{ik}, n_{ik} \geq 0, \tag{3.1n}$$

$$x_{ek} \in \{0,1\} \tag{3.1o}$$

The upper level problem contains the constraints on the vehicle flows x_{ek} along the flow graph edges $\mathcal{E}$ that identify feasible routings. Constraints (3.1b) through (3.1e) enforce flow conservation through the network and require each vehicle to visit both the pickup and drop-off locations of any requests that they are routed to. Constraints (3.1f) and (3.1g) dictate the order of the pickup and drop-off actions taken along the vehicle's route as a consequence of the edges the vehicle uses to traverse the network, and (3.1h) requires that a request's pickup location must be visited before the drop-off location (i.e. the drop-off location comes later in the vehicle's location order). Constraint (3.1m) ensures that the order variable of the final visited passenger location on the vehicle's route will be at least 1, which will be important for methods used later in this chapter. Constraints (3.1i) and (3.1j) are vehicle capacity constraints which ensure that the number of passengers in any vehicle throughout its route never exceeds its capacity.

Constraint (3.1k) ensures that any vehicle/passenger combination that was not assigned in a previous timestep ($\omega_{ik} = 0$) is treated as a new assignment ($\Delta_{ik} = 1$), which is assigned a cost γ_2 in the objective function. Constraint (3.1l) ensures that exactly 1 vehicle is assigned to every request.

The objective function used in this formulation is a weighted combination of several components. The first, with weight γ_1, is the total distance traveled by vehicles in the assignment. Reducing the total distance traveled is an important consequence of dynamic ridesharing services in mitigating the impacts of on-demand rides on emissions and other undesirable externalities. Without a penalty for distance traveled, a system may aim to use as many vehicles as possible to serve the incoming requests to minimize the chances of delay, but the operator or regulator of the service may instead prefer for the service to pool trips which are shareable with only a small penalty to passenger delay in order to reduce these externalities. The weight used in the objective function represents the distance-based variable costs to the operator for vehicle operations, such as energy consumption or any road pricing schemes imposed by regulators.

The second, with weight γ_2, is the number of requests which are assigned to a new vehicle in this assignment compared to the last assignment. All incoming requests will always have $\Delta_{ik} = 1$ for any assignment as we do not allow for the rejection of requests. Possible assignments will therefore differ in this portion of the objective function only in the case where one or more requests have their assigned vehicle change. This component of the objective function therefore represents the costs the operator faces in altering a previously committed vehicle's route once that vehicle is on its way, which in practice could reflect payment to the vehicle's driver for the inconvenience of changing route or other costs the passenger imposes on the system such as a loss of trust.

The third and fourth components of the objective, with weights γ_3 and γ_4 respectively, are the expected delays faced by each passenger beyond their latest acceptable pickup time (γ_3) and drop-off time (γ_4). Allowing for different weights on the two reflects the finding from the previous chapter that passengers treat delay beyond the estimated pickup time differently than that beyond the estimated arrival time at their destination when deciding to book a ride from a DRS service.

The weights reflect the costs the operator faces if it fails to meet the time estimates given to the passenger upon booking. For a private operator, the costs borne by the service will largely be long-term costs such as those arising from a loss of future demand due to a lack of trust in on-time performance. For a public service operator or a private operator subject to some regulation by a private entity, these costs could instead directly reflect the passengers' loss of utility from the delay. In the first case, future research is needed to further our understanding of the dynamics by which a lack of trusts develops based on experienced delays.

These delays are a function of the random variables t_e describing the distribution of link travel times for each edge in the network. In this formulation, there are several reasons why this has been chosen as the objective component in place of other representations of travel time variability observed in the literature. Delay is a linear combination of the decision variables, which is important for computational reasons. When compared to a chance-constrained formulation or one that aims to maximize the on-time arrival probability, we note that these alternative representations of uncertainty have the quality that a request which is very unlikely to arrive on time may be delayed to a great extent in order to better allow other requests to arrive on time. However, we view this behavior as more undesirable than the inverse which may arise in this case, as we hypothesize that the magnitude of the delay has a larger influence on loss of trust or loss of passenger utility than the binary indicator of whether the trip arrived on-time. Further research is needed to investigate this assumption to develop a better understanding as to what approach is best from a passenger utility maximization perspective.

One final alternative that was considered was the collective requirement violation index (CRVI) proposed by [46]. However, this has a few undesirable properties for the specific context of this online assignment problem. First, it requires knowledge of or assumptions on the distributional form or families of distributional forms that the route travel times may take. Additionally, it has the property of abandonment, where a random variable whose expected value falls outside of the required bounds has a cost of $+\infty$. In an online reoptimization context, this has troubling consequences. For example, a request that has already been picked up whose expected arrival time was originally within its time constraint may face delays as its assigned vehicle progresses along its route to the point where its expected arrival time now falls outside the time bound. In this case, the cost of an assignment including that request will now be unboundedly large, and yet it must still be included in its assigned vehicle's route. In this case, the expected delay metric that we use for the objective will instead still aim to choose assignments that minimize the extent of this delay, which is more desirable behavior.

In order to compute the expected delay for an assignment, we use a sample average approximation approach, which eliminates the need for a rigorous definition of the uncertainty space of the travel times and instead the use of any method which can generate a set of possibilities for them, such as sampling from historical data or ensemble-based traffic forecasting. We definite lower-stage problems $s = \{1, ..., S\}$ which each reflect a realization of the uncertain travel times on the road network which occurs with probability p_s. We represent the travel time on each flow network edge $e \in \mathcal{E}$ under the s-th realization as $\tilde{t}^s_e$. Each of these problems are formulated as follows:

$$
\begin{aligned}
Q_s\left(x\right) = \min_{\tau, \eta} \quad & \gamma_3 \sum_{i \in \mathcal{V}^p} \eta^s_i + \gamma_4 \sum_{i \in \mathcal{V}^d} \eta^s_i && && \text{(3.2a)} \\
\text{s.t.} \quad & \tau_{jk} - \tau_{ik} \geq \tilde{t}^s_{(i,j)} - M\left(1 - x_{(i,j)k}\right) , && \forall \left(i, j\right) \in \mathcal{E}, \forall k \in \mathcal{V}^s, && \text{(3.2b)} \\
& \eta_i \geq \sum_{k \in \mathcal{V}^s} \tau_{ik} - \ell_i, && \forall i \in \mathcal{V}^p \cup \mathcal{V}^d, && \text{(3.2c)} \\
& \eta_i, \tau_{ik} \geq 0 && && \text{(3.2d)}
\end{aligned}
$$

Constraint (3.2b) relates the arrival time of each vehicle at each node in the network to the network flow variables from the upper level problem. If a vehicle $k \in \mathcal{V}^s$ flows along the edge from vertex i to vertex j, then that vehicle's arrival time at the downstream vertex j is equal to its arrival time at the upstream vertex i plus the realized travel time along edge (i, j), $\widetilde{t}^s_{(i,j)}$. We use a big-M formulation to ensure that vehicles' arrival times are not impacted by the edges that they do not flow along. Constraint (3.2c) relates the passenger delay at a specific pickup or drop-off location to the vehicle arrival times at that location and its associated time constraint. This constraint combined with the non-negativity constraint and the minimization objective function is equivalent to $\eta^s_i = \max\left(\max_{k\in\mathcal{V}^s} \tau^s_{ik}, 0\right)$.

The objective function of each subproblem is the contribution of the passenger delays in the subproblem to the expected values in the original objective function. We can therefore introduce a variable θ_s to the objective function for each subproblem and add constraints relating the objective value of each subproblem to its corresponding variable in the form:

$$\theta_s \geq Q_s\left(x\right) \quad \forall s \in \{1, ..., S\} \tag{3.3}$$

The objective function can therefore be written as a function of the introduced variables, each weighted by the associated probability of the realization of travel times corresponding to the related subproblem:

$$\min_{x,P,n,\Delta,\theta} \gamma_1 \sum_{e\in\mathcal{E}} \sum_{k\in\mathcal{V}^s} d_e x_{ek} + \gamma_2 \sum_{i\in\mathcal{V}^p} \sum_{k\in\mathcal{V}^s} \Delta_{ik} + \sum_{s=1}^{S} p_s \theta_s \tag{3.4}$$

This problem can then be solved using an iterative Benders decomposition approach. At first, the upper-level problem is solved without any constraints imposed by the lower-level subproblems. When an integral candidate solution $\left(\hat{x}^t, \hat{P}^t, \hat{n}^t,\ \hat{\Delta}^t, \hat{\theta}^t\right)$ is found, the $\hat{x}^t$ are passed into the lower-level subproblems and used to generate cuts on the feasible region of the upper-level problem. This procedure repeats until a solution to the upper-level problem is found that is feasible under all cuts generated by the lower-level problems, which is guaranteed to be optimal.

The Benders cuts added by each subproblem are found by taking the dual of the lower-level subproblem, which can be written as follows:

$$
\begin{aligned}
Q_s^t\left(\hat{x}^t\right) = \max_{\pi,\,\lambda} \quad & \sum_{e\in\mathcal{E}}\sum_{k\in\mathcal{V}^s} \pi_{ek}^s\left(\tilde{t}_e^s - M\left(1-\hat{x}_{ek}^t\right)\right) - \sum_{i\in\mathcal{V}^p\cup\mathcal{V}^d} \ell_i\lambda_i^s && \text{(3.5a)}\\
\text{s.t.} \quad & \sum_{e\in\mathcal{E}_i^-}\pi_{ek} - \sum_{e\in\mathcal{E}_i^+}\pi_{ek} + \lambda_i \geq 0, \quad \forall i\in\mathcal{V}^p\cup\mathcal{V}^d, && \text{(3.5b)}\\
& \lambda_i \leq \gamma_3, \qquad \forall i\in\mathcal{V}^p, && \text{(3.5c)}\\
& \lambda_i \leq \gamma_4, \qquad \forall i\in\mathcal{V}^d, && \text{(3.5d)}\\
& \pi_{ek}, \lambda_i \geq 0 && \text{(3.5e)}
\end{aligned}
$$

Solving this subproblem for a given solution $\hat{x}^t$ yields the optimal values of the dual variables $\hat{\pi}^{s,t}$ and $\hat{\lambda}^{s,t}$, which we can use to generate an optimality cut on the upper-level problem as a function of the upper-level decision variables x:

$$
\theta_s \geq \sum_{e\in\mathcal{E}}\sum_{k\in\mathcal{V}^s} \hat{\pi}_{ek}^{s,t}\left(\tilde{t}_e^s - M\left(1-x_{ek}\right)\right) - \sum_{i\in\mathcal{V}^p\cup\mathcal{V}^d} \ell_i\hat{\lambda}_i^{s,t} \tag{3.6}
$$

Note that the lower-level subproblem is guaranteed to be feasible for a feasible solution to the upper-level problem, and therefore the only cuts that will be generated are these optimality cuts. The potential for feasibility cuts in this problem is discussed further in section 3.4.1.

This formulation and solution method guarantees an assignment of passengers to vehicles in the network that gives the minimal average weighted passenger delay over the sample of network travel times used to generate the subproblems, and further incorporates operator costs for vehicle distance traveled and passenger reassignment. This contributes to the literature by extending stochastic programming methods used in offline vehicle routing problems to the online dynamic ridesharing assignment problem, which heretofore did not adequately incorporate the impacts of travel time stochasticity and its impact on passenger delay. In practice, however, this solution method can be computationally expensive to solve as the Benders decomposition is known to often lead to slow convergence even when incorporated into a branch-and-cut framework. The following section discusses further extensions to this formulation which aim to reduce the computational burden of the problem while retaining optimality.

3.4 Problem Extensions to Improve Tractability

Mitigating the computational complexity of the problem to increase the speed at which optimal solutions can be found is important for this method to be used in practical applications given the speed at which online reoptimization methods are needed to operate. Outlining the steps involved in the solution method highlights some areas where complexity may arise:

- First, an integral candidate solution to the upper-level problem needs to be found. The number of constraints in the upper-level problem is large and grows quickly as the scale of the problem increases, as many of them apply once to each edge-vehicle combination. The number of edges in the flow network can be approximated by $K + KN + 4N^2$, where K is the number of vehicles and N is the number of requests (it will be marginally less than this if some of the requests have already been picked up by vehicles). Furthermore, many of these constraints are logical "big-M" constraints, which are known to result in poor computational performance for mixed integer linear programs as the LP relaxation is loose. Improving the speed at which integral solutions can be found would reduce the computational burden spent on this element.

- Once an integral solution has been found, an optimal solution to each subproblem needs to be found. As all variables in the subproblems are continuous, each one is not a large computational burden, but due to the sample average approximation approach we may find it desirable to have a large number of travel time samples and hence a large number of subproblems which need to be solved for each integral solution. Reducing the number of integral solutions that must be explored in this manner in order to guarantee an optimal solution would reduce the number of these subproblems that need to be solved and therefore improve the solution efficiency.

- Returning to the issue of the "big-M" constraints, once the subproblems are solved the optimality cuts added to the upper-level problem are also of this variety. This may result in increasing difficulty to find integral solutions as the feasible space is explored, generating optimality cuts that are added to the problem. Reformulating the added cuts in a way which increases the tightness of the LP relaxation of the resulting MIP would help to prevent this problem.

To address the first issue, we harness methods from the literature on Combinatorial Benders (CB) [62] to add a second class of lower-level subproblem which is used to generate feasibility cuts on the upper-level problem in contrast to the optimality cuts generated by the previously described stochastic subproblems. This method aims to improve the speed at which integral solutions can be found by eliminating the logical constraints from the upper-level problem.

To address the second issue, we adopt two approaches. First, we add to the upper-level problem constraints which provide a lower bound on the expected objective contribution of each subproblem as a function of the upper-level decision variables, which reduces the need to explore some of the feasible region after a higher-quality solution has already been found. Second, we harness the problem structure to formulate some additional optimality cuts which can be added to the upper-level problem after solving the subproblems which provide lower bounds on the delay contribution for similar solutions without needing to solve a new set of subproblems for those solutions.

The third issue raised is one that remains unaddressed by these extensions. Formulating approaches to resolve this issue would be a key contribution by future literature to improve the efficiency of this solution method.

3.4.1 Combinatorial Benders Decomposition

Passenger-vehicle assignment problems require constraints on feasibility to obey certain logical conditions, such as the requirement that a vehicle's route must visit the node corresponding to a passenger's pickup location before it can visit the node corresponding to that passenger's drop-off location. Implementing these conditions on the feasibility space in a mixed integer linear program often requires the use of "big-M" constraints that enforce an if-then condition by multiplying an integer variable by a large value of M. This has the issue of reducing the quality of the LP relaxation of the integer program, which leads to poor performance of branch-and-cut solution methods to IPs

To address these issues, we separate out these logical constraints into a subproblem which is used to generate Combinatorial Benders (CB) cuts that eliminate infeasible solutions from the search space without sacrificing the quality of the LP relaxation of the upper-level problem. The method described below is applied to this formulation using the framework established by [62].

For a given integral candidate solution $\hat{x}^t$, the logical constraints used to create the combinatorial subproblem used in this method are exactly constraints (3.1f) through (3.1j) and constraint (3.1m) from the original upper-level problem. We remove these constraints from the upper-level problem and formulate a system of linear equations using the candidate solution:

$$\begin{aligned}
\text{SLAVE}\left(\hat{x}^t\right) = & P_{jk} - P_{ik} \leq 1 - M\left(\hat{x}^t_{(i,j)k} - 1\right) && \forall\,(i,j) \in \mathcal{E}, \forall k \in \mathcal{V}^s && (3.7a)\\
& P_{jk} - P_{ik} \geq 1 - M\left(1 - \hat{x}^t_{(i,j)k}\right) && \forall\,(i,j) \in \mathcal{E}, \forall k \in \mathcal{V}^s && (3.7b)\\
& P_{i+N,k} - P_{ik} \geq \sum_{e \in \mathcal{E}_i^+} \hat{x}^t_{ek} && \forall i \in \mathcal{V}^p, \forall k \in \mathcal{V}^s && (3.7c)\\
& n_{jk} - n_{ik} \geq q_j - M\left(1 - \hat{x}^t_{(i,j)k}\right) && \forall\,(i,j) \in \mathcal{E}, \forall k \in \mathcal{V}^s && (3.7d)\\
& n_{ik} \leq C_k && \forall i \in \mathcal{V}^p \cup \mathcal{V}^d, \forall k \in \mathcal{V}^s && (3.7e)\\
& P_{\mathcal{V}^0,k} = 0 && \forall k \in \mathcal{V}^s && (3.7f)\\
& P_{ik}, n_{ik} \geq 0
\end{aligned}$$

The goal of the combinatorial subproblem is to identify whether the candidate solution $\hat{x}^t$ is feasible according to logical constraints on ordering pickup and dropoff locations within a route and on vehicle capacity. If Problem (3.7) is infeasible, then combinatorial Benders cuts are added to the upper level problem to remove the candidate solution from the feasible space of the problem. This is done by finding a minimal infeasible system (MIS) of constraints. This MIS can be found by solving the following LP, based on the dual of an LP using the system of linear equations above as constraints [62]:

$$
\begin{aligned}
Z^t\big(\hat{x}^t\big) = \max_{\substack{\mu^u,\mu^\ell,\nu,\\ s,\kappa}} \quad & \sum_{e\in\mathcal{E}\setminus\mathcal{E}_s^-} \left[\mu_e^u \left(\sum_{k\in\mathcal{V}^s} \big(1 - M\left(\hat{x}^t_{(i,j)k} - 1\right)\big)\right)\right] && \text{(3.8a)}\\
& + \sum_{e\in\mathcal{E}\setminus\mathcal{E}_s^-} \left[\mu_e^\ell \left(\sum_{k\in\mathcal{V}^s} \big(1 - M\left(1 - \hat{x}^t_{(i,j)k}\right)\big)\right)\right] \\
& + \sum_{(i,j)\in\mathcal{E}} \sum_{k\in\mathcal{V}^s} \big(-n_{jk} - M\left(1 - \hat{x}^t_{(i,j)k}\right)\big)\, \nu_{(i,j)k} \\
& + \sum_{i\in\mathcal{V}^p} s_i - \sum_{i\in\mathcal{V}} \sum_{k\in\mathcal{V}^s} C_k \kappa_{ik} \\
\text{s.t.} \quad & \sum_{e\in\mathcal{E}_i^+} \big(\mu_e^u - \mu_e^\ell\big) + \sum_{e\in\mathcal{E}_i^-} \big(\mu_e^\ell - \mu_e^u\big) + s_i = 0, \quad \forall i \in \mathcal{V}^p, && \text{(3.8b)}\\
& \sum_{e\in\mathcal{E}_i^+} \big(\mu_e^u - \mu_e^\ell\big) + \sum_{e\in\mathcal{E}_i^-} \big(\mu_e^\ell - \mu_e^u\big) + s_{i-R} = 0, \forall i \in \mathcal{V}^d, && \text{(3.8c)}\\
& \sum_{e\in\mathcal{E}_i^-} \nu_{ek} - \sum_{e\in\mathcal{E}_i^+\setminus\mathcal{E}_{\mathcal{V}0}^-} \nu_{ek} - \kappa_{ik} = 0, \;\; \forall i \in \mathcal{V}^p \cup \mathcal{V}^d, \forall k \in \mathcal{V}^s, && \text{(3.8d)}\\
& \sum_{e\in\mathcal{E}_{\mathcal{V}0}^-} \mu_e^\ell = 0, && \text{(3.8e)}\\
& \sum_{e\in\mathcal{E}_{\mathcal{V}0}^-} \mu_e^u = 0, && \text{(3.8f)}\\
& \mu_e^u, \mu_e^\ell, \nu_{ek}, s_i, \kappa_{ik} \geq 0
\end{aligned}
$$

If there exist any feasible solutions to this dual problem, then the corresponding primal problem is infeasible and the set of constraints in Problem (3.7) corresponding to the dual variables which take nonzero values for an extreme ray of the above LP are a minimal infeasible system. We find an extreme ray by adding a constraint that fixes the value of the objective function, Equation (3.8a), to 1, and finding the optimal solution to the resulting problem. The dual variables μ_e^u and μ_e^ℓ correspond to Equations (3.7a) and (3.7b) respectively, the s_i correspond to Equation (3.7c), the ν_{ek} correspond to Equation (3.7d), and the κ_{ik} to Equation (3.7e).

The minimal infeasible system of constraints found in this manner allows us to add a feasibility cut to the upper level problem. Because each equation in Problem (3.7) includes at most one decision variable from the upper level problem, all of which are binary variables, the MIS dictates a (minimal) set of upper level variables which have taken values in the current candidate solution that result in an infeasible solution. Therefore, in any feasible solution, at least one of those decision variables must take the opposite binary value as in the current candidate solution.

Defining $\mathbb{I}^*(m)$ to take a value of 1 if variable m has a nonzero value in the optimal solution to Problem (3.8) and 0 otherwise, we can write the combinatorial feasibility cuts added to the upper level problem as follows:

$$\begin{aligned}\sum_{e\in\mathcal{E}}\left[\mathbb{I}^*\big(\hat{\mu}_e^{u,t}\big)\left(1-\sum_{k\in\mathcal{V}^s}x_{ek}\right)+\mathbb{I}^*\big(\hat{\mu}_e^{\ell,t}\big)\left(1-\sum_{k\in\mathcal{V}^s}x_{ek}\right)\right]&\\ +\sum_{e\in\mathcal{E}}\sum_{k\in\mathcal{V}^s}\mathbb{I}^*\big(\hat{\nu}_{ek}^t\big)(1-x_{ek})\geq 1&\end{aligned}\tag{3.9}$$

This simplified constraint takes advantage of the structure of the problem and the cases of infeasibility to eliminate some terms. The first reduction is made through observing that the dual decision variables will take nonzero values only if they correspond to x_{ek} with a value of 1 in the current candidate solution (or if there is at least one vehicle k such that $x_{ek}=1$ in the case of the μ_e. This is because the big-M formulation of the constraints ensures that they will always be satisfied if $x_{ek}=0$. We therefore know that requiring one of these x_{ek} to take the opposite value can be written as a constraint of the form $\sum\limits_{e,k}(1-x_{ek})\geq 1$.

We also take advantage of the fact that if a solution is infeasible, it is impossible to make a feasible solution by keeping all $x_{ek}=1$ corresponding to constraints in the MIS fixed and changing the value of a different x_{ek} from 0 to 1. For the case where a solution is infeasible because a passenger is dropped off before being picked up, this is self-evident. For the case where a solution is infeasible because it violates a capacity constraint, this follows from the observation that the MIS will include all constraints corresponding to the edges incident on pickup nodes traversed by a vehicle prior to arriving at the node where the capacity constraint violation occurs.

The solution procedure using the CB method is altered as follows. Whenever an integral candidate solution $\hat{x}^t$ is found to the upper-level problem, we formulate and attempt to solve $Z^t(\hat{x}^t)$. If a feasible solution is found, a feasibility cut of the form found in Equation (3.9) is added to the upper-level problem. If the problem is infeasible, we then use $\hat{x}^t$ to solve the stochastic subproblems and add optimality cuts to the upper-level problem from the resulting solutions as before.

This separation of the logical constraints from the upper-level problem allows for easier identification of integral candidate solutions due to a tighter LP relaxation. The solution procedure using the combinatorial subproblem to check feasibility before generating optimality cuts is guaranteed to converge to the optimality, and both the combinatorial subproblem and the constraints generated by it forgo any big-M type formulation, reducing the number of these types of constraints significantly.

3.4.2 Heuristic Objective Bound in Upper-Level Problem

If the objective weights placed on passenger delay are high relative to the other components of the objective function, this solution method may have a poor understanding of the quality of the integral candidate solutions it passes to the lower-level subproblems before many optimality cuts are generated by these subproblems. For a problem with a high dimensional search space, this could result in a prolonged search process where the upper-level problem identifies many integral candidates for which it estimates 0 objective contribution from the lower-level problems only for computational effort to be expended generating optimality cuts that will eliminate them from consideration. This next section describes the approach taken to remedy this issue by adding heuristics to the upper-level problem that use the problem structure to place lower bounds on the objective contribution from each subproblem. These lower bounds thus reduce the search space of high-quality candidate solutions without expending computational effort on subproblem solution iterations, although it does come at a tradeoff of increasing the complexity of the upper-level problem.

To construct this lower bound, we look at what can be inferred about delay using only the decision variables and data available to the upper level problem. Because the CB approach described in the previous section limits which variables are included in the upper-level problem, this will require a separate formulation for the basic upper-level problem and for the version of the upper-level problem with many constraints removed for the CB formulation.

Because the delay contribution η_i^s of a specific pickup or drop-off location i in subproblem s is equal to the greater of either $\tau_i^s - \ell_i$ and 0 (where τ_i^s is the arrival time at location i in subproblem s of whichever vehicle k is assigned to this request, and ℓ_i is the associated time constraint), we can see that $z_i^s \geq \tau_i^s - \ell_i$.

If we consider $\mathcal{E}_k^{i-}$ to be the set of edges traversed by vehicle k prior to its arrival at location $i \in \mathcal{V}^p \cup \mathcal{V}^d$, we can write the arrival time of vehicle k at location i in travel time scenario s as:

$$\tau_{ik}^s = \sum_{e \in \mathcal{E}_k^{i-}} \tilde{t}_e^s \tag{3.10}$$

We can then consider a lower bound on the cumulative delay over vehicle k's entire route informed by the above equation, using $\mathcal{V}_k$ to represent the locations visited by vehicle k $\left(\mathcal{V}_k = \left\{i \in \mathcal{V}^p \cup \mathcal{V}^d | \sum_{e \in \mathcal{E}_i^-} x_{ek} \geq 1\right\}\right)$:

$$\sum_{i \in \mathcal{V}_k} \eta_i^s \geq \sum_{i \in \mathcal{V}_k} \tau_{ik}^s - \sum_{i \in \mathcal{V}_k} \ell_i = \sum_{i \in \mathcal{V}_k} \sum_{e \in \mathcal{E}_k^{i-}} \tilde{t}_e^s - \sum_{i \in \mathcal{V}_k} \ell_i \tag{3.11}$$

Thinking about the sum $\sum_{i \in \mathcal{V}_k} \sum_{e \in \mathcal{E}_k^{i-}} \tilde{t}_e^s$, the set $\mathcal{E}_k^{i-}$ (the edges already traversed by vehicle k when it arrives at location i) adds a single edge at each location $i \in \mathcal{V}_k$: the edge used to travel from the previous location along that vehicle's route to the current one. For instance, if $\mathcal{V}_k = \{o_1, o_2, d_2, d_1\}$, then $\mathcal{E}_k^{o_1-} = \{(k, o_1)\}$, $\mathcal{E}_k^{o_2-} = \{(k, o_1), (o_1, o_2)\}$, $\mathcal{E}_k^{d_2-} = \{(k, o_1), (o_1, o_2), (o_2, d_2)\}$, and $\mathcal{E}_k^{d_1-} = \{(k, o_1), (o_1, o_2), (o_2, d_2), (d_2, d_1)\}$.

In this example, we can see that the edge (k, o_1) appears four times, the edge (o_1, o_2) appears three times, and so on. Thus this sum is a linear combination of the travel times along each edge that vehicle k traverses, with the weight associated with each edge equal to the number of nodes left to visit after traversing that edge. If we label these weights δ_{ek} (and set $\delta_{ek} = 0$ for any edge e that vehicle k does not traverse) we thus obtain:

$$\sum_{i \in \mathcal{V}_k} \eta_i^s \geq \sum_{e \in \mathcal{E}} \delta_{ek} \tilde{t}_e^s - \sum_{i \in \mathcal{V}_k} \ell_i \tag{3.12}$$

And using the fact that each request is assigned to exactly one vehicle, we can then provide a bound on the total delay for all requests under a given assignment:

$$\sum_{i \in \mathcal{V}^p \cup \mathcal{V}^d} \eta_i^s \geq \sum_{k \in \mathcal{V}^s} \sum_{e \in \mathcal{E}} \delta_{ek} \tilde{t}_e^s - \sum_{i \in \mathcal{V}^p \cup \mathcal{V}^d} \ell_i \tag{3.13}$$

However, the objective function as presented in Problem (3.1a) includes separate weights for the delay resulting at pickup locations vs. drop-off locations. We therefore need to find values of the introduced δ variables which satisfy the following property:

$$\gamma_3 \sum_{i \in \mathcal{V}^p} \eta_i^s + \gamma_4 \sum_{i \in \mathcal{V}^d} \eta_i^s \geq \sum_{k \in \mathcal{V}^s} \sum_{e \in \mathcal{E}} \delta_{ek} \tilde{t}_e^s - \sum_{i \in \mathcal{V}^p \cup \mathcal{V}^d} \ell_i \tag{3.14}$$

The question is then how to introduce these multiplier variables δ_{ek} to the problem in such a way that they take desired values so that the right-hand side of Equation (3.14) is equal to the left-hand side when all locations in the network face nonzero delay.

For the case where $\gamma_3 = \gamma_4$, the integer variables P_{ik} present in Problem (3.1) when the CB decomposition is not employed, scaled by a factor of γ_3, are equal to exactly the values of the multipliers we need. For an edge (i, j), the value of P_{jk} for whichever vehicle k travels along that edge in a feasible solution is equal to the number of pickup and drop-off location vertices downstream of that edge that the vehicle will travel to along its route. For the case where $\gamma_3 \neq \gamma_4$, as long as $\gamma_3 > 0$ and $\gamma_4 > 0$, we can reformulate Constraints (3.1f) and (3.1g) such that P_{ik} takes the value of the total objective weight of the delay at all locations downstream of location i on vehicle k's route, rather than its numerical order within the route. In this case, we replace these equations in the original problem with the following:

$$
\begin{aligned}
P_{jk} - P_{ik} \leq& \gamma_3 - M\left(x_{(i,j)k} - 1\right) && \forall j \in \mathcal{V}^p, \forall\,(i,j) \in \mathcal{E}_j^-, && \text{(3.15a)}\\
P_{jk} - P_{ik} \leq& \gamma_3 - M\left(1 - x_{(i,j)k}\right) && \forall j \in \mathcal{V}^p, \forall\,(i,j) \in \mathcal{E}_j^-, && \text{(3.15b)}\\
P_{jk} - P_{ik} \leq& \gamma_4 - M\left(x_{(i,j)k} - 1\right) && \forall j \in \mathcal{V}^d, \forall\,(i,j) \in \mathcal{E}_j^-, && \text{(3.15c)}\\
P_{jk} - P_{ik} \leq& \gamma_4 - M\left(1 - x_{(i,j)k}\right) && \forall j \in \mathcal{V}^d, \forall\,(i,j) \in \mathcal{E}_j^- && \text{(3.15d)}
\end{aligned}
$$

These constraints still serve the function of ensuring that a pickup location must come first on a vehicle's route before the corresponding drop-off location in conjunction with Constraint (3.1h). Additionally, using their values as the multiplier for the travel time of the incident link reflects the total objective contribution of that link's travel time to the objective function in terms of delay. Beyond these constraints, we additionally add a constraint to the upper-level problem to ensure that δ_{ek} takes the value of the relevant P_{jk}, and that it takes the value of 0 if vehicle k does not traverse edge e:

$$
\delta_{(i,j)k} \geq P_{jk} - M\left(1 - x_{(i,j)k}\right) \quad \forall\,(i,j) \in \mathcal{E}, \forall k \in \mathcal{V}^s \tag{3.16}
$$

Finally, we add a set of constraints which provide a lower bound on θ_s, which is the objective value contribution from the lower-level problem representing scenario s.

$$
\theta_s \geq \sum_{k \in \mathcal{V}^s} \sum_{e \in \mathcal{E}} \delta_{ek} \tilde{t}_e^s - \sum_{i \in \mathcal{V}^p \cup \mathcal{V}^d} \ell_i \quad \forall s \in \{1, ..., S\} \tag{3.17}
$$

This bound will be exact when $\tau_i^s \geq \ell_i \; \forall i \in \mathcal{V}^o \cup \mathcal{V}^d$. In this case, every location faces delay beyond its minimum travel time, so $z_i^s = \tau_i^s - \ell_i \geq 0$. For cases where vehicles are able to arrive at some locations before their time constraints elapse and these locations face no delay, this constraint will be an underestimate of the delay as the full values of the time constraints at these locations are subtracted by the final term even when they are greater than the vehicle arrival times at those locations in the first term. However, this bound serves to shrink the feasible region of the problem by eliminating low-quality solutions without first having to expend computational effort to generate optimality cuts which rule them out.

For the CB formulation, the combinatorial subproblem removes the P_{ik} variables from the upper-level problem, meaning the approach described above won't apply. With access only to the vehicle flow decision variables that are relevant to computing delay, we therefore adopt a looser bound than the one in Equation (3.17):

$$\theta_s \geq \sum_{k \in \mathcal{V}^s} \sum_{e \in \mathcal{E}} x_{ek} \tilde{t}_e^s - \sum_{(i,j) \in \mathcal{V}^0} x_{(i,j)k} \ell_i \quad \forall s \in \{1, ..., S\} \tag{3.18}$$

Without knowledge of what order each vehicle k uses to visit the locations in its route, the only certainty about the resulting delay from that route is that the vehicle's arrival time at the final location visited on its route will be equal to the sum of all travel times on the edges in its route. We use this fact to construct the bound shown in Equation (3.18). The final term sums the time bounds of all vertices from which an edge is used for a vehicle to flow from that vertex to the sink node, where all vehicles' routes end. The first term is an unweighted sum of the total travel time of all traversed links in the network.

There are many situations in practice in which the CB heuristic bound may yield a very inaccurate result. For instance, a vehicle which arrives late to all locations along its route save the final one, but arrives early to the final location, will have a negative contribution to the right-hand side of Equation (3.18). Although the total reduction of the feasible space is much smaller, this once again serves as a computationally inexpensive way to constrain the feasible region without expending computational effort on subproblem solutions. The final section of this chapter will investigate experimentally how the CB formulation performs computationally given this tradeoff between a more tractable upper-level problem and a worse heuristic objective bound present in that problem.

3.4.3 Additional Optimality Cuts

When using a single integral candidate solution to the upper-level problem to solve the stochastic subproblems in the original formulation, we generate a single optimality cut for each subproblem. This optimality cut places a lower bound on the weighted delay that is contributed to the objective function by the travel time scenario associated with that subproblem as a function of the vehicle-edge flow decision variables.

However, we can harness knowledge of the problem structure to place additional lower bounds on the objective function contribution that are functions of different sets of decision variables besides those associated with the candidate solution that was explored. These additional bounds can be used as optimality cuts that are added to the upper-level problem in addition to the traditional cut generated by Equation (3.6), potentially reducing the number of candidate solutions that need to be explored.

In this section, we describe three types of additional optimality cuts that can be generated from the solution to one of the lower-level stochastic subproblems and used to refine the feasible region of the upper-level problem. For these methods, we use the notation $\hat{x}^t_{ek}$ to refer to the optimal values of x in the t-th iteration of the upper-level problem, and $\hat{pi}^{s,t}_{ek}$ to refer to the optimal values of pi in the t-th iteration of the s-th lower-level dual subproblem, as described in Problem (3.5).

Vehicle Swap Cuts

The first cut generation method is to swap the assigned routing between two vehicles which would each take longer to serve the other's routing than their own under the current candidate solution. The optimal solution to the primal problem given a solution from the upper level problem $\hat{x}^t$ will give the arrival times of vehicles at each node i and the resulting delay if that arrival time exceeds the node's time constraint ℓ_i. If a vehicle k is assigned to travel from its starting location k to location i (which must be an element of $\mathcal{V}^p$ because vehicles will never travel directly to a dropoff location), and $z_i > 0$, it follows that if any other vehicle k' were assigned to travel directly to i from its starting location k', and $\widetilde{t}^s_{(k',i)} \geq \widetilde{t}^s_{(k,i)}$, meaning that it takes longer for vehicle k' to travel to vertex i than it does vehicle k, then the resulting delay z_i' would be at least as great as the observed delay under the current candidate assignment ($z_i' \geq z_i$). We also can see that if the vehicle k' were given the same subsequent route from vertex i that the downstream delays would also be greater than or equal to the delays observed in this current assignment.

In the dual stochastic subproblem, described in Problem (3.5), each dual variable π_{ek} relates to a specific edge and vehicle traveling on that edge, and $\hat{\pi}_{ek}^{s,t} > 0$ if $\hat{x}_{ek}^{t} = 1$ and the vehicle that travels along that edge serves at least one location which faces a delay beyond its time constraint downstream of that edge. In fact, the numerical value of $\hat{\pi}_{ek}^{s,t}$ will correspond to the total objective weight of all locations downstream of this edge that the vehicle serves which experience delay, with pickup locations having a weight γ_3 and drop-off locations γ_4.

From this, we can evaluate the delay resulting from an alternative assignment w which swaps the routes of vehicles k and k' – that is, $\hat{x}_{ek'}^{w} = \hat{x}_{ek}^{t}$ and $\hat{x}_{ek}^{w} = \hat{x}_{ek'}^{t}$. We are interested in placing bounds on $Q_s^w(\hat{x}^w)$, that is the optimal value of the total objective contribution for subproblem s as a result this assignment, using only knowledge of the current solution to the current assignment $\hat{x}^t$ and subproblem variables $\hat{\pi}^{s,t}$. This allows us to gain some knowledge regarding this alternative assignment without needing to dedicate computational time to solving it directly.

For an edge (k, i) where $\hat{\pi}_{(k,i)k}^{s,t} > 0$, we can determine that $\hat{x}_{(k,i)k}^{t} = 1$ and vehicle k will serve at least one request that experiences nonzero delay for either its pickup, drop-off, or both. Then, we also know that if $\hat{x}_{(k',i)k}^{w} = 1$, $\hat{\pi}_{(k',i)k'}^{s,w} \geq \hat{\pi}_{(k,i)k}^{s,t}$ is guaranteed to hold as long as $\tilde{t}_{(k',i)}^{s} \geq \tilde{t}_{(k,i)}^{s}$ because any downstream delay for locations served by vehicle k would be at least as great as in the current candidate solution. Under the alternative assignment, it would also be the case that for any downstream $\hat{\pi}_{ek}^{s,w} > 0$ it would also hold that $\hat{\pi}_{ek'}^{s,w} \geq \hat{\pi}_{ek}^{s,t}$.

The inverse of each statement made above also holds in the case where $\tilde{t}_{(k,j)}^{s} \geq \tilde{t}_{(k',j)}^{s}$, where j is a pickup node such that $\hat{x}_{(k',j)}^{t} > 0$, or in other words the node to which vehicle k' is traveling directly from its starting location. We also note that in the case where $\hat{x}_{(k',\mathcal{V}^0)k'}^{t} = 1$, then it is guaranteed that $\tilde{t}_{(k,\mathcal{V}^0)}^{s} = \tilde{t}_{(k,j)}^{s} \geq \tilde{t}_{(k',j)}^{s} = \tilde{t}_{(k',\mathcal{V}^0)}^{s}$, as both $\tilde{t}^s_{(k,\mathcal{V}^0)}$ and $\tilde{t}^s_{(k',\mathcal{V}^0)}$ are equal to 0 by definition.

The procedure for generating these cuts is formalized as follows:

Given an optimal solution to the upper level problem $\hat{x}^t$, and given a positive value of the optimal dual solution for a lower-level problem (i.e. $Q^t_s(\hat{x}^t) > 0$), we take the optimal values of the primal dual variables $\hat{\pi}^{s,t}, \hat{\lambda}^{s,t}$. For any edge $e \in \left\{(k^*, i) \in \mathcal{E} | k^* \in \mathcal{V}^s, i \in \mathcal{V}^p, \hat{\pi}^{s,t}_{(k^*,i)k^*>0}\right\}$ we identify the set of all other edges with a vehicle node as a source and the same node as a target, which we label $\mathcal{E}'_{(k^*,i,)} := \{(k', i) \in \mathcal{E} | k' \in \mathcal{V}^s \backslash k^*\}$. Then for each $(k', i) \in \mathcal{E}'_{(k^*,i)}$, we find the vertex downstream of the starting location of vehicle k' which we label $j \in \mathcal{V}^p \cup \mathcal{V}^0$. If $\tilde{t}^s_{(k',i)} \geq \tilde{t}^s_{(k^*,i)}$ and $\tilde{t}^s_{(k^*,j)} \geq \tilde{t}^s_{(k',j)}$, then we add a supplemental optimality cut:

$$\begin{aligned} \theta_s \geq & \sum_{e \in \mathcal{E}} \hat{\pi}^{s,t}_{ek^*} \left(\tilde{t}^s_e - M\left(1 - x_{ek'}\right)\right) + \sum_{e \in \mathcal{E}} \pi^{\hat{s},t}_{ek'} \left(\tilde{t}^s_e - M\left(1 - x_{ek^*}\right)\right) \\ & + \sum_{e \in \mathcal{E}} \sum_{k \in \mathcal{V}^s \backslash \{k^*, k'\}} \hat{\pi}^{s,t}_{ek} \left(\tilde{t}^s_e - M\left(1 - x_{ek}\right)\right) - \sum_{i \in \mathcal{V}^p \cup \mathcal{V}^d} \ell_i \hat{\lambda}^{s,t}_i \end{aligned} \tag{3.19}$$

Location Swap Cuts

Another small alteration of the route that can lead to an additional optimality cut is swapping the order of two locations within one vehicle's route. For two locations to be able to be swapped, they must be associated with different requests, as a vehicle can't travel from a request's drop-off location to that same request's pick-up location without violating Constraint (3.1i). Let us assume we have two stop locations i and j that will be swapped within the vehicle's route, and a location v_{start} preceding i in the route and a location v_{end} following j. Then if the travel time directly from v_{start} to j is at least as large as the sum of the travel times from v_{start} to i and from i to j, we know that the delay resulting from a route that swaps i and j will be at least as large as the delay resulting from the current route. This could be expected to occur when i is along the shortest route from v_{start} to j, as then the two will be equivalent.

For each vehicle k, we use that vehicle's route to define two sets of edges containing the edges used in the swap, and one set containing the remainder of the edges in the network:

$$
\begin{aligned}
\mathcal{E}^k_{\text{base}} &:= \left\{ \left(v_{\text{start}}, i\right), \left(i, j\right), \left(j, v_{\text{end}}\right) \right\} = e_1, e_2, e_3 \\
\mathcal{E}^k_{\text{swap}} &:= \left\{ \left(v_{\text{start}}, j\right), \left(j, i\right), \left(i, v_{\text{end}}\right) \right\} = e'_1, e'_2, e'_3 \\
\mathcal{E}^k_{\text{rem}} &:= \mathcal{E} \setminus \left\{ \mathcal{E}_{\text{base}} \cup \mathcal{E}_{\text{swap}} \right\}
\end{aligned}
$$

Then, if $\tilde{t}^s_{e'_1} \geq \tilde{t}^s_{e_1} + \tilde{t}^s_{e_2}$, we can add the following supplemental optimality cut which places a lower bound on the total delay that would result from the resulting assignment with the swapped route:

$$
\begin{aligned}
\theta_s \geq & \sum_{e \in \mathcal{E}^k_{\text{rem}}} \sum_{k \in \mathcal{V}^s} \hat{\pi}^{s,t}_{ek} \left(\tilde{t}^s_e - M \left(1 - x_{ek}\right)\right) + \sum_{z=1}^{3} \hat{\pi}^{s,t}_{e_z k} \left(\tilde{t}^s_{e'_z} - M \left(1 - x_{e'_z k}\right)\right) \\
& - \sum_{i \in \mathcal{V}^p \cup \mathcal{V}^d} \ell_i \hat{\lambda}^s_i
\end{aligned}
\tag{3.20}
$$

Vehicle Exclusion Cuts

Finally, we can additionally generate constraints for any subset of vehicles in the network such that those constraints reflect the lower bound on the delay if only those vehicles are considered. These types of cuts can be useful because of the big-M optimality cuts added to the upper-level problem, which mean that changing only some of the x_{ek} from 1 to 0 but leaving others the same will result in a cut of the form $\theta_s \geq -M$, which leads to the search space needing to evaluate many similar assignments. If we consider a subset of vehicles $\mathcal{L} \subset \mathcal{V}^s$, and the set of locations visited by those vehicles $\mathcal{V}_{\mathcal{L}} = \{v \in \mathcal{V} | \exists e \in \mathcal{E}^-_v, \exists k \in \mathcal{L} : x_{ek} = 1\}$, then we can formulate a constraint:

$$
\theta_s \geq \sum_{e \in \mathcal{E}} \sum_{k \in \mathcal{L}} \hat{\pi}^{s,t}_{ek} \left(\tilde{t}^s_e - M \left(1 - x_{ek}\right)\right) - \sum_{i \in \mathcal{V}_{\mathcal{L}}} \ell_i \hat{\lambda}^{s,t}_i \tag{3.21}
$$

This constraint effectively dictates that the total delay for the system will be at least as great as the total delay incurred by all requests served by vehicles in $\mathcal{L}$. This will always hold for any subset $\mathcal{L}$ because delay is nonnegative, so adding additional vehicles to that subset can only increase the right-hand side of the constraint. However, adding constraints for all subsets of vehicles could add up to $2^{|\mathcal{V}^s|} - 1$ constraints which for large networks is impractical. To reduce the number of constraints added in each iteration, the experimental results that follow will use only a select group of subsets of $\mathcal{V}^s$ to generate these cuts. Those are:

- Subsets of $\mathcal{V}^s$ that include exactly one vehicle ($|\mathcal{L}| = 1$).
- Subsets of $\mathcal{V}^s$ that exclude exactly one vehicle ($|\mathcal{L}| = |\mathcal{V}^s| - 1$).

Furthermore, we only implement those subsets which generate nonzero delay, that is:

$$\exists e \in \mathcal{E}, \exists k \in \mathcal{L} : \hat{\pi}_{ek}^{s,t} > 0$$

This results in at most $2K$ optimality cuts of this form added to the upper-level problem from each stochastic subproblem solved.

Together, these three additional optimality cut formulations yield more insight into the problem structure and search space from solving any one set of stochastic subproblems. This does create a tradeoff between expending computational effort on extracting this information from the solution to each subproblem and expending more effort on finding more integral candidate solutions in the upper-level problem. To the extent that these cuts effectively eliminate poor candidates from the feasible space of the upper-level problem, these cuts have the potential to improve the performance of the solution method. In the experiments conducted in the following section, we investigate to what extent these cuts are able to improve upon the baseline formulation's performance and how it is affected by the problem scale and the number of subproblems used.

3.5 Computational Time Experiments

Armed with a variety of approaches to improving the computational efficiency of the optimization problem, we designed and executed a series of experiments to understand how these different methods perform under realistic conditions. While the stochastic optimization method described in this chapter was designed for implementation in an online dynamic ridesharing system with fixed-interval reoptimization, the computational challenges faced mean that real-time operation as currently formulated would require a prohibitive amount of computational power. To understand the scale of this challenge, we generated snapshots of system states that may be encountered by such a ridesharing system in a synthetic testbed network. These were then used as inputs to several instantiations of the stochastic optimization problem, each using different combinations of computational interventions and parameters. Each of these problems was then solved for this single snapshot, outputting an assignment of vehicles to passengers. Statistics regarding the solution time and the performance of the resulting assignment were recorded.

The array of optimization problem instances solved were selected to answer several questions. First, what is the impact of each of the computational interventions described in the previous section on the time required to solve the problem to optimality and how do different combinations of these interventions interact? Additionally, for larger problems where near-real-time solutions are infeasible, how well do inexact solution methods generate assignments that improve upon existing methods, and how do different strategies in improving the computational performance affect problems at a larger scale? Finally, how does the number of subproblems in the optimization formulation (corresponding to the number of samples drawn from the historical travel time distribution) impact the performance of these different interventions?

3.5.1 Small-Scale Problems with Exact Solutions

The first set of experiments created were small scale instantiations with only 3 vehicles and 5 requests randomly distributed throughout a 10km x 10km street grid. For these

Extensions Used			**Number of Samples**				
Combinatorial Benders (CB)	Heuristic Obj. Bound	Additional Opt. Cuts	5	10	25	50	100
No	Yes	No	28.8	42.1	59.8	82.8	194.7
No	No	Yes	90.3	170.8	389.3	770.9	1606.2
No	Yes	Yes	27.5	38.3	67.5	116.8	245.4
Yes	Yes	No	160.9	284.6	672.9	1374.0	2655.2
Yes	Yes	Yes	113.5	205.8	559.3	1267.8	2512.7

Table 3.1: Average solution time (in seconds) using different optimization extensions across different numbers of travel time samples used in the stochastic optimization, for 3 vehicles and 5 requests.

small-scale problems, even with a large number of samples drawn, the optimization problems were solved to optimality within a comparatively short time frame. All of the computational methods tested still solve the problem to exact optimality, and so the optimal solutions for each problem instantiation were the same. We therefore compare for these problems only the time needed to reach this optimal solution and other statistics corresponding to the solution procedure.

Table 3.1 shows the average computation time necessary for several different instances of the optimization problem to reach the optimal solution over 10 runs (2 iterations with different randomly drawn travel time samples for 5 sets of initial conditions). Five different combinations of the computational interventions described previously were implemented: only adding a heuristic objective bound to the upper level problem; only adding additional optimality cuts in the stochastic subproblems; using a combination of a heuristic objective bound and additional optimality cuts; using combinatorial benders (CB) with a heuristic objective bound but no additional optimality cuts; and using CB with both a heuristic objective bound and additional optimality cuts.

The best performances were obtained by methods that used a heuristic objective bound without adding the CB decomposition on top of the existing stochastic subproblems. The addition of extra optimality cuts based on the results of the stochastic subproblems had a slight impact that differed depending on the sample size used for the sample average approximation approach. For small sample sizes (5 to 10 travel time samples drawn), using both the heuristic objective bound as well as these additional cuts saved a small amount of time (4.5% for 5 samples and 9.0% for 10). However, for larger sample sizes of 25 and larger, the addition of these extra optimality cuts hampered the computational performance compared to the pure heuristic objective method. This finding makes sense as the computational time to generate each additional set of optimality cuts is replicated for each stochastic subproblem solved, meaning that as the number of samples drawn increases it requires more computational time to add all of these optimality cuts. These findings suggest that the benefit of investing this computational effort on generating additional cuts, further exploiting the information gained from a smaller set of candidate solutions, is lower than the benefit of investing that effort into further exploration of the search space and generating a larger set of candidate solutions.

This dynamic is clarified further by Table 3.2, which shows the average number of Benders iterations needed for the solution procedure to reach optimality for each optimization method and sample size combination across each of the 10 runs. Comparing these two methods (heuristic objective bound only vs. heuristic objective bound and additional optimality cuts), we can see that the extra optimality cuts do reduce the number of subproblems that need to be solved for each sample size investigated in these experiments by between 10 and 40%. The fact that the method using only a heuristic objective bound reached an optimal solution faster even without these additional cuts indicates that the computational time of solving those additional subproblem instances was faster than the time needed to generate the additional cuts for the smaller numbers of subproblems.

Meanwhile, we find that using the CB formulation in problems of this small scale results in significantly slower computation. Comparing the results for the CB methods to those that included the logical constraints in the upper level problem, it appears there may be two main factors contributing to this decrease in performance. First, the CB formulation is unable to use the logical constraints to generate a tighter objective bound because those constraints are contained within the combinatorial subproblem rather than the upper level problem. Second, it appears that the low number of

Extensions Used			Number of Samples				
CB	Heuristic Obj. Bound	Additional Opt. Cuts	5	10	25	50	100
No	Yes	No	61.6	56.3	77.7	59.2	84.8
No	No	Yes	436.5	423.4	446.6	444.6	447.4
No	Yes	Yes	50.0	46.7	52.3	53.1	50.0
Yes	Yes	No	1302.2	1293.7	1400.8	1405.3	1414.3
Yes	Yes	Yes	376.8	354.6	357.6	391.4	398.9

Table 3.2: Average number of Benders iterations needed to reach the optimal solution with 3 vehicles and 5 requests.

constraints needed to represent the logic for a small system of only 3 vehicles and 5 requests was not prohibitive to the solver's ability to find candidate solutions, meaning that the benefit of only incorporating logic "as needed" in the search procedure was not significantly helpful in generating feasible solutions.

These factors combined meant that the branch-and-cut search used to solve the problem was unable to eliminate suboptimal solutions as quickly while also identifying several infeasible solutions as candidates. These impacts can be seen by the large number of Benders iterations needed to solve the problem to optimality, which reflects both more optimality-cut-generating iterations due to exploring suboptimal candidates as well as feasibility-cut generating iterations using the CB framework. While using both additional optimality cuts and the upper-level heuristic objective bound in the CB framework lead to optimal solutions after a fewer number of Benders iterations than a method using only the additional optimality cuts without an objective bound heuristic, the actual computational time needed to solve these iterations was higher despite that.

The fast performance of problems at this scale with low sample sizes in sub-1-minute intervals indicates some potential for this method to be implemented in real-time systems. This 30 to 60 second interval could be used to collect requests as input for the next iteration of the optimization procedure. For larger fleets of vehicles with high demand, this could require decomposing the overall assignment problem into smaller distinct subregions with very small numbers of vehicles and requests. The best ways to decompose this problem and whether the benefits of incorporating travel time uncertainty explicitly into the optimization problem outweigh the costs of this decomposition are potential areas for further research in the future.

3.5.2 Larger-Scale Problems with Time Budget

The curse of dimensionality is a well-known problem in optimization contexts, and the proposed formulation for dynamic ridesharing assignment under travel time uncertainty suffers from it as well. The previous experiments demonstrate the potential of the novel formulation for near real-time applications with extremely small problem sizes using a subset of the computational interventions devised here. However, the number of feasible assignments explode for problems at larger scale. Specifically, the number of feasible assignments for R requests and K vehicles, assuming that no vehicle is assigned more than 2 requests, is:

$$\sum_{n=R-K}^{\left\lfloor \frac{R}{2} \right\rfloor} \left[\frac{3^n}{n!} * {}_{R}P_{2n} * {}_{K}P_{R-n} \right] \tag{3.22}$$

Because of the nature of this formulation where each combination of vehicle and requests is feasible regardless of the probability of on-time arrival (which is necessary in practice in order to handle cases where the probability of on-time arrival is 0), this means the number of constraints grows extremely quickly which renders the problem very computationally time-intensive. In order to understand the potential for this methodology to handle problems of larger scale, we ran a second set of experiments with larger problem scales (5 vehicles, 8 requests) and used the optimization instances to generate good, but not optimal, solutions by introducing a bound on the allowed runtime.

Because the solution time is now a parameter passed into the optimization problem it can no longer be used to measure the performance of the solution procedures. With this time bound introduced, we focus instead on the objective value of the resulting solutions themselves. As opposed to the small-scale experiments where each problem instance was solved to optimality, we instead investigate the objective function value reached after different lengths of time using different problem formulations to see how quickly different approaches are able to find good solutions, even if they are not necessarily optimal.

To evaluate how good these solutions are, we use as a benchmark the method described by [7] for online dynamic ridesharing using identical inputs as our novel formulation, with the exception of using the median travel time on each link when evaluating shortest paths between locations in the network. We then compute the average delay faced by passengers under an assignment produced by a solution to the novel formulation and compare it to that produced the benchmark solutions. This average is computed over the full range of historical travel times used to generate the travel time samples used as input to the stochastic formulation and the median travel times used as inputs for the benchmark. If our new method can reduce the average delay faced by passengers of the system even when given a limited computational budget, that indicates strong potential for improvement upon current dynamic ridesharing methods by incorporating travel time uncertainty information into the decision-making process.

In addition to practical questions regarding the scalability of this formulation, the larger scale experiments are also useful in understanding how the different computational interventions scale. Compared to a small problem with fewer constraints and a lower-dimension feasible region, a larger problem may face much greater costs in solving the stochastic subproblems in each Benders iteration and in finding candidate solutions through branch-and-cut processes on the LP relaxation. This may mean that approaches like creating additional optimality cuts with each pass through the stochastic subproblems or moving the logical constraints out of the upper stage problem and into a combinatorial subproblem may yield improvements, even though those approaches did not appear to yield significant benefits in the smaller scale problem.

Table 3.3 shows how different solution methods with different numbers of travel time samples used and different time constraints perform compared to the benchmark solutions. For only 10 travel time samples drawn, the two methods tested that do not use any additional optimality cuts added during the stochastic subproblem solution

Extensions Used			**Number of Samples and Time Limit**			
CB	Heuristic Obj. Bound	Additional Opt. Cuts	10 20 mins.	10 1 hr.	100 20 mins.	100 1 hr.
No	Yes	No	0.7%	16.7%	-59.2%	4.9%
No	Yes	Yes	-12.7%	16.9%	-123.3%	-31.6%
Yes	Yes	No	4.0%	19.2%	-40.2%	9.8%
Yes	Yes	Yes	-17.8%	12.0%	-98.2%	-10.5%

Table 3.3: Average objective improvement compared to benchmark method [7] using different optimization extensions, sample sizes, and time budgets for scenarios with 5 vehicles and 8 requests.

phase provide small improvements on the benchmark after only 20 minutes. In particular, the method using a combinatorial Benders decomposition saw a 4% average improvement in solution quality over the benchmark after this length of time, the greatest of any method. This finding is of interest because of the poor performance observed when using CB for smaller scale problems. This contrast indicates that for larger scale problems with a larger search space and more logical constraints added to the problem, the benefit of CB is much greater in creating a LP relaxation that can be more quickly solved to integrality when compared to a problem of smaller scale.

If given one hour of computational time to work on the optimization problem, every method tested provided a substantial improvement on the benchmark despite the fact that none were able to prove that the best solution found by that time limit was the global optimum. This is a significant finding as it demonstrates that our stochastic programming formulation can identify solutions to larger scale problems that improve upon existing methods that do not account for the stochasticity of travel times, even if not solved to optimality. Even when only using a handful of travel time samples and not solved to optimality, the potential for improvement in passengers' delay in these larger scale scenarios are significant enough that our formulation can identify improvements. Although for practical purposes, our methodology may fall short of large-scale implementation due to the computational complexity issues that are not fully resolved by the improvements identified above, this highlights the need for methods that recognize the importance of travel time uncertainty in delivering

Extensions Used			**Number of Samples and Time Limit**			
CB	Heuristic Obj. Bound	Additional Opt. Cuts	10 20 mins.	10 1 hr.	100 20 mins.	100 1 hr.
No	Yes	No	275.9	293.9	154.4	182.0
No	Yes	Yes	173.3	189.7	66.9	87.1
Yes	Yes	No	1344.5 (6.3%)	3238.4 (3.1%)	242.4 (21.9%)	567.5 (13.7%)
Yes	Yes	Yes	290.9 (21.7%)	590.9 (15.1%)	108.3 (39.0%)	165.1 (31.9%)

Table 3.4: Average number of iterations of the Benders method completed within the computational budget allocated to different solution methods under different sample sizes, for scenarios with 5 vehicles and 8 requests. Methods using the CB decomposition also list, in parentheses, the percentage of these iterations in which the combinatorial subproblem is solved rather than the stochastic subproblem.

DRS passengers to their destinations on time.

With a larger sample size of 100 travel time scenarios, the increase in computational time needed to solve the stochastic subproblems associated with one iteration of the Benders method leads to a large decrease in performance relative to the benchmark. None of the methods tested were able to provide an improvement over the benchmark solution after 20 minutes of computational time, and only the two that did not include additional optimality cuts improved upon the benchmark after 1 hour. This can be demonstrated by looking at the average number of iterations solved within the computational budget allocated to each method's solution procedure, as shown in Table 3.4. The number of iterations solved by each of the methods investigated decreases substantially when 100 travel time samples are used rather than 10.

The data in Table 3.4 gives other insights into the ways in which the computational interventions tried affect the solution procedure. Looking at the methods that do not use the Combinatorial Benders decomposition, we see a relatively small increase in the number of iterations solved between the 20 minutes and 1-hour mark. For the CB formulations, however, this is not the case as even the smallest increase from 20 minutes to 1 hour among these methods (52.4% for the method using CB, heuristic objective bound, and additional optimality cuts) is greater than the largest increase among other methods. This indicates that much of this additional time given to the problems is spent searching for integral candidate solutions to the LP relaxation among the methods that don't use the CB decomposition. In this case, the CB decomposition is helpful in making those solutions less difficult to find, which results in a greater ability to explore a larger number of candidate integral solutions, thereby improving the ability to find greater improvements.

Note also that although some number of the Benders iterations for the CB method are spent solving the combinatorial subproblem rather than the stochastic subproblems, these iterations alone do not fully account for the difference in iterations solved between methods that use CB and those that don't. The data indicates that the percentage of iterations in which the combinatorial subproblem is solved always decreases for any given method when given 1 hour to find solutions rather than 20 minutes. This indicates that in this additional time allotted, the relative increase in stochastic subproblem iterations solved is at least as large as the relative increase in total iterations solved.

3.6 Conclusions

Building upon existing works which investigate the dial-a-ride problem with stochastic travel times and online, large-scale assignment algorithms for dynamic ridesharing systems that assume deterministic travel times, we formulate a two-stage stochastic programming for online reoptimization of ridesharing assignment that incorporates travel time uncertainty into the decision-making process. This formulation is the first in the space of online large-scale ridehailing assignment which aims to minimize the total passenger delay as opposed to increasing the probability of on-time arrival.

This decision is based on the knowledge gained from the previous chapter's findings that passenger trust in the on-time arrival of a ridesharing system is one of the largest determinants of travelers' willingness to use these pooled ride services, and the assumption that greater magnitude delays which can result from maximizing on-time arrival probability are worse for this trust than a greater quantity of small magnitude delays. This choice of objective also allows the formulation to use a sample average approximation approach which makes no assumptions as to the distributional forms of travel time uncertainty in the network of operation, which is a highly desirable quality given the breadth of different distributions that the literature on this topic have been found to fit best in different scenarios.

Given the computational complexity of the problem at large scales, we further define several interventions which can be used to enhance the computational performance while still obtaining exactly optimal solutions. These include Combinatorial Benders (CB) decomposition, adding a heuristic objective bound in the upper-level problem, and generating additional optimality cuts using the results of each stochastic subproblem. Each of these approaches aim to address different aspects of the computational complexity, and have differing use cases in which they may be most impactful.

We then investigate the computational performance of the formulation in differing problem scales with different combinations of these computational interventions implemented. We find that the heuristic bound is effective at small problem scales, to the point where a parallelized decomposition of a large-scale system could potentially be run at near real-time speeds. Meanwhile, the CB approach is found to be effective at larger problem scales, and we demonstrate that this method is capable of finding routings which significantly reduce average passenger delay compared to benchmark deterministic methods even when time constraints prevent solution to full optimality, although the computational performance degrades quickly as the problem increases in scale. In the following chapter, we expand upon these findings by using a variety of experimental scenarios to illustrate the performance of this novel method compared to existing benchmarks in a wide variety of operational regimes and with a more complete set of evaluative metrics.

This formulation provides valuable contributions to the literature on online ridesharing assignment problems. While the solution methods for the proposed formulation that we implement are not fully able to overcome the computational challenges inherent in dealing with large-scale systems and accounting for travel time uncertainty in a rigorous manner, this research provides the groundwork for future exploration of this problem and builds upon the body of evidence that failure to consider these aspects of the problem can lead to significantly worse performance in practice. We also identify the elements of the formulation which lead to these computational challenges, provide discussion on how they might be addressed, implement some of these methods in our solution procedure, and provide numerical results on their relative performance to each other.

Building upon this work, the first area that will be important for future research to address will be the computational efficiency of the solution method. The specific methods proposed in this work could be extended in several ways. More work could be done to identify whether specific subsets of the additional optimality cuts are the most beneficial, and the method could also be adjusted to include more logic to selectively add only additional cuts which will be the most helpful in refining the search space. These alterations could retain most of the benefits of this method while significantly reducing the computational burden, given that these cuts are added for each subproblem solved. Additionally, if a reformulation of the optimality cuts added (whether the base subproblem cut or the additional cuts) could be found which creates a tighter cut than the current "big-M" formulation of these cuts, that would have the potential to yield similar performance improvements as were obtained by the CB decomposition approach thanks to its reduction of these types of constraint from the upper-level problem.

Another potential area for further exploration is further investigation of other methods to address the issue of reliability in dynamic ridesharing systems through passenger-vehicle assignment methods. In the approach presented in this chapter, we aim to minimize the average passenger delay beyond certain time bounds. Extensions or adaptations of this method could also consider, for instance, the early arrival of passengers (which may come at costs to users, though likely smaller than those for late arrival). This could be achieved through additional constraints in the stochastic subproblems. Another direction which could be explored is additional recourse actions that can be taken, as the stochastic optimization approach that we use would be suitable for such actions. This could include explicit consideration of future changed assignments in the stochastic subproblems, or potentially even larger changes such as allowing multiple vehicles to be assigned to the same request as a way for the operator to hedge against uncertainty at costs to total distance traveled.

Additionally, as we will discuss in the subsequent chapter of this dissertation, average delay is one metric regarding reliability, which relates to passengers' willingness to use the system as discussed in the prior chapter. Other metrics of interest could be the rate of late arrivals or the magnitude of the maximal delay under any assignment. It is likely that passengers' trust in the on-time performance of a DRS system is impacted by each of these reliability metrics in different ways and to different extents, and therefore exploring methods that can be used to address performance in other areas is of interest. Future research to adapt the formulation presented here could extend the toolbox available to operators looking to optimize assignment under travel time uncertainty.

Considering the larger area of research into dynamic ridesharing assignment problems under travel time uncertainty, given the recent emergence of interest in this field there remains many unexplored areas which warrant further investigation. We discuss the tradeoff between on-time arrival probability and total delay present in these systems, but an integrated formulation, perhaps using ideas such as the Collective Requirement Violation Index (Jaillet, Qi and Sim 2016), could balance these two aspects. Additionally, the scope of this problem is presently limited to assigning a set of requests which has already entered the system following the acceptance of a trip offer presented to the traveler previously. Future investigation into how to jointly optimize this trip offer, including the associated time bounds and trip price, and the assignment of vehicles to passengers under the uncertainty of whether these offers will be accepted, would have the potential to unify much of the different areas of work on this problem and provide large benefits in terms of service reliability, thereby increasing willingness of travelers to adopt these pooled services.

With few exceptions, the existing work on on-demand ridesharing assignment makes the strong assumption of deterministic travel times, when in reality travel times are highly uncertain due to a large number of sources of variability, and this variability is likely to be correlated across space and time. This assumption allows for these methods to quickly generate “optimal” assignments for large-scale systems with high quantities of passengers and vehicles present. However, failing to account for travel time uncertainty in the assignment method leads to the risk of violating passengers’ time constraints, especially when multiple passenger trips are pooled into one vehicle. For these ridesharing systems, a lack of faith in the system to satisfy passengers’ time constraints is one of the main reasons why travelers often choose exclusive rides, which are associated with greater vehicle distance traveled, congestion, and emissions. In order for these dynamic ridesharing systems to successfully compete with exclusive trip services and lead to increasingly sustainable transportation systems, future research must continue to devise operational strategies for these systems which incorporate travel time uncertainty into decision-making and increase the reliability of the system.

Chapter 4

Simulation

4.1 Introduction

The two-stage stochastic programming approach to dynamic ridesharing assignment under uncertainty presented in Chapter 3 aims to improve upon existing deterministic methods for solving this problem by incorporating the knowledge that travel times are stochastic and knowing the exact arrival times at request locations is impossible. Evaluating this method in a realistic testbed is important to develop an understanding of what benefits and drawbacks it offers relative to existing deterministic methods, and how these depend upon the surrounding environment of operation and choices made when implementing this method. This thorough evaluation, however, requires developing a complex framework for simulation of the system in a variety of settings motivated by empirical knowledge of how travel time uncertainty behaves in the real urban settings that dynamic ridesharing systems operate within.

While the literature regarding dynamic ridesharing assignment methods under stochastic travel times is not deep, previous studies do indicate that these methods outperform deterministic competitors in experimental settings where travel time uncertainty is incorporated [16, 52, 54]. However, these studies evaluate this relative performance only on a limited scope of distributional forms, and many of the assumptions made in constructing these distributions are contrary to empirical evidence regarding the manner in which travel times are distributed in real-world urban networks.

Therefore, it remains an open question how these methods' performance is affected by operating within networks that obey more realistic travel time distributions, and how some of these distributional parameters affect the degree of demonstrated benefit over deterministic methods.

However, travel time distributions in urban networks are complex to represent, and there exists no consensus among the literature regarding the distributions which best model empirical travel time observations. As opposed to the assumption made by many researchers who consider the ridehailing assignment or dial-a-ride problem under travel time uncertainty that travel times on urban network links are Normally distributed and independent, research has demonstrated that link travel times are asymmetric with a heavy right tail [63] and exhibit significant correlation across space [59, 60]. Although these properties are widely accepted and critical in accurately representing travel time distributions according to realistic understanding, no one distributional form has gained widespread acceptance among the literature as a definitive representation of realistic forms. Indeed, many distributions have been proposed to fit empirical data, including the Lognormal [64], shifted Lognormal [65], Gamma [58], Compound Gamma [66], Generalized Beta [67], Stable [68], and Burr XII distributions [63]. It remains an open question as to how distributions with the important properties of correlation and heavy right tails influence the degree of any benefits that are provided by stochastic assignment methods over deterministic ones.

In addition, the formulation provided in the previous chapter includes parameters that require operator decisions in order to implement this assignment method in a real-world network. Beyond the computational considerations discussed at length in the previous chapter, which do not affect the optimality of the solution found by the method, we are also interested in how the decisions made with regards to other parameters that do influence which routings are chosen by the optimization method, namely sample size and objective weights, affect performance. Knowledge of these trends may be influential for operators deciding how to implement a stochastic assignment framework in their dynamic ridesharing system in order to maximize the benefit gained.

Existing works evaluate the performance of ridesharing assignment methods, whether deterministic or stochastic, largely using either real-world case studies or smaller-scale synthetic networks. However, none of these works consider the full range of aspects described above in formulating their experimental evaluation methods.

Many works evaluate the performance of the assignment method on real-world case studies. The most common of these is the Manhattan road network, which uses open-source data provided by the NYC Taxi and Limousine Commission [7, 16, 42, 43]. Of these works, only one models travel times as stochastic in their evaluation of the proposed assignment methods, though it models link travel times with independent Normal distributions [16]. Other methods are evaluated using road networks and travel time distributions taken from Vienna [54], Chicago [51], and Taipei [53]. However, we find no examples in the literature of research which evaluates the benefit of stochastic assignment methods across a variety of urban networks, and so it is unclear to what extent the results of these studies can be generalized to other cities, which can vary significantly in the spatial distribution of demand and travel times.

Other works use synthetic test networks to evaluate system performance under artificial conditions which can be more explicitly controlled to draw more directed and potentially generalizable insights [17, 46]. However, the parameters of these networks are often chosen in ways that fail to capture dynamics of realistic urban networks, such as uniform demand or unrealistic travel time distributions. Further, we find no examples of works which incorporate travel time correlation into their evaluation of stochastic assignment methods. We do note that some works on the most reliable path problem have evaluated their methods under correlated travel times [65, 69], although these works have not been extended into a full ridesharing assignment method as of this writing.

In this chapter, I implement a novel evaluation framework that incorporates realistic understanding of travel time distributions under a variety of network conditions intended to represent meaningful variety in real-world urban settings. This framework addresses the limitations found in the literature, allowing the research to explore how previously overlooked factors such as the level of correlation of travel time distributions in nearby links or the distribution of demand within the network impact the degree to which the stochastic assignment method improves upon the performance of a deterministic benchmark method.

In this chapter, I first describe this evaluation framework, discussing the considerations made when designing the synthetic networks which are used for evaluation of the assignment formulation proposed in the previous chapter. I then provide an overview of the configurations of the benchmark method and the stochastic assignment method implemented in the simulation environment, and the scope of experiments conducted. I then present the results of these simulation experiments and examine how the stochastic assignment method performs relative to the benchmark method across different network configurations and model implementations. I finally discuss these observations and their potential import in the practical implementation of these stochastic assignment methods, and the broader research field.

4.2 Methodology

4.2.1 Scenario Design

In order to gain a thorough understanding of the advantages offered by a stochastic assignment method for dynamic ridehailing systems under a wide variety of network conditions, we first formulate a general model for link travel time distributions that will be able to reflect idealized versions of a variety of theorized behaviors. These distributions are not aimed to be used as functions that can fit real world data, but instead reflect general principles of travel time distributions to create a realistic test bed for the operation of dynamic ridesharing systems in an environment with stochastic travel times. Using different inputs to these models can therefore be used to generate distributions that represent archetypal city structures and scenarios of interest for experimental evaluation.

We implement a network with L x H nodes arranged in a grid network, with directed links connecting any two adjacent nodes. Each node represents an intersection, and the nodes form the set of locations that requests use as their origin and/or destination locations. We represent each intersection by the coordinates $(x,\ y)$. Each link is therefore either between the intersections at (x, y) and $(x+1, y)$ or between the intersections at (x, y) and $(x, y+1)$ for some integer values of x and y. We also define a spacing parameter S, which in our test networks is fixed at the value $S = 10$. All links that connect the intersections (x, nS) and $(x+1, nS)$ for some integer values of x and n, along with those that connect the intersections (nS, y) and $(nS, y+1)$ for some integer values of y and n, are defined as arterial links. These arterial links represent higher capacity roads within the network which serve greater traffic volumes between different regions within the network. We denote the set of arterial links with $\mathcal{H}$.

Each contiguous region that is bounded by arterial links is defined as a "neighborhood," with the links within that neighborhood representing local roads with lower traffic volumes, lower capacity, and often lower speed limits than the arterial links. This division into local neighborhoods with connective arterial links allows us to create differentiation between regions of the network as well as increase the diversity of individual link travel time distributions.

Travel Time Distributions

We model the vehicle speeds on each link in the network as a multivariate Normal distribution, with a vector of means $\boldsymbol{\mu}$ and a covariance matrix $\boldsymbol{\Sigma}$. Normal distributions, as noted previously, have been demonstrated to be a poor fit for travel times in the literature [57, 58]; however, as speed is inversely proportional to travel time, representing speed with a Normal distribution generates the characteristic heavy right tail seen in the corresponding travel time distribution as observed in empirical data. Empirical studies have demonstrated that more complex distributions such as Weibull, Gamma, or Burr may provide better empirical fit [63]. However, we note that the literature generally lacks agreement on any single distribution which can best be used to represent travel times, and the smaller set of parameters for the Normal distribution

Parameter	Definition
$\bar{\boldsymbol{\mu}}$	Base mean speed for all links in the network.
$\bar{\boldsymbol{\sigma}}$	Base standard deviation of speed for all links in the network.
$\hat{\boldsymbol{h}}_{\boldsymbol{\mu}}$	Adjustment factor for mean speed on high-capacity through roads.
$\bar{\boldsymbol{h}}_{\boldsymbol{\sigma}}$	Adjustment factor for speed variance on high-capacity through roads.
$\boldsymbol{f}\left(\boldsymbol{r}\right)$	Function that relates a link's radius r from the center of the network to an additive adjustment to that edge's mean speed.
$\boldsymbol{g}\left(\boldsymbol{r}\right)$	Function that relates a link's radius r from the center of the network to an additive adjustment to that edge's speed standard deviation.
$\hat{\boldsymbol{\mu}}_{\boldsymbol{N}}$	Additive adjustment to a link's mean speed based on the neighborhood N of that link.
$\hat{\boldsymbol{\sigma}}_{\boldsymbol{N}}$	Additive adjustment to a link's speed standard deviation based on the neighborhood N of that link.
$\hat{\boldsymbol{h}}_{\boldsymbol{\phi}}$	Adjustment factor for covariance between two links' speeds depending on whether one, both, or neither of the links are high-capacity through roads.
$\boldsymbol{D}_{\max}$	Maximum distance (in terms of the number of links between them) at which two links in the network can have nonzero covariance.

Table 4.1: Definitions of the parameters used to generate travel time distributions for the simulation scenarios.

allows for a clearer interpretation of the impact of adjusting each.

We specify the multivariate Normal distribution for the link speeds in the network by assigning each link e in the road network a mean speed μ_e and a standard deviation of speed σ_e. The vector of means $\boldsymbol{\mu}$ is therefore the list of each individual link's mean speed. The covariance matrix $\boldsymbol{\Sigma}$ uses each link's speed variance σ_e^2 along the diagonal, and for off-diagonal terms we define the covariance between any two links $i,\ j$ in the network to be ϕ_{ij}.

We parameterize each link's speed mean and standard deviation as follows:

$$\mu_e = \left(\bar{\mu} + \hat{\mu}_{N(e)} + f\left(r_e\right)\right) * H_\mu\left(e\right) \tag{4.1}$$

$$\sigma_e = \left(\bar{\sigma} + \hat{\sigma}_{N(e)} + g\left(r_e\right)\right) * H_\sigma\left(e\right) \tag{4.2}$$

These similar equations identify four different parameters which govern link speeds. The first is a network baseline: $\bar{\mu}$ is the base mean speed for any link on the network, and $\hat{\sigma}$ is the base speed variance. The second is a neighborhood specific term, $\hat{\mu}_N$ or $\hat{\sigma}_N$, which applies a flat adjustment to speed or standard deviation to every link between any two nodes within each neighborhood, the degree of which can vary by neighborhood in the network. The third term is a function of the normalized radial distance of the link from the center of the network, denoted by r_e. This distance is measured by taking the Manhattan distance of the geographical midpoint of the link from the midpoint of the network grid, and dividing by the maximum such distance of any node within the network. In the case of a network where each link is a uniform length ℓ, then this divisor is equal to $\frac{L+H}{2}$. Finally, the sum of the first three terms is multiplied by an adjustment factor $H(e)$, the value of which depends on whether the edge is an arterial link or a local neighborhood link.

The three modifications to the base mean or variance each aim to capture different effects of network structure on travel time distributions. The neighborhood effect captures the influence of different regional traffic behaviors. For example, the mean speed could be influenced by differing speed limits or road widths within that region of the network. The variability could be affected by factors such as different traffic volumes due to different population or job densities within different neighborhoods, as high traffic volumes can lead to higher levels of congestion and therefore greater link speed standard deviations. The radial distance factor captures this same idea of differing traffic volumes but in a continuous manner radiating from the center of the network outwards, assuming that the area of greatest demand lies in the center of the network (therefore leading to slower average speeds and greater travel time variability) rather than neighborhood-level distributions. Finally, the arterial factor reflects the difference in traffic volume as well as capacity that are features of these larger connective roadways between neighborhoods.

The covariance between two links i and j is formulated as the product of the standard deviations of each individual link, multiplied by a correlation factor that decreases with distance:

$$\phi_{ij} = \sigma_i * \sigma_j * H\left(i, j\right)^{d+1} \tag{4.3}$$

Spatial correlation between link speeds is an important factor in realistically modeling travel time stochasticity in urban networks, as previous studies find that independently distributed link travel times are a poor fit to actual travel time data [59, 60]. The decision to base the level of covariance of two links' speeds on those individual links' standard deviations reflects the fact that the overall variance of travel time along a path consisting of multiple links is dominated by the covariances between links in the network as opposed to the individual links' variance terms as the number of links along the path increases. Therefore, incorporating the variance of the individual links into their covariance increases the degree to which the parameters in Equation (4.2) affect the path travel times, which are what is used as inputs to our stochastic assignment method.

The product of these individual link standard deviations is further adjusted by a term $H(i, j)$ which takes a value on the interval $[0, 1)$ dependent on whether link i, link j, both, or neither are arterial links. This term reflects that the travel time of distinct segments of the larger, more heavily traveled roads in a network may exhibit greater correlation with one another due to effects such as congestion queues or green signal timing which affect several links at the same time. On the other hand, the travel times of vehicles on less traveled local roads may have less correlation with one another as speeds on those roads are more a product of highly localized factors such as individual driver behavior, stop signs, and other such effects that are unlikely to impact a large area simultaneously.

This arterial adjustment is exponentiated by a factor of the distance between the two links, d, measured by the number of links traversed between the downstream node of one link and the upstream node of the other along the shortest path through the network. We further define a parameter $D_{\max}$ for the network which defines the upper limit of distance at which two links exhibit a correlation – if the shortest path distance between the two links i and j is greater than or equal to $D_{\max}$, then ϕ_{ij} is set equal to zero (which is mathematically equivalent to setting d to a very large number, as the base $H(i,j)$ is defined to be less than 1).

We define formally the arterial-level functions used as follows, to be clearer about the specific parameters that we set values to:

$$H_{\mu}(e) = \begin{cases} \bar{h}_{\mu} & e \in \mathcal{H} \\ 1 & e \notin \mathcal{H} \end{cases} \tag{4.4}$$

$$H_{\sigma}(e) = \begin{cases} \bar{h}_{\sigma} & e \in \mathcal{H} \\ 1 & e \notin \mathcal{H} \end{cases} \tag{4.5}$$

$$H(i,j) = \begin{cases} \bar{h}_{\phi}(1,1) & i \in \mathcal{H} \cap j \in \mathcal{H} \\ \bar{h}_{\phi}(0,0) & i \notin \mathcal{H} \cap j \notin \mathcal{H} \\ \bar{h}_{\phi}(1,0) & \text{otherwise} \end{cases} \tag{4.6}$$

While these parameterizations of link travel time distributions are not modeled on actual road speeds and the literature does not provide insight onto specific values to choose for each of these parameters, we do note a few findings from previous studies which inform the choices of specific values in the equations above. Based on empirical findings that correlations between travel times on road segments extends beyond immediately adjacent links to further upstream and downstream segments, and even parallel or intersecting segments [60], the value of $D_{\max}$ is kept greater than 1. We believe this to be the first instance in the literature of dynamic ridesharing simulation on networks with correlated travel times that extend beyond immediately adjacent links, which is significant given the empirical grounding that this has in real world data.

Second, we observe that the mean and standard deviation of distance-normalized travel times on network links are positively correlated, and this correlation can be accurately captured by a linear relationship between the two [66]. This indicates that for any choice of parameters, the corresponding terms between Equations (4.1) and (4.2) should be chosen such that they are inversely proportional. To give an example, if links within a specific neighborhood of the network have a greater-than-average speed, which means faster travel times, they should also have lower-than-average variability, as the reduction in travel times is directly proportional to the reduction in travel time variability.

This parametrization of the link speed distributions results in 12 parameters that can be used to generate scenarios for the experiments. 8 of these are singular values ($\bar{\mu},\ \bar{\sigma},\ \bar{h}_{\mu},\ \bar{h}_{\sigma},\ \bar{h}_{\phi}\left(1,1\right),\ \bar{h}_{\phi}\left(1,0\right),\ \bar{h}_{\phi}\left(0,0\right)$, and $D_{\max}$), two are univariate functions (f and g), and two are vectors with a number of elements equal to the number of neighborhood subregions in the network ($\hat{\boldsymbol{\mu}}$ and $\hat{\boldsymbol{\sigma}}$). Of these parameters, we can create three distinct groupings that can be used to characterize the influence that they have on the travel time distribution and what considerations to make when deciding on their values for the experimental scenarios.

1. Urban form and activity distribution: $f,\ g,\ \hat{\boldsymbol{\mu}},\ \hat{\boldsymbol{\sigma}}$

   These parameters can be used to impact how speed means and variances are distributed spatially throughout the network, either through creating a city center that behaves differently compared to the periphery or by adding neighborhood-specific influences to create spatial speed variability in a different pattern. These can be used to generate diverse network structures to understand how performance differs across several city archetypes.

2. Traffic conditions and congestion impact: $\bar{\mu},\ \bar{\sigma},\ \bar{h}_{\mu},\ \bar{h}_{\sigma}$

   These parameters control the base speed and speed variance throughout the network that are modified by the other adjustments, as well as the adjustments to speed mean and variance that occur on the arterial roads and the degree to which speeds on adjacent links of different types of roads covary. These can be used to adjust how congestion influences the network for any type of structure.

3. Parameters of interest to optimization performance: $\bar{h}_{\phi},\ D_{\max}$

   These parameters control the degree to which link speeds vary across the network, and how strongly they covary with the speeds of other nearby links. While these may relate in some sense to city structure or traffic conditions, at a much more fundamental level we are interested in how each assignment method's performance is impacted by travel time variability and the correlation of network travel times between links. Therefore, rather than using these to create different scenarios to consider, these will be varied within each scenario and used to evaluate how the performance of the assignment methods responds.

With the specification of the link travel time distributions above, we then created a distinct set of networks to use in the experimental evaluation of the stochastic assignment method against the benchmark method. In creating these networks, we had two main goals. First, we wanted to reflect distinct archetypes of theoretical urban form, as understanding whether these different urban forms influence the effectiveness of stochastic assignment is of interest to public policy practitioners and operators of dynamic online ridehailing services in understanding the benefits these methods may have in different real-world urban areas.

Second, the network configurations need to be able to result in a large variety of similar quality solutions to the assignment problem in regards to the relative locations of the vehicles, request origins, and request destinations within the network. In order for these experiments to have a high degree of specificity in identifying the benefits of the stochastic assignment method over the benchmark method, it requires that we be able to find meaningful differences between the two. If there only exists a small number of feasible routes which are able to deliver travelers to their destinations near their desired arrival times, the likelihood of the stochastic assignment method yielding an optimal route that is significantly different than that of the benchmark method is reduced. It was therefore important to seed the environment with conditions that made it likely to have a large number of plausibly good solutions to the assignment problem.

Synthetic Networks

We created two synthetic networks to use for experimental evaluation of the assignment methods: one based on the monocentric model of urban formation, and

Parameter	**Monocentric**	**Polycentric**
$\bar{\boldsymbol{\mu}}$	7.5 m/s	10 m/s
$\bar{\boldsymbol{\sigma}}$	2.0 m/s	2.0 m/s
$\hat{\boldsymbol{h}}_{\boldsymbol{\mu}}$	1.5	1.5
$\bar{\boldsymbol{h}}_{\boldsymbol{\sigma}}$	2.0	1.5
$\boldsymbol{f}\left(\boldsymbol{r}\right)$	$+\dfrac{5}{1+e^{-8(r-0.5)}}$ m/s	0
$\boldsymbol{g}\left(\boldsymbol{r}\right)$	$-r$	0
$\hat{\boldsymbol{\mu}}_{\boldsymbol{N}}$	0	High activity: -2.0 m/s Moderate activity: $+0.0$ m/s Low activity: $+2.0$ m/s
$\hat{\boldsymbol{\sigma}}_{\boldsymbol{N}}$	0	High activity: $+1.0$ m/s Moderate activity: $+0.0$ m/s Low activity: -1.0 m/s

Table 4.2: Parameter values for the link travel time distributions on the monocentric and polycentric model networks.

the other on the polycentric model. These competing models, though not precisely defined, arose from theories of urban agglomeration to explain observed trends in the densification of cities in developed economies (Kloosterman and Musterd 2001). Polycentric urban regions feature multiple concentrations of density within them, as opposed to a monocentric city which exhibits a clear contrast between a single dense center and the suburban hinterland. Examining how the performance of the stochastic assignment method compares to deterministic methods under each of these two urban models will therefore be revealing as to what real world conditions or cities will result in this methodology providing the largest benefit.

Table 4.2 shows the parameter values chosen to use for the links speed distributions of each synthetic network. The monocentric model networks have no neighborhood adjustments; instead, the differences between roads' travel times and speed variability are driven by their proximity to the center of the network where the greatest concentration of traffic is. On the other hand, the differences in the polycentric model are driven solely by the distinct neighborhoods of the region which are separated into high, moderate, and low activity with corresponding levels of congestion, with more congestion leading to lower speeds and greater variability.

Parameter	**Low Correlation**	**Medium Correlation**	**High Correlation**
$\bar{\boldsymbol{h}}_{\phi}\,(\mathbf{0,0})$	0.1	0.5	0.7
$\bar{\boldsymbol{h}}_{\phi}\,(\mathbf{1,0})$	0.4	0.7	0.85
$\bar{\boldsymbol{h}}_{\phi}\,(\mathbf{1,1})$	0.6	0.9	0.99
$\boldsymbol{D}_{\max}$	2	3	5

Table 4.3: Parameter values for each level of correlation implemented in the synthetic networks.

Beyond the parameters for the individual link travel time distributions, we also needed to specify parameters for the correlation of travel times between nearby links in the network. we specified three levels of correlation that travel times on either of the two network configurations could exhibit. These different levels were created by adjusting the parameters used in Equation (4.3) as well as the maximum correlation extent parameter $D_{\max}$, using the values displayed in Table 4.3. These scenarios can each be applied to either of the synthetic networks to generate better understanding of how the interaction of travel time correlation and urban form influence the performance of the stochastic assignment method.

In terms of the specific configuration of each of the synthetic networks implemented in the simulation, we were limited by the computational performance of the stochastic method to investigate only small scenarios with a handful of vehicles and requests. With the low dimensionality of the problem, it was therefore difficult to generate significant difference between various solutions. The approach we used to do this was to constrain the size of the network as well as where vehicle starting locations, request origins, and request destinations could be located. These had the effect of limiting the difference in distance between two vehicles and any given request or destination location, thereby increasing the number of plausible solutions in which different vehicles could be assigned to serve the same request with relatively small variation in the satisfaction quality of the request's time bounds, though it also limits the difference in quality between said solutions. This tradeoff was made as the quantity of potential solutions was more important for us in being able to detect a difference between solution methods than the absolute magnitude of that difference.

For the monocentric model, we created two test network configurations using the same speed distribution parameter values shown in Table 4.2. The first of these has a width of 60 links, each 250m long, and a height of 10 links of the same length. The through road spacing parameter was set to 10 links, leading to an arrangement of 6 distinct neighborhoods arranged linearly. The starting locations of the vehicles were constrained to within neighborhood 1 (the leftmost), request origins were located within neighborhoods 2 and 3, and each request's destination was placed in neighborhood 4 and 5. We label this network Monocentric(10,60). The linear arrangement of locations in this network is intended to maximize the number of solutions to the optimization problem which have similar performance in terms of passenger delay, as the differences between path travel times among vehicle and request combinations will be small.

The second monocentric model network has a width of 30 links and a height of 30 links, each 250m long again. The same spacing parameter for through roads is used, leading to 9 neighborhoods in the network. The starting locations of vehicles are constrained to within the four neighborhoods at the corners of the network, request origins are located within the outer eight neighborhoods, and destinations are located entirely within the central neighborhood. The arrangement of supply and demand in this network is therefore intended to reflect patterns of inbound commuting in a monocentric region, with vehicles and passengers starting in the periphery of the network and aiming to travel to the center of the network, where congestion is highest. However, this may generate scenarios where the initial locations of the vehicles are spread in such a way that only a small number of feasible routings perform well. This neighborhood is labeled as Monocentric(30,30).

For the polycentric model network, we again use a square network with a length of 30 links of 250m in each dimension, leading to 9 neighborhoods as in the Monocentric(30,30) network. The neighborhoods in the corners of the network were designated as the regions of higher activity within the network, representing clusters of businesses, entertainment, shopping, or other areas of interest in the polycentric model. Vehicles started in the central neighborhood of the network, while origins were distributed throughout the outer eight neighborhoods, and destinations located only within the four corner neighborhoods. We also mandated that a request's origin and destination could not lie within the same neighborhood, necessitating cross-network travel for the vehicles. We label this network Polycentric(30,30). The pattern of supply and demand is somewhat the inverse of the Monocentric(30,30) network, with vehicles beginning in the center of the network and serving trips which wish to arrive in the corner neighborhoods.

Figure 4-1 depicts each of these three model networks used in the simulations. To give intuition regarding the distributions of link travel times used in each, it also shows for each network one sample of link travel times using the high correlation parameters and one sample using the low correlation parameters (as shown in Table 4.3). From these images we can see clearly the differences between the monocentric travel time distributions, which feature slow speeds in the center of the network and high speeds in the periphery, and the polycentric distributions, which feature different travel time distributions within each neighborhood. We can further see the impact of the degree of travel time correlation. Networks with higher correlation of travel times not only feature a greater degree of "clumping" of links with high travel times, but also have a greater number of links with extreme travel times, both slow and fast.

While Figure 4-1 depicts only a single sample of travel time distributions, we further explore these distributions on each of the different networks by plotting the full distribution of travel times across all 1,250 samples that we use for validation of the experimental results later in this chapter. These distributions are shown in Figure 4-2. These plots validate the realism of the shape of the travel time distributions, with a heavy right tail indicating the potential for large delays. They also reflect the intuition gained regarding the spread of travel times as a function of the network travel time correlation, with travel times in the low correlation regime demonstrating tighter distributions, with fewer travel times at the extremes and a higher peak frequency near the mode of the distribution. Meanwhile, travel times on networks with higher correlation extend further in both directions, but with a reduced peak frequency.

In addition to origin and destination locations, each request is also associated

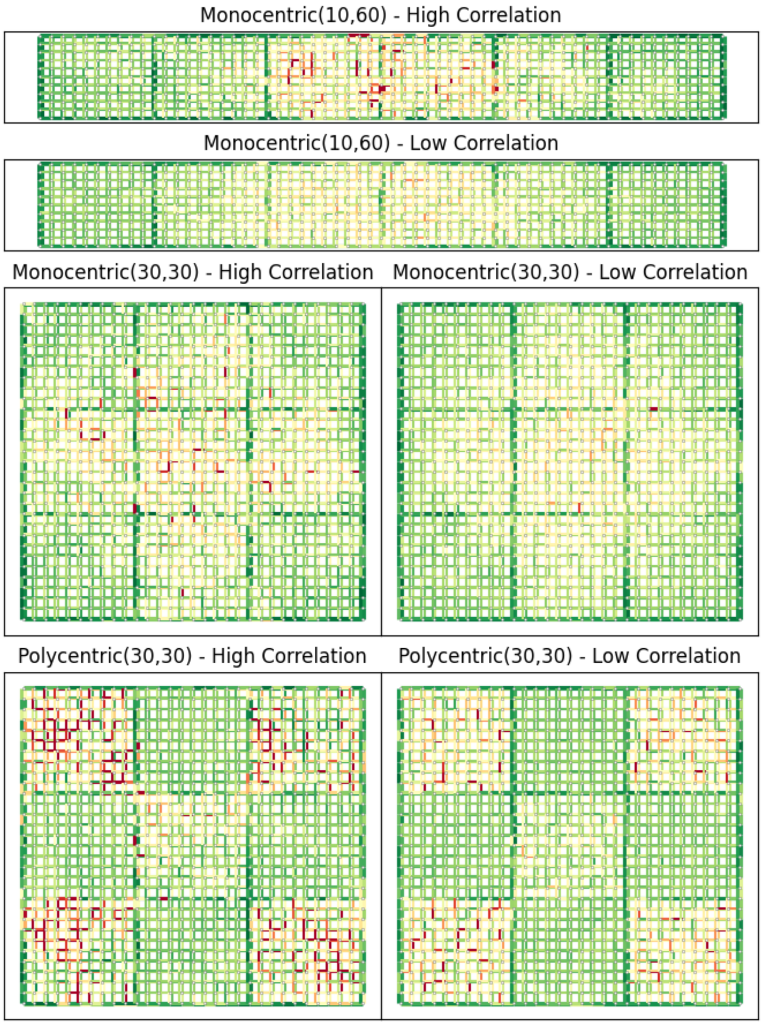

Monocentric(10,60) - High Correlation
Monocentric(10,60) - Low Correlation
Monocentric(30,30) - High Correlation
Monocentric(30,30) - Low Correlation
Polycentric(30,30) - High Correlation
Polycentric(30,30) - Low Correlation

Figure 4-1: Images depicting the network configuration of each of the three model networks used for the experimental analysis, for both high and low levels of travel time correlation. Each link in the network is colored according to the travel time on that link in a random network travel time sample drawn from the appropriate distribution, with green representing low travel times (high speeds) and red representing high travel times (low speeds). The thicker lines denote the links designated as high-capacity throughways.

with a pickup time constraint and a drop-off time constraint. Each of these time constraints represents the time after which a vehicle arriving at the corresponding location counts as delay – if the vehicle arrives at that location prior to the time constraint, it is treated as not violating the constraint and incurs no penalty in the objective function of our stochastic assignment formulation. For the requests in the experimental scenarios, we decided upon a maximum acceptable wait time for pickup

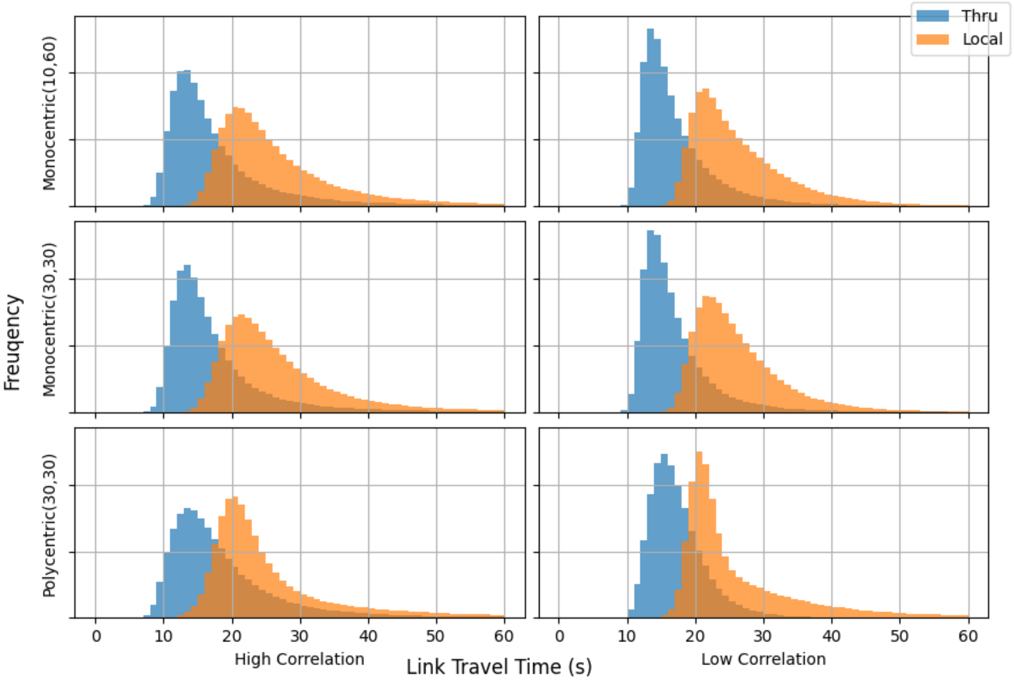


Figure 4-2: Distribution of link travel times for each network structure for low and high degrees of travel time correlation. The blue distribution shows the travel times of the higher capacity through roads, while the orange distribution shows the travel times of the smaller local streets.

of 5 minutes, and a maximum acceptable detour factor for the whole trip of 1.25. The latest acceptable drop-off was determined by multiplying this time detour factor by the median travel time along the shortest path between the request's origin and destination and adding the result to the acceptable pickup wait of 5 minutes. These time bounds are comparable to the delays allowed in the formulations of other assignment methods in the literature [7, 16]. Additionally, given the small number of vehicles in our experiments, it was impossible to satisfy all of them in many of the simulated cases, which helped generate a larger variety of plausible paths with different performance.

4.2.2 Experimental Plan

With the scenarios in place to generate initial conditions that can be used as inputs for the assignment problem, we now develop the methodology that will be used to evaluate how the stochastic assignment method proposed in the previous chapter compares to existing deterministic methods on these experimental networks. We first describe how we implement a benchmark method from the literature as a baseline for comparison. We then explain the implementation of our stochastic assignment method for these experiments. Finally, we outline the set of experiments which are run using these two methods to generate the set of experimental results.

The deterministic method we implement as a benchmark for comparison is that described by [7]. This assignment method is presently near the state of the art in terms of large-scale online assignment for dynamic ridesharing systems. Additionally, in comparison to some of the other methods discussed in the literature review of the previous chapter, it has the important property that it is capable of bundling multiple incoming requests in one iteration of the assignment process into a single trip that is served by a single vehicle, as opposed to other methods which are capable only of assigning at most one passenger to each vehicle in any given iteration [42]. This is an important property given the setup of our experiments as described later.

As the benchmark method relies on deterministic travel times in its constraints, we use the median of each link's travel time as the input for this method. We note that, given the extent of correlation between links in the network, the median path travel time is not necessarily equal to the sum of the median travel times along each link in the path, and using the median path travel time to evaluate the cost of each trip would likely improve the benchmark method's performance. However, the method implemented in the original paper makes no note of this fact and uses a Tabu search method that assumes individual link travel times can be summed into total path travel times in evaluating link costs, which we have reproduced for comparison.

The objective function used for the benchmark method is to minimize the total delay at the passengers' drop-off locations beyond the earliest possible drop-off time (assuming an instantaneous pickup), as formulated in the original paper. However, the constraints in the original paper will reject any requests which are not able to be dropped off within the desired time window. As our stochastic assignment method makes no such constraint, and indeed in many of the simulation scenarios it is impossible to satisfy all passengers' desired time windows, we instead force the benchmark method to assign each request to some vehicle by relaxing this constraint.

For the stochastic assignment method, the decisions faced in implementing it are the number of samples used to generate the stochastic subproblems (i.e., how many travel time scenarios are considered in the optimization process), and the objective weights of different components. We vary the sample size from 10 to 250 to understand how this parameter impacts performance. We expect to see that increasing sample size leads to an improvement in terms of average delay, which is the objective of the method, and potentially other metrics as well, though there may be diminishing returns to this benefit at larger sample sizes as the sample average converges towards the true distribution average.

In terms of objective weights, there are four weighting parameters which can be used to modify the objective function. γ_1 is a distance-based cost that gives weight the total travel distance of all vehicles in the network in the minimization. γ_2 is the cost of new assignments of passengers to vehicles, which differentiates between candidate solutions in terms of how many passengers that were previously assigned to a vehicle change their vehicle assignments. γ_3 is the cost of delay at passengers' pickup locations, and γ_4 is the cost of delay at passengers' drop-off locations.

As the experiments conducted involve only one iteration of the assignment method, γ_2 would not serve any role in differentiating potential solutions as there are no previously assigned passengers who could be re-assigned to new vehicles. We therefore fix that parameter to zero for most experiments, with one exception described later. We also fix $\gamma_3 = \gamma_4$ in these experiments. This reflects that passengers' trust in the system performance and their utility from the trip is influenced both by the excess wait time that they experience beyond the predicted wait time, and any delay they suffer in arriving at their final destination. As we have no a priori knowledge of how the magnitudes of these influences compare, we reduce the dimensionality of the experiment space by constraining these two to be equal to each other.

We are left with only two parameters in the objective function. We fix the delay weight $\gamma_3 = \gamma_4$ to be equal to 1 in all cases – delay is measured in units of seconds, so the units of the delay weight can be thought of as objective cost per second. We then vary the distance-based cost weight, γ_1, by setting a parameter W such that $\gamma_1 = W\gamma_3 = W\gamma_4 = W$. As we measure distance in units of meters, the units of the objective weight for total distance traveled are in objective cost per meter, which means the units of W are seconds per meter, or $(\mathrm{m/s})^{-1}$. In the simulation experiments, we use three values for W: 0 s/m, 0.05 s/m, and 0.1 s/m. The first reflects the system cost being measured solely in terms of passenger delay. The second and third indicate some amount of distance-based cost being added into the evaluation of solutions in addition to delay – at 0.05 s/m, traveling 20 extra meters to save 1 second of average delay is an equivalent tradeoff; at 0.1 s/m, the tradeoff is equivalent at 10 extra meters rather than 20. Using a value of travel time savings of \$15/hour, these values would translate to distance-based costs of \$0.21/km or \$0.42/km, respectively.

Using these implementations of the benchmark and stochastic assignment methods, we conducted a series of experiments on the array of synthetic networks generated. For each network configuration – a combination of the network structure, monocentric or polycentric, and the degree of correlation – we generated 20 system states using the demand generation and vehicle location placement procedures described earlier, with 3 vehicles and 5 requests in the system. For each of these system states, we used this as input to the benchmark method to obtain an optimal routing from that method.

For each sample size used from 10 to 250, we then created 10 samples of that size by drawing randomly from a training set of 1,250 realizations of the network speed distribution. Each of those 10 samples was used as input for the stochastic assignment method for each of the 20 system states being evaluated. This resulted in 200 optimal solutions obtained for each model specification in terms of sample size and objective weight, with 10 of those solutions corresponding to each of the 20 optimal solutions obtained from the benchmark method.

In addition to these experiments with 3 vehicles and 5 requests, we also explored situations with slightly greater numbers of actors in the system to gain a better understanding of how performance may scale to more complex situations with different balances of supply and demand. In these experiments, we used 6, 8, and 10 vehicles with 8 requests. To better reflect the iterative nature of DRS assignment in practice, these experiments choose 4 of the requests randomly that are pre-assigned to the vehicles nearest their origins. These pre-assignments are treated as fixed, achieved by setting $\gamma_2 >> \gamma_3$ in the stochastic method (the benchmark method assumes all previous assignments are fixed by default). These experiments, which were more computationally intensive due to the larger scale, were run only over 10 scenarios generated at random on the Monocentric(30,30) network, with a sample size of 50 travel time draws for the stochastic method.

4.2.3 Evaluation

Once we have obtained the solutions for the benchmark method and each implementation of the stochastic method under each scenario, we then aim to evaluate comprehensively the benefits and tradeoffs made by incorporating the stochasticity of travel times into the assignment method. We do this by comparing the performance of the resulting solutions using a validation set of 1,250 realizations of the network speed distributions which are distinct from, though generated in the same method as, the training set used to solve the problems in the original case. This avoids giving the stochastic assignment method an unrealistic advantage from being able to optimize exactly against the set of travel times that are used for evaluation.

We compare the performance of the assignments generated by each method in terms of reliability and road use. Reliability, which we investigate through the facets of average passenger delay, variance of passenger delay, and frequency of delays, affects how passengers perceive the system's ability to meet their needs for an effective and timely door-to-door on-demand transportation service. Road use, which is measured through the total vehicle distance traveled to serve all locations on the vehicles' routes, generates externalities such as congestion and emissions. Minimizing these externalities through reducing road use is one of the main purposes driving the creation of pooled services, and so evaluating how effective each method is in accomplishing this goal is of interest.

The three metrics we use to evaluate reliability (average delay, variance of delay, and frequency of delay) each approach the issue of reliability through a different perspective. Our focus in particular upon the metrics related to delay is motivated by the observation that, while getting passengers to their destination early is always a desirable goal and may help to engender good will among the users of the service, our previous findings indicate that passengers are much more reluctant to use the service if they perceive it to do a worse job at delivering them to their destination on time. However, the results presented in Chapter 2 fail to provide a clear picture on the relative importance of each of these three factors in how passengers' perceptions of reliability are formed. We therefore examine how all three of them are affected by the different solution methodologies.

Average delay is computed by taking the delay at each location in the network by computing the arrival time of the assigned vehicle at that location minus the location's associated time constraint, and averaging across all locations served (weighted, if necessary, by the relative importance of pickup and drop-off delay as described by the parameters δ_3 and δ_4 in the objective function defined in Chapter 3). An early arrival before that time constraint is treated as a delay of zero. This is then divided by the number of locations in the network. This is the exact quantity that the stochastic assignment method incorporates in its objective function.

Lower average delay means that vehicles are on average less late in picking up and dropping off passengers relative to the provided time bounds on those actions. In our experiments, these time bounds are intended to reflect the estimated pickup and drop-off times provided by the DRS service when the passenger booked a ride. In other applications, trip requests may allow passengers to define their own time bounds reflecting at what point they become "late" to their destination, or how much waiting and/or in-vehicle travel time they are willing to tolerate. Reducing average delay therefore demonstrates reduced excess waiting and journey time relative to passenger expectations, whether those expectations are provided by the system itself or communicated by the passenger to the system operator.

In addition to the average magnitude of delay, the variability of passengers' wait and journey times (which we describe collectively as arrival times, referring to the vehicles' arrivals at passenger locations) across all travel time scenarios is also an important measure of system reliability. We measure this by taking the standard deviation of the individual pickup or drop-off times associated with each passenger location in each of the 1,250 travel time scenarios in the validation set, and averaging across all passenger locations. In contrast with the average delay, which gives a sense of how late the average passenger will arrive, this measure of average arrival time variability provides information on reliability by indicating how much deviation from this average a passenger may expect to experience on their individual journey.

Reducing the variability of arrival times corresponds to an increase in the perceived dependability of the system to deliver them closer to the mean delay. A system with greater variability may have some passengers which experience wait or journey times much longer than the provided estimates, which may serve to deteriorate trust in the system more quickly, all else equal. We also note that as opposed to the other metrics we use for evaluation of reliability, namely average delay and late arrival rate, the variability of arrival times depends less on the specific time bounds set by the system. While average delay may be reduced through supply side changes outside of the scope of these experiments, such as providing more cautious time estimates, the variance of arrival times is much more difficult to change outside of altering the method in which vehicles are assigned to passengers by accounting for travel time variance.

The final reliability metric evaluated is the frequency of late arrivals, which we also call the late arrival rate. Whereas average delay measures the extent of delay experienced by passengers who do not arrive on time, late arrival rate measures how many passengers face delay. It is obtained by counting the number of locations at which the vehicles arrive later than the corresponding time bound under each travel time scenario divided by the total number of locations in the network.

Reducing the late arrival rate indicates that more passengers arrive to their destinations within the allotted time bounds. However, it does not provide any information about the magnitude of delay experienced by those passengers who do face excess wait or journey time. In a system that reduces late arrival rate while increasing the average magnitude and/or variability of delay, it will be the case that the smaller number of passengers who experience delays are facing more severe delays. Over the long run, this dynamic may engender more distrust in the system's reliability compared to the alternative due to the lower probability but higher impact nature of these delays.

Finally, we evaluate the total vehicle distance traveled to provide service in order to understand the relative road use of each assignment method. Reduced travel distance, indicating less road use, indicates that the system is more efficiently using the space dedicated to it to serve the same number of passengers. While this does not directly affect the passenger experience, unlike the reliability metric, it is of key interest to policymakers and regulators who seek to understand the externalities of DRS on the transportation system as a whole. It is also of interest to the system operator, who faces costs from road use from per-distance costs, either from fuel and vehicle depreciation or from externally provided incentives such as road pricing.

Improved performance in each of these metrics reflects their own unique advantages. Determining what to prioritize among the three reliability metrics is a complex question, as is the decision of how to weigh the passengers' interest in reliability and the system-level objective of road use. For the system operator, both are important at different scales, as the per-distance costs that can be reduced by minimizing road use are immediate while the long-term costs associated with passengers' perceptions of reliability are harder to quantify and become relevant over time through repeated interaction with the system. Similarly, policymakers and regulators aiming to improve the sustainability of the transportation system in which DRS operates must weigh increasing system efficiency through reduced road use against increasing system efficiency through providing a reliable DRS service that may attract passengers from less efficient single-occupancy exclusive ridehailing trips. Deciding how to weigh each of these factors is beyond the scope of this research – instead, we present system performance in each of these metrics to provide context to these decisions with information about the relative tradeoffs that these methods provide.

4.3 Results

In this section, we analyze the results of the simulation experiments across each of the network types, levels of travel time correlation, and optimization parameters. Overall, we evaluated solutions on 8 different combinations of network structure and correlation, using 3 different objective function parameter settings and 7 different sample sizes. For each combination of these factors, we generated 10 different solutions using different travel time samples on each of 20 unique scenarios involving different locations for request origins and destinations and vehicle starting positions. These combinations resulted in a total of 30,800 simulation runs along with 260 unique benchmark solutions (as benchmark solutions didn't vary based on the simulation parameters).

This section is divided into three main parts, each with a different focus of analysis. The first section focuses on the performance of the stochastic assignment method in relation to the specific scenarios evaluated in the simulations, both in absolute terms and relative to the benchmark method. This investigates what factors related to the spatial distribution of supply and demand are most critical in terms of system performance, and how different network configurations and solution methods influence system performance in an absolute sense. In the second section, we evaluate how the relative performance of the stochastic method in comaparison to the benchmark method depends upon the network configuration and optimization parameters across all simulated scenarios. This analysis provides insight as to the relative benefits and drawbacks of using the novel stochastic assignment method over the deterministic benchmark, and how different operation environments affect the magnitude of these changes. Finally, we zoom out from the complex tradeoffs between each of these four metrics and compare the performance of the stochastic method to the benchmark method in terms of the Pareto efficiency of solutions. This analysis demonstrates the capacity of each method to find solutions which are best for some combination of priorities among these metrics without imposing any judgment as to which are more or less important than others.

4.3.1 Assignment Performance by Scenario

We first analyze how each assignment method performs across the different scenario contexts which we implemented in the simulation framework. As performance may vary greatly across the different individual scenarios that we simulated, due to the differences in the initial locations of vehicles and the request origin and destination locations, analyzing the average performance across all scenarios may obscure finer details of the data. This analysis therefore provides information regarding the spread of system performance across the simulated scenarios, as well as how the spatial distribution of vehicles and passengers in each scenario affects performance.

To give a sense of the range of scenario performances, Tables 4.4, 4.5, 4.6, and 4.7 show the performance of the solutions found by the benchmark method and the stochastic method with W=0 for two of the 20 scenarios evaluated in each network configuration. For each network configuration, each table shows the performance in the scenario which resulted in the lowest value of one metric, labeled "Min", and the performance in the scenario which resulted in the greatest value of that metric, labeled "Max".

The average delay experienced by passengers under the generated assignments ranges from a minimum of under 1 minute to a maximum of about 6.5 minutes depending on the scenario. Given the small size of the synthetic networks used to increase the number of good performance routings, the low upper end of this spectrum makes sense, as does the fact that the smallest network, Monocentric(10,60), has the smallest delays at the upper end.

Meanwhile, all scenarios feature at least some delay for some passengers, succeeding at the goal of limiting how many optimal zero-delay solutions exist. While the smallest network, Monocentric(10,60), has the lowest maximum delays, it also has the highest minimum delays. This can be explained by evaluating the scenarios in terms of the average distance from each request's origin to the initial location of the nearest vehicle. The specific configuration of supply and demand in the Monocentric(10,60) network leads to this average distance to be higher than in the other networks. As

Average Delay (minutes)			Bench-mark	Stochastic Method Sample Size						
Scenario	Corr.	Scenario		10	25	50	100	150	200	250
Mono (10,60)	Low	Min	0.53	0.53	0.53	0.50	0.50	0.50	0.50	0.50
		Max	2.96	2.74	2.72	2.71	2.71	2.71	2.70	2.70
	Med.	Min	1.17	0.88	0.86	0.96	0.90	0.95	0.95	0.95
		Max	4.41	4.50	4.52	4.41	4.44	4.41	4.41	4.41
	High	Min	2.20	2.16	2.12	2.13	2.13	2.12	2.10	2.10
		Max	4.87	4.95	4.88	4.85	4.85	4.86	4.85	4.85
Mono (30,30)	Low	Min	0.45	0.39	0.38	0.38	0.38	0.38	0.38	0.38
		Max	3.99	3.63	3.60	3.60	3.60	3.60	3.60	3.60
	High	Min	1.17	1.18	1.18	1.17	1.17	1.17	1.17	1.17
		Max	5.99	5.58	5.58	5.58	5.58	5.58	5.58	5.58
Poly (30,30)	Low	Min	1.09	0.82	0.81	0.79	0.79	0.79	0.79	0.79
		Max	4.54	4.24	4.21	4.21	4.21	4.21	4.21	4.21
	Med.	Min	1.09	0.81	0.80	0.79	0.79	0.79	0.79	0.79
		Max	4.54	4.21	4.21	4.21	4.20	4.20	4.20	4.21
	High	Min	2.23	1.65	1.64	1.65	1.65	1.63	1.62	1.62
		Max	6.51	6.16	6.09	6.11	6.07	6.08	6.07	6.06

Table 4.4: Average delay, in minutes, for each network configuration for the individual scenario which generated the minimal and maximal average delay. Stochastic method results indicate the solutions found when $W = 0$.

the time limit on pickups of 5 minutes is fixed and does not change depending on the network configuration, starting vehicles further away from the pickup locations therefore corresponds to incurring more delay.

Finally, we note that the stochastic methods almost always yield better performance in terms of average delay than the benchmark solution. Additionally, this reduction in delay generally increases as the number of samples drawn from the travel time distribution increases, demonstrating greater reliability in finding optimal or near-optimal solutions with a larger sample size. However, there are some cases in which the stochastic method does not reduce average delay significantly. In these cases, the smallest sample sizes used sometimes result in an increase in delay, although greater sample sizes are able to match the performance of the benchmark method even if they do not improve upon it.

Looking at performance in terms of the frequency of late arrivals beyond the requests' time constraints, as shown in Table 4.5, we observe what would be quite

Late Arrival Rate			**Bench-mark**	**Stochastic Method Sample Size**						
Scenario	Corr.	Scenario		10	25	50	100	150	200	250
Mono (10,60)	Low	Min	39%	39%	41%	40%	40%	41%	41%	40%
		Max	85%	86%	86%	86%	86%	86%	86%	86%
	Med.	Min	39%	46%	50%	48%	48%	46%	46%	46%
		Max	100%	100%	100%	100%	100%	100%	100%	100%
	High	Min	54%	53%	53%	52%	52%	52%	52%	52%
		Max	90%	90%	89%	90%	90%	90%	90%	90%
Mono (30,30)	Low	Min	27%	36%	36%	36%	36%	36%	36%	36%
		Max	79%	99%	99%	99%	99%	99%	99%	99%
	High	Min	39%	46%	44%	43%	43%	43%	43%	43%
		Max	79%	95%	98%	98%	98%	98%	98%	98%
Poly (30,30)	Low	Min	36%	29%	26%	26%	26%	26%	26%	26%
		Max	66%	74%	74%	74%	74%	74%	74%	74%
	Med.	Min	36%	28%	26%	26%	26%	26%	26%	26%
		Max	66%	74%	75%	74%	74%	74%	74%	74%
	High	Min	44%	40%	38%	37%	37%	37%	37%	37%
		Max	82%	82%	81%	82%	83%	83%	83%	83%

Table 4.5: Frequency of late arrivals for each network configuration for the individual scenario which generated the minimal and maximal late arrival rate. Stochastic method results indicate the solutions found when $W = 0$.

poor performance from a real-world system, with between 25% and 100% of passengers being delayed beyond their latest acceptable time windows. This is a result of the aggressive time bounds used for the passengers, which accomplish the design goal of yielding few solutions that result in no delay.

In contrast to average delay, we do not observe much improvement in terms of late arrival frequency from the stochastic method compared to the benchmark method. While there are some scenarios where the stochastic method results in reductions in both delay and late arrivals, there are others where late arrival rate is equal between the two method or even increases. We also note that the impact of sample size on late arrival frequency varies between scenarios, with some scenarios featuring stable performance, others showing a decrease in late arrival rate as sample size increases, and a few others showing the opposite.

Table 4.6 shows how the average variance of arrival times of vehicles at each passenger location differs across individual scenarios on each netowrk. As expected,

Average Variance of Arrival Times (minutes)			Benchmark	Stochastic Method Sample Size						
Scenario	Corr.	Scenario		10	25	50	100	150	200	250
Mono (10,60)	Low	Min	0.63	0.72	0.72	0.72	0.72	0.72	0.72	0.72
		Max	1.33	1.58	1.66	1.70	1.62	1.70	1.74	1.74
	Med.	Min	2.51	2.74	2.56	2.55	2.55	2.54	2.54	2.54
		Max	6.42	5.78	5.58	5.39	5.39	5.39	5.39	5.39
	High	Min	7.58	7.91	7.62	7.60	7.61	7.62	7.58	7.58
		Max	16.7	17.5	16.9	17.1	16.7	16.7	16.7	16.7
Mono (30,30)	Low	Min	0.34	0.47	0.47	0.47	0.47	0.47	0.47	0.47
		Max	1.45	1.45	1.45	1.45	1.45	1.45	1.45	1.45
	High	Min	3.42	3.42	3.38	3.33	3.31	3.31	3.31	3.33
		Max	10.7	10.4	10.4	10.4	10.4	10.4	10.4	10.4
Poly (30,30)	Low	Min	1.64	1.48	1.47	1.46	1.46	1.47	1.44	1.41
		Max	4.39	4.26	4.22	4.22	4.22	4.22	4.22	4.22
	Med.	Min	1.64	1.42	1.44	1.44	1.42	1.41	1.42	1.44
		Max	4.39	4.25	4.22	4.22	4.22	4.22	4.22	4.22
	High	Min	7.47	5.52	5.59	5.52	5.45	5.67	5.74	5.74
		Max	15.0	14.2	14.1	14.7	14.2	14.2	14.2	14.2

Table 4.6: Average arrival time variance, in minutes, for each network configuration for the individual scenario which generated the minimal and maximal average arrival time variance. Stochastic method results indicate the solutions found when $W = 0$.

a greater level of travel time correlation on each network results in an increase in the variance of arrival times. This is because, for long paths consisting of multiple links, the covariance of individual link travel times is the greatest contributor to overall path travel time variance.

Across the three network structures used, we observe greater arrival time variability on the Monocentric(10,60) and Polycentric(30,30) networks relative to the Monocentric(30,30) network. We also, as with late arrival rate, find mixed outcomes when comparing the various stochastic methods' performance. For some scenarios, the stochastic method results in a greater average arrival time variance than the benchmark method, while in others it results in a decrease in this metric. Sometimes, increasing sample size results in an increase in arrival time variance, while other times it results in a decrease. However, we consistently find that across all six scenarios observed in the Polycentric(30,30) network, the stochastic method reduces average

Total VKT			**Bench-mark**	**Stochastic Method Sample Size**						
Scenario	Corr.	Scenario		10	25	50	100	150	200	250
Mono (10,60)	Low	Min	28.0	27.7	27.6	27.7	27.5	27.5	27.5	27.6
		Max	39.0	42.8	42.8	42.8	42.8	42.8	42.8	42.8
	Med.	Min	25.0	25.9	26.5	26.5	26.5	26.5	26.5	26.5
		Max	38.8	39.5	39.0	39.1	38.8	38.8	38.8	38.8
	High	Min	28.0	28.0	28.1	28.1	28.0	28.0	28.0	28.0
		Max	39.0	39.3	39.6	39.3	39.4	39.5	39.6	39.6
Mono (30,30)	Low	Min	27.0	27.9	27.6	28.1	27.6	27.9	27.8	27.8
		Max	46.8	46.8	46.8	46.8	46.8	46.8	46.8	46.8
	High	Min	27.0	27.3	27.3	27.1	27.1	27.1	27.1	27.0
		Max	46.8	46.8	46.8	46.8	46.8	46.8	46.8	46.8
Poly (30,30)	Low	Min	27.0	27.0	27.0	27.0	27.0	27.0	27.0	27.0
		Max	41.2	43.1	42.8	43.0	43.0	43.0	43.0	43.0
	Med.	Min	27.0	27.0	27.0	27.0	27.0	27.0	27.0	27.0
		Max	41.2	42.3	43.0	43.0	43.0	43.0	43.0	43.0
	High	Min	27.0	27.5	27.0	27.0	27.0	27.0	27.0	27.0
		Max	41.2	42.8	43.0	43.0	43.0	43.0	43.0	43.0

Table 4.7: Total vehicle kilometers traveled (VKT) for each network configuration for the individual scenario which generated the minimal and maximal total travel distance. Stochastic method results indicate the solutions found when $W = 0$.

arrival time variance, which is not true of the other two networks.

Finally, Table 4.7 displays how the total distance that vehicles travel when serving their assigned routes varies across individual scenarios. Note that the results shown are only in situations where the objective weight for this metric is set to 0 in the objective function of the stochastic method. Because this metric is not part of the objective, we do observe increases in VKT for many scenarios for the stochastic method relative to the benchmark. In fact, the only scenario of the ones presented in which we observe a reduction in VKT from the stochastic method is the minimum scenario for the Monocentric(10,60) network with low travel time correlation.

These tables are meant to provide a sense of the range that each metric of interest takes across the scenarios evaluated in the simulation experiments. The trends observed can't be generalized to all scenarios within the same network configuration and travel time correlation level. For that reason, we turn to aggregate analysis across all 20 scenarios to gain further insights.

Factors Influencing Scenario-Level Performance

Earlier, we mentioned that the magnitude of delay (and thereby the frequency of late arrival) may be strongly influenced by the average distance between each request's origin and the starting location of the nearest vehicle. We explored how this physical measure of each scenario corresponded to the performance of solutions within that scenario to develop our understanding of how performance is influenced by the network configuration. In addition to this measure, we also conducted the same analysis using the average distance between each request's origin and its destination, as well as the average minimal trip distance for each request using the nearest vehicle (the sum of the distance between each request's origin and the starting location of its nearest vehicle plus the distance between each request's origin and its destination). However, we did not observe strong patterns connecting performance to these measures, so the results are excluded for brevity.

Figure 4-3 shows how the average delay and frequency of late arrival for passengers in a given scenario varies with this measure of average distance from each request's origin to the starting location of the closest vehicle to it. As expected, we observe an increasing trend where greater distances between requests and vehicles result in greater delays faced by passengers, as well as a higher frequency of late arrivals. This trend appears to hold across all three networks used in the simulation experiments and all levels of travel time correlation. We also do not observe significant differences between the different objective weights used. However, we do not that scenarios with greater travel time correlation appear to result in increased average delay, all else equal, which matches the initial findings observed in the data from Table 4.4. High correlation also appears to increase the rate of late arrivals, although the difference is not as great compared to other levels of correlation as the difference in delay.

Given the strong relationship between this measure and the performance of the system in each scenario, we are also interested in how this measure relates to the difference in performance between solutions found by the stochastic method and those found by the benchmark method. Figure 4 shows the results of this analysis. We

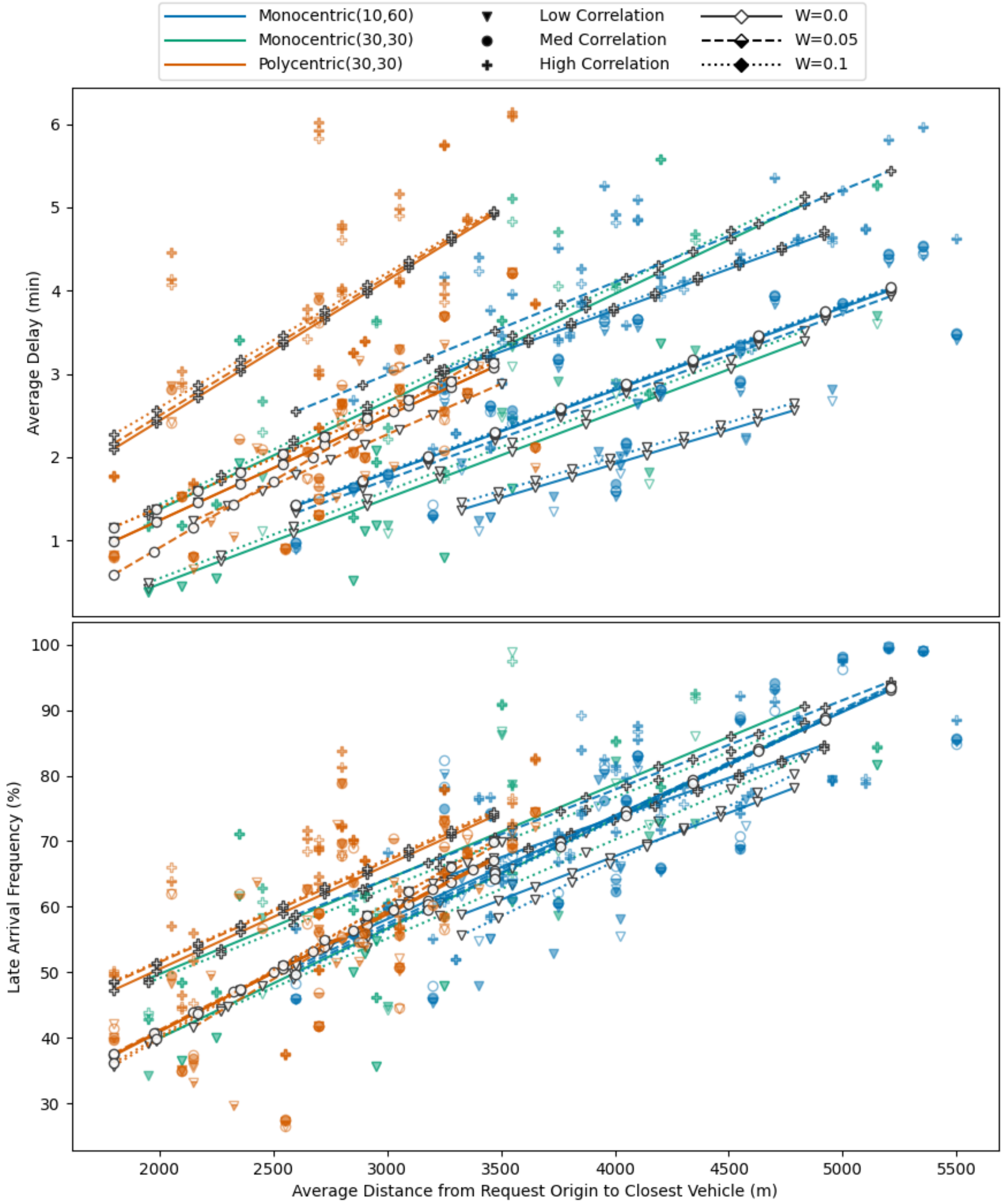


Figure 4-3: A scatter plot of the average delay (top) and the frequency of late arrivals (bottom) obtained by stochastic method solutions against the average distance from each request's origin to the starting location of the nearest vehicle. Each point corresponds to a single scenario, with delay averaged across all solutions found for that scenario across all sample sizes. Points are categorized by network type, level of travel time correlation, and objective weight parameter W. Lines of best fit are also shown for all data corresponding to a single category.

observe a much weaker relationship between the relative locations of passengers and vehicles within each scenario and the magnitude of changes obtained by the stochastic method relative to the benchmark. This appears to indicate that the degree of improvement that can be obtained by the stochastic method depends on deeper structure of each specific scenario.

Although within each network type, the average distance between request origins and their closest vehicles does not appear to strongly impact the change in performance relative to the benchmark, there does appear to be meaningful differences between the specific networks implemented in the simulation in terms of the reduction in average passenger delay. While the Polycentric(30,30) and Monocentric(30,30) appear to have similar changes in delay relative to the benchmark method at any given level of average vehicle-passenger distance, the Monocentric(10,60) network appears to differ significantly. While it does on average have greater distance between the request origins and initial vehicle locations, solutions on this network result in less improvement over the benchmark method even compared to scenarios with similar values in this metric in the other two networks. This difference is especially notable compared to the Monocentric(30,30) network, which uses the same travel time distribution. This appears to indicate that the network shape (long vs. square) may be more important for the potential benefits of the stochastic method relative to the benchmark method than the specific form of the travel time distribution on either type of network.

Given the strong relationship between this measure and the performance of the system in each scenario, we are also interested in how this measure relates to the difference in performance between solutions found by the stochastic method and those found by the benchmark method. Figure 4-4 shows the results of this analysis. We observe a much weaker relationship between the relative locations of passengers and vehicles within each scenario and the magnitude of changes obtained by the stochastic method relative to the benchmark. This appears to indicate that the degree of improvement that can be obtained by the stochastic method depends on deeper structure of each specific scenario.

Although within each network type, the average distance between request origins and their closest vehicles does not appear to strongly impact the change in performance relative to the benchmark, there does appear to be meaningful differences

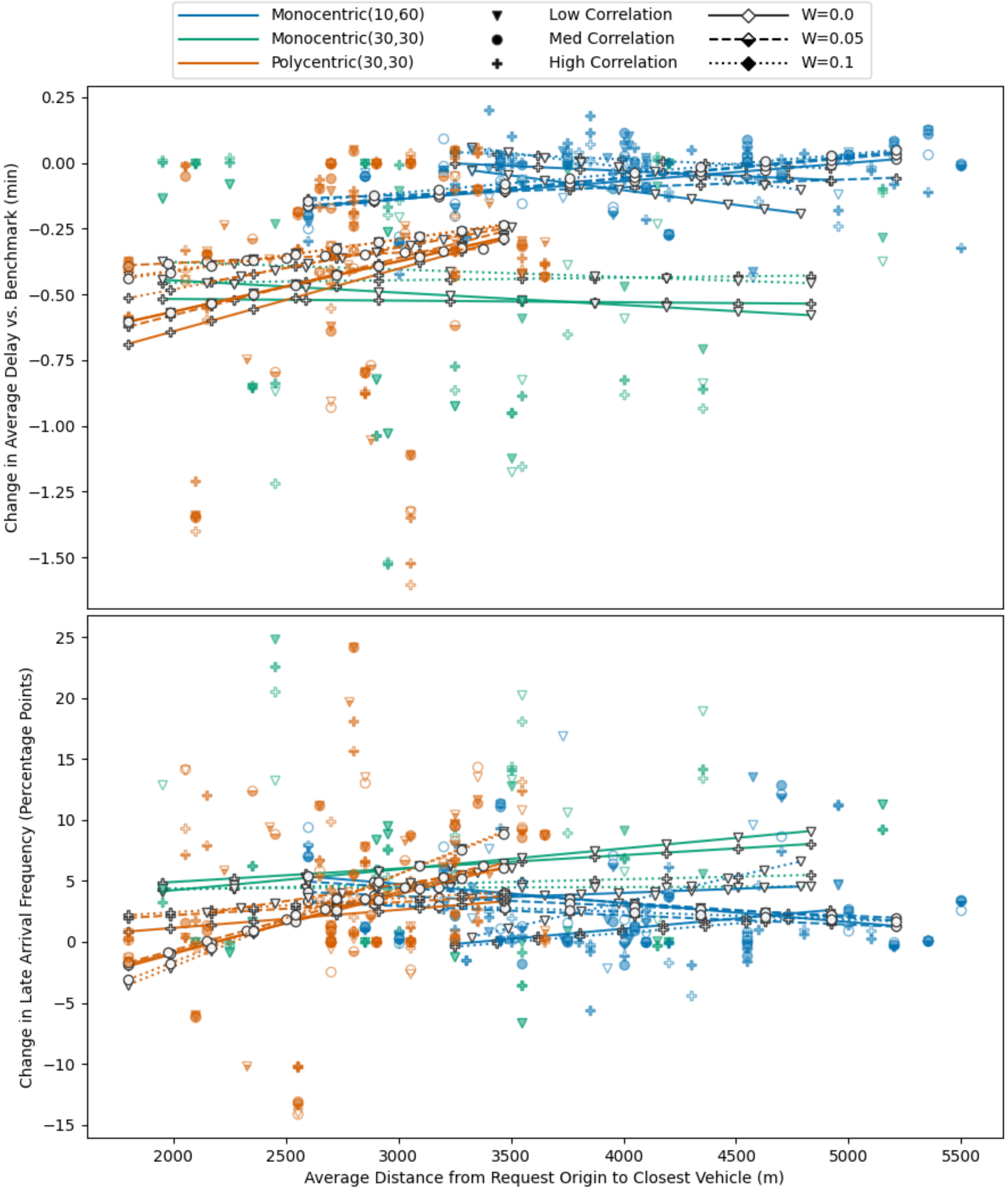


Figure 4-4: A scatter plot of the change in average delay (top) and the change in frequency of late arrivals (bottom) obtained by stochastic method solutions relative to the benchmark solution against the average distance from each request's origin to the starting location of the nearest vehicle.

between the specific networks implemented in the simulation in terms of the reduction in average passenger delay. While the Polycentric(30,30) and Monocentric(30,30) appear to have similar changes in delay relative to the benchmark method at any given level of average vehicle-passenger distance, the Monocentric(10,60) network appears to differ significantly. While it does on average have greater distance between the request origins and initial vehicle locations, solutions on this network result in less improvement over the benchmark method even compared to scenarios with similar values in this metric in the other two networks. This difference is especially notable compared to the Monocentric(30,30) network, which uses the same travel time distribution. This appears to indicate that the network shape (long vs. square) may be more important for the potential benefits of the stochastic method relative to the benchmark method than the specific form of the travel time distribution on either type of network.

4.3.2 Stochastic Performance Relative to Benchmark Method

Having investigated how the low-level details of supply and demand positioning in the network affect the performance of the assignment methods, we turn our attention to understanding the higher level performance across each of these simulated instances. In this section, we aim to answer the question of how the stochastic method performs relative to the benchmark method in aggregate across all simulated scenarios, and how the network configuration and solution method affect these relative differences. To do this, we evaluated the relative change in several metrics of interest discussed in the previous section when using the stochastic method compared to the corresponding benchmark solution for each of the 30,800 individual simulations. For each combination of network type, degree of correlation, objective weight parameter, and sample size, we then averaged this relative performance across all runs using that configuration.

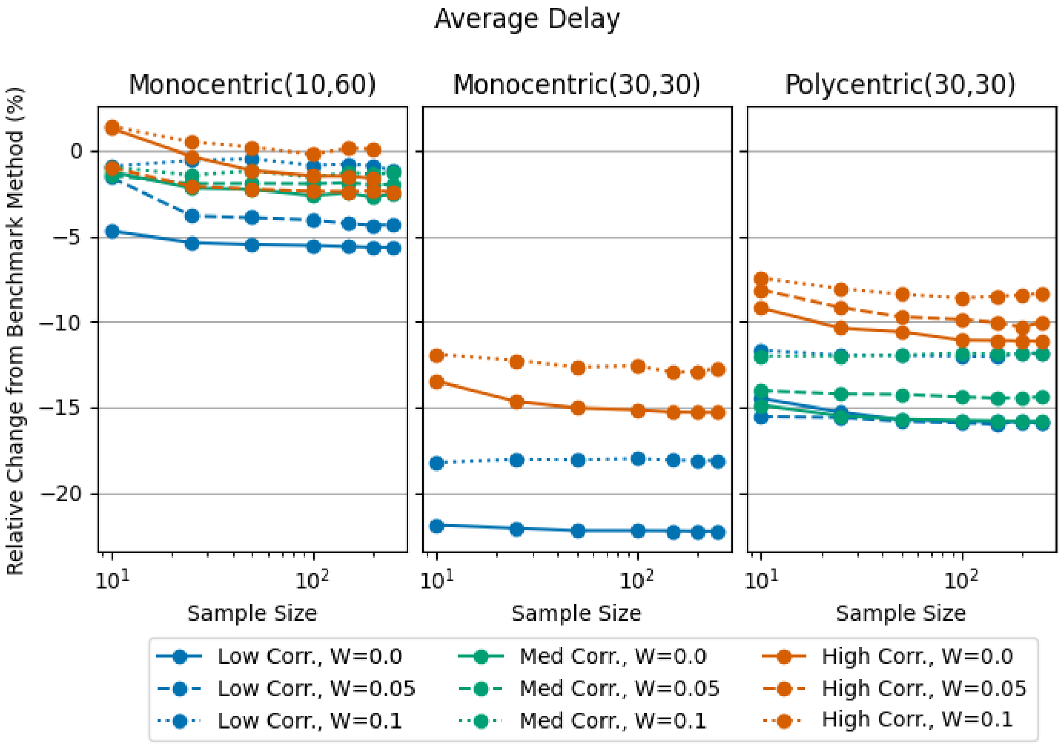


Figure 4-5: Improvement in average delay at each passenger origin/destination location from the benchmark method. The left subplot shows results from the Monocentric(10,60) network; the middle shows results from the Monocentric(30,30) network; and the right shows results from the Polycentric(30,30) network. Each line reflects a combination of network correlation level and objective weights used in the stochastic programming formulation. The x-axis shows the number of travel time samples used as input to the stochastic programming model, and the y-axis shows the average improvement over the benchmark model's performance in the same network configuration, as a percentage.

Average Delay

Figure 4-5 shows the change in the average delay experienced by passengers resulting from using our stochastic optimization formulation relative to the benchmark method's performance in identical scenarios. For the Monocentric(10,60) network, the stochastic method ranges from decreasing average delay by just over 5.0% to increasing it by roughly 1.3% depending on the specific model configuration and operating environment. Notably, the only situations which observe an increase in average delay under the stochastic optimization method are those in highly correlated networks, either where the objective weight for total distance traveled is high ($W = 0.1$) or at low sample sizes. For both of the other networks, we observe a significant decrease in average delay across all combinations of inputs to the model, with the greatest reductions yielded by the Monocentric(30,30) network.

Across all three network configurations, there are a few consistent similarities in the effects of different model inputs on average delay. For any given level of network correlation on either test network, for any sample size used as input to the stochastic assignment method, placing a greater weight on total travel distance in the objective function results in worse perforamnce in terms of average delay. This is a natural consequence of incorporating a secondary objective into the optimization method, reducing the extent to which the primary objective (in this case, average delay), is prioritized.

Increasing the sample size of travel times given to the stochastic assignment method generally leads to improvements in average delay. These improvements are largest at low sample sizes – increasing from 10 to 25 to 50 samples leads to larger improvements than increasing sample size from 50 to 250 in most cases.

Additionally, we find that increasing sample size has the largest impact comparatively for networks with the highest degree of correlation in travel times. This makes intuitive sense as a small sample from a distribution with high covariance will be less likely to capture the full range of possible outcomes, whereas for a distribution with less internal correlation a smaller sample size may be sufficient to obtain solutions with good performance over the full possibility space. However, the differences between performance at various sample sizes are significantly less than the differences due to network structure, travel time correlation, or objective weights.

Relative to its performance on the larger and more dispersed (30,30) networks, the stochastic method demonstrates less improvement in average delay on the Monocentric(10,60) network. For the other networks, we find that scenarios with greater levels of link travel time correlation feature the lowest degree of improvement over the benchmark method in terms of average delay compared to scenarios with less correlation.

These findings may be related if we think of the degree of choice available in various

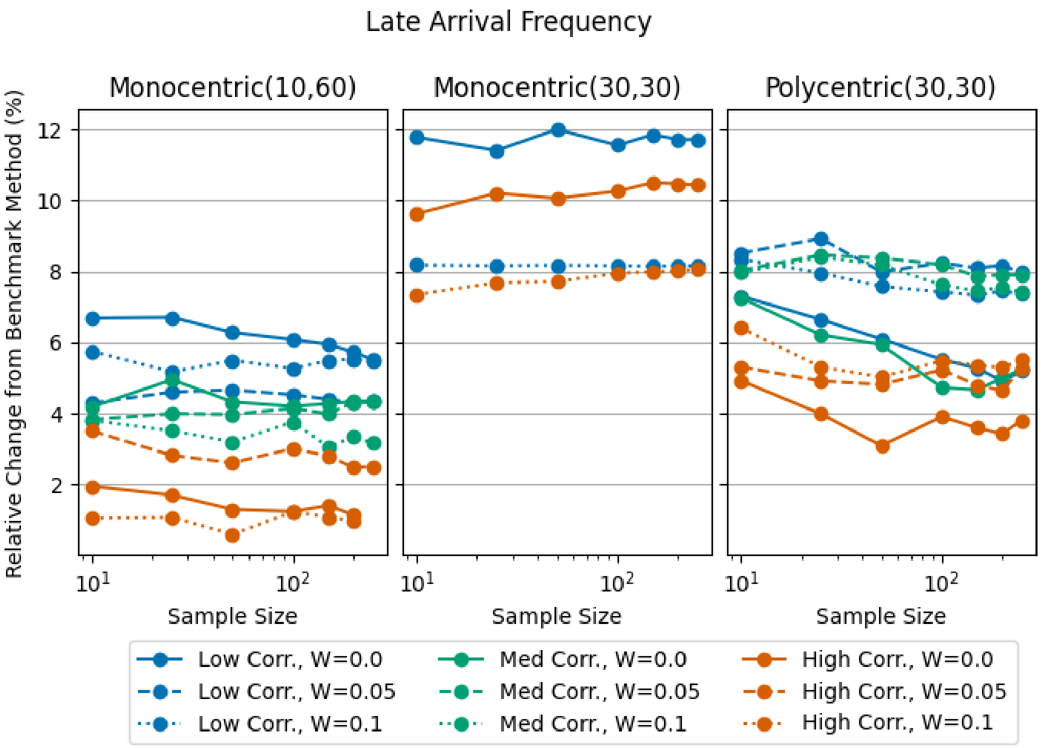


Figure 4-6: Change in late arrival rate from the benchmark method.

networks as an indicator of the potential for improvement over the benchmark method. The Monocentric(10,60) network physically constrains the potential paths between vehicles and requests and reduces the magnitude of any differences between potential solutions, reducing the possible extent of improvement by the stochastic method. Similarly, increasing travel time correlation, which has the effect of increasing travel time variability as demonstrated in Figure 4-2, means that every solution faces greater risk of delay. This reduces the differences between solutions, limiting the potential improvement from the stochastic method.

The average delay across all locations in the network was the primary objective of the stochastic assignment formulation, and so the improvements observed over the benchmark method demonstrate the comparative advantage of this method in reducing the magnitude of passenger delays under stochastic travel time environments.

Late Arrival Rate

Figure 4-6 shows how the frequency of late arrivals of the system's vehicles at locations within the network changes when using the stochastic assignment method relative to the benchmark for each set of experiments. We first observe that, on average, the stochastic assignment method always increases frequency of late arrivals relative to the benchmark performance. This finding holds in aggregate, taking all simulated scenarios for any given network configuration and solution method, though there are individual scenarios for which the stochastic method does reduce the late arrival frequency compared to the benchmark, as seen in Figure 4-4.

For the Monocentric(10,60) network, the increase in late arrival rate ranges from around 1 to 7%; in the Polycentric(30,30) network, it ranges from roughly 3 to 9%; meanwhile, the Monocentric(30,30) network generates the greatest increases in late arrival frequency, ranging from 7 to 12%. The fact that the networks with the greatest relative reduction in average delay yield the greatest increases in late arrival frequency indicates a tradeoff between average delay and late arrival frequency that arises in DRS assignment problems. This tradeoff is explored further in Section 4.3.3.

The level of travel time correlation in the network plays a very large influence on the change in late arrival rate attained by the stochastic assignment method. In the Monocentric(10,60) network, operations in networks with a low correlation between link travel times have the greatest increase in late arrival rate compared to the benchmark, while those with high correlation have the lowest, and medium correlation predictably falling between the two extremes. For the Polycentric(30,30) network, the separation between low and medium correlation is not as clear, but once again scenarios with high levels of travel time correlation result in the least increase in late arrival rate in comparison to the benchmark.

The objective weight used to weight total travel distance appears to have opposite impacts on late arrival frequency between the two types of network travel time distributions. For both monocentric model networks, placing a greater weight on total travel distance and reducing the extent to which delay is minimized generally reduces the frequency of late arrivals, with the exception of highly correlated travel time scenarios in the Monocentric(10,60) network. This follows intuitively from the tradeoff between the magnitude and frequency of delay, with less emphasis being placed on reducing the magnitude of delay allowing for smaller increases in the frequency of delay. However, in the Polycentric(30,30) network, the smallest increases in late arrival frequency arise from methods using W=0.0. This could indicate a tradeoff that arises between total travel distance and late arrival rate on this network type, although it is unclear what differences between the polycentric and monocentric travel time distributions could lead to this different behavior.

While increasing sample size generally leads to a reduction in late arrival frequency, the magnitude of this effect varies considerably among different network structures, travel time correlation levels, and objective weights. However, there are very few cases where increasing sample size leads to an increase in late arrival rate. The fact that increasing sample size reduces both the magnitude and frequency of delay despite the difficulty in minimizing both of these quantities indicates the power of larger sample sizes in stochastic applications.

Arrival Time Variability

Figure 4-7 depicts the change in the average variance of arrival times across all locations in the network obtained by the stochastic assignment method relative to the benchmark. This metric is obtained by finding the variance of vehicle arrival times at each location in the network, then averaging that variance across all locations. It therefore indicates the degree to which an average passenger may expect their pickup and drop-off times to vary under a given scenario.

We find that, with the exception of low correlation scenarios on the Monocentric(10,60) network, the stochastic method decreases arrival time variability across all network structures, correlation levels, and objective weight values. In some cases, these changes are close to zero, but many scenarios result in moderate decreases of 5 to 10%, with scenarios featuring highly correlated travel times on the Monocentric(30,30) network featuring reductions of up to 15%.

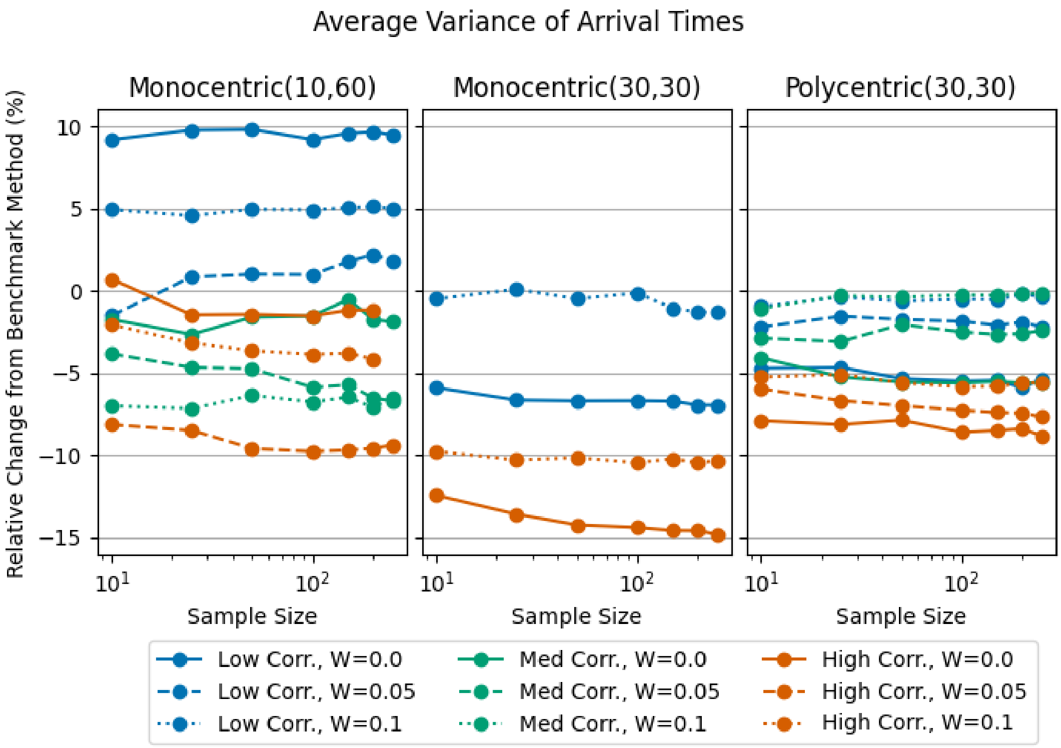


Figure 4-7: Change in the average variance of arrival times relative to benchmark method.

This indicates that not only does the stochastic assignment method reduce the average delay faced by passengers beyond the time bounds imposed upon the optimization, but also that it reduces the degree to which their arrival times vary, independent of anny imposed time bounds. This finding is very important because, as noted previously, the magnitude and frequency of delay are dependent upon the time bounds imposed upon the system, but variance of arrival times is independent of this factor and is harder to alter through means outside of assignment and routing methods.

For the larger (30,30) networks, the scenarios with higher correlation result in greater decreases in arrival time variability than those with lower correlation. The greater path travel time variability associated with this increased correlation results in greater arrival time variance experienced by passengers, but it reflects well upon the stochastic assignment method that it is able to make larger relative improvements in these environments. On the Monocentric(10,60) network, both the medium and high correlation scenarios result in similar performance for the stochastic assignment method, while the scenarios with low correlation result in increases in average arrival time variance.

The objective weight also appears to have differing impacts on reduction in arrival time variability for the different network scales. For the Monocentric(10,60) network, placing no objective weight on total travel distance results in the worst performance in terms of this metric. Meanwhile, on both (30,30) networks, increasing the objective weight placed on travel distance decreases the magnitude of reduction in arrival time variability. This indicates that there may exist a tradeoff between the extent to which it is possible to minimize total travel distance and arrival time variability, perhaps arising from choices between longer, less variable routes on smaller roads and shorter routes along more large throughways with greater travel time variability. This explanation could also provide insight as to why the Monocentric(10,60) network behaves differently, as fewer routes will travel along the throughways due to their existence mainly along the periphery of the network or perpendicular to the major direction of travel.

Total Distance Traveled

Figure 4-8 displays the relative change in total vehicle distance traveled for each simulation environment when using the stochastic method over the benchmark method. We find that instances of the stochastic assignment method which place some degree of objective weight on minimizing this metric achieve reductions in total distance traveled relative to the benchmark, while those that aim only to minimize average passenger delay result in much smaller reductions or even increases.

For both the Monocentric(10,60) and Polycentric(30,30) network, using $W = 0$ results in modest increases in total vehicle distance traveled, while for the Monocentric(30,30) network it achieves a reduction, though this reduction is much smaller than when using $W = 0.1$. This difference between the network structures can also be seen when placing greater weight on total travel distance in the objective function, as the Monocentric(30,30) network is able to achieve reductions on the order of 4 to

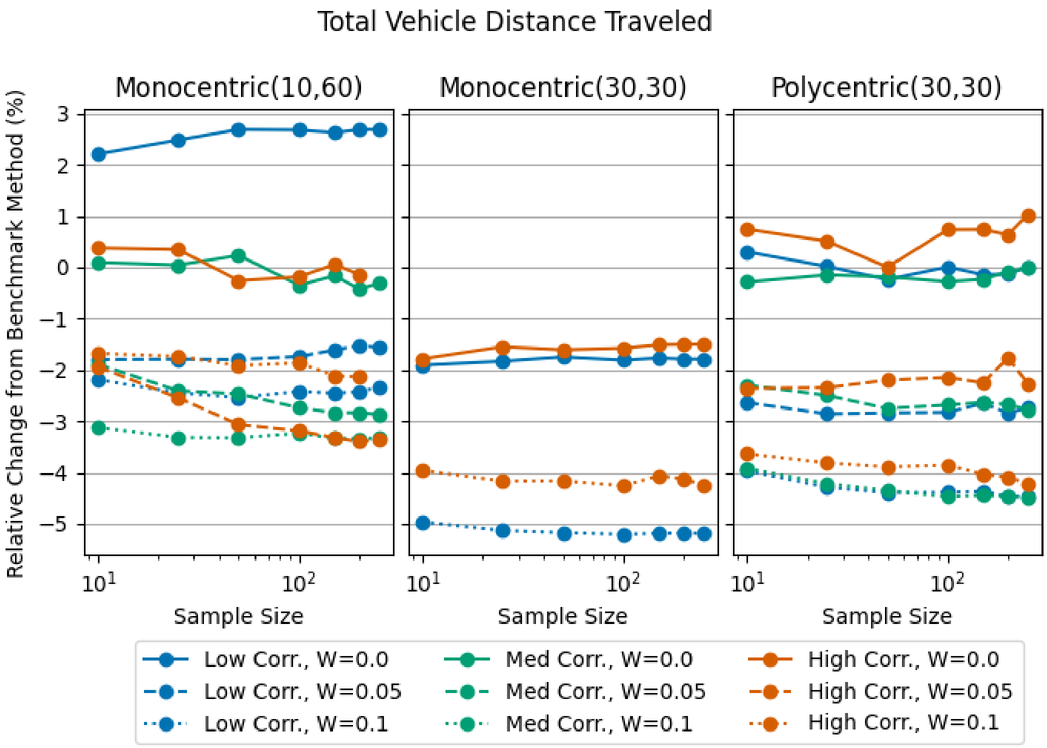


Figure 4-8: Change in the total distance traveled by vehicles relative to the benchmark method.

5%, while the other two network structures result in reductions of at most 4%.

The impact of the parameter W on these results, and the potential for increased total distance traveled when $W = 0$, indicates that total vehicle distance traveled must be explicitly captured in the goals of the assignment method in order to achieve reductions. As opposed to the joint reductions achieved in average delay and arrival time variability by many of the stochastic assignment instances, we find that reducing average delay on its own does not yield reductions in total distance traveled on the aggregate. However, the fact that methods using $W > 0$ are able to achieve reductions in both total distance and some delay metrics in many cases indicates the potential to improve performance for both of these goals at the same time if the correct method is chosen to do so.

Differences between travel time correlation levels within the individual network structures are generally small, which may be expected as travel time correlations affect only the service time components of operation as measured by the other metrics rather

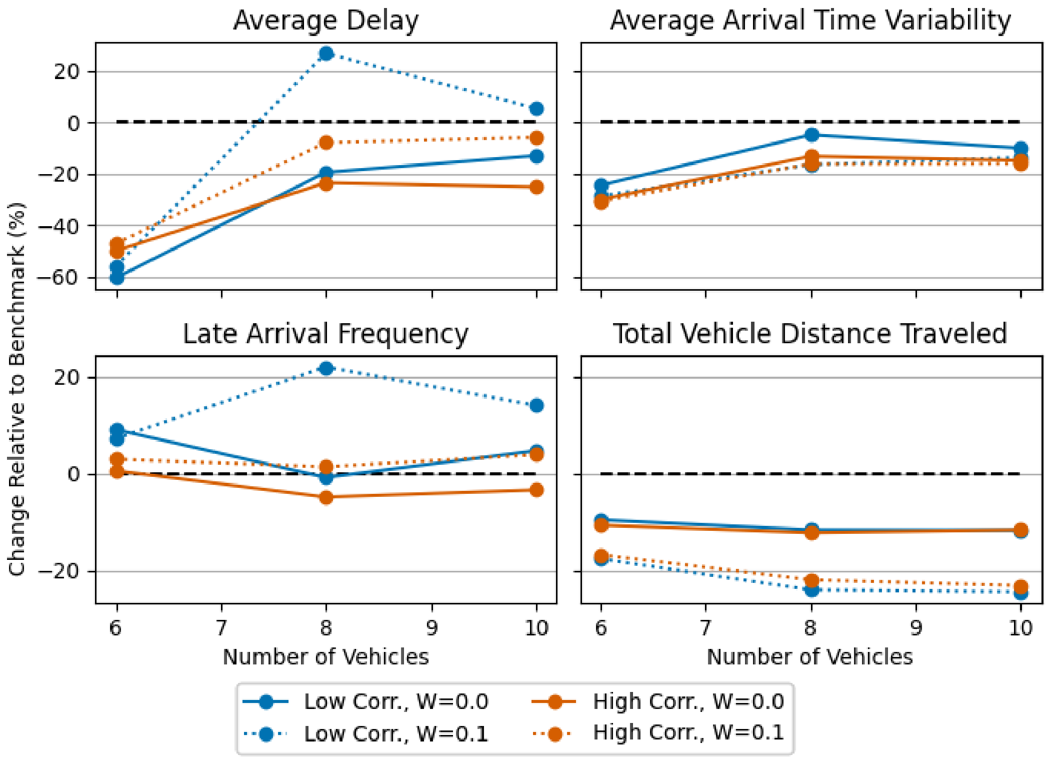


Figure 4-9: Relative change in metrics of interest compared to benchmark method for scenarios with 8 requests and 6, 8, or 10 vehicles. All simulations were conducted with 50 travel time samples on the Monocentric(30,30) network.

than route distance, which stays constant regardless of the travel time distributions on the roads.

Surplus Vehicle Supply

The previous results demonstrate the relative benefits and drawbacks of the stochastic assignment method relative to the deterministic benchmark in situations with small numbers of vehicles and passengers, and in situations where the number of passengers exceeds the number of vehicles available. This latter situation forces the vehicles to pool, or at least to sequence multiple passenger trips in the same vehicle, and also leads to natural increases in all of the metrics we investigate in this section.

Figure 4-9 displays the relative performance across all four metrics of interest obtained by the stochastic assignment method in larger-scale situations with 8 requests and 6, 8, or 10 vehicles. In these scenarios, 4 of the requests, chosen at random, are pre-assigned to their nearest vehicle, and each assignment method is run while holding those previous assignments fixed. The variation of the number of vehicles available to serve the requests provides insight as to how the findings of the previous sections are changed as the balance of supply to demand shifts from excess demand, to balanced supply and demand, to excess supply.

We find similar performance with 6 vehicles as with the previous set of experiments, although the magnitude of the reductions in delay, arrival time variability, and total distance traveled appear to be larger with a greater number of actors in the system. The greater number of possibilities associated with the assignment problem at this problem scale creates greater room for the stochastic assignment method to improve upon the benchmark in situations where the number of requests exceeds the available supply of vehicles. Note also that the frequency of late arrivals does not appear to increase significantly beyond the levels seen in 4-6, indicating that these larger improvements do not come with associated downside in terms of the tradeoff with late arrival rate.

With 8 and 10 vehicles, when the system capacity meets or exceeds the number of requests seeking the performance of the stochastic assignment method relative to the benchmark still behaves mostly the same, although some surprising results appear. For assignments generated using $W = 0$, we find that these assignments are actually able to reduce the frequency of late arrivals relative to the benchmark method, which we had not seen in the previous set of experiments at all. However, for the method using $W = 0.1$ in low correlation scenarios, the method produces increases in average passenger delay relative to the benchmark in addition to large increases in late arrival frequency. This result is surprising, but can be explained by the lower magnitude of average delay in these contexts and the large reduction in total distance traveled achieved. It would appear that the high objective weight placed on total travel distance created an incentive for the optimization method to sacrifice average delay in exchange for distance savings, and that these two metrics were at odds with one another in these low correlation scenarios.

The strong performance of the stochastic method in these experiments, matching or exceeding its performance in the smaller scale tests in conditions less favorable to pooling, demonstrates the potential for this method in a variety of contexts and at larger scales.

4.3.3 Multi-Objective Efficiency

The results of the previous section demonstrate that our stochastic optimization method improves performance of the DRS system for some metrics while resulting in worse performance in others. Notably, there appears to be a tradeoff between lowering the variance of passengers' delays (and, in situations where the number of requests outnumber the available fleet, the magnitude of delays as well) and the frequency with which passengers experience some nonzero delay. Given the nature of this tradeoff, any assessment of which method is better will depend on the perceived goals of the system. Regardless of system goals or the choice of objective function for evaluation, however, we can evaluate the performance of both methods in terms of their ability to deliver an optimal solution for some choice of objective function by examining the Pareto efficiency of solutions.

For each assignment generated by the simulation, we evaluate the assignment by four metrics that we previously established as representative of important aspects of system performance - average passenger delay, late arrival rate, average arrival time variability, and total vehicle distance traveled. An assignment is Pareto efficient if no other possible assignment of vehicles to passengers could result in a decrease in at least one of these metrics while holding the rest constant. If one assignment results in a decrease in one of these metrics compared to a second assignment, and all other metrics either hold constant or also decrease, then the first assignment Pareto dominates the second – the second is likewise Pareto dominated by the first. The set of all Pareto efficient assignments for a given scenario is the Pareto frontier. Regardless of how many Pareto efficient solutions there are, identifying one is a sign that the optimization procedure has identified a strong solution in the uncertain context of the simulation environment.

While some scenarios may have only a single Pareto efficient solution, indicating one assignment which outperforms all others, other scenarios may have a larger Pareto frontier consisting of several solutions which feature tradeoffs between average delay and late arrival frequency. In the first case, finding a Pareto efficient solution is optimal regardless of the system operator's or individual's preference in regards to this tradeoff. In the latter case, however, this preference could mean that some Pareto efficient solutions are preferred to others, or even that some Pareto inefficient solutions are preferred to some Pareto efficient ones. Analyzing Pareto efficiency therefore allows us to judge the effectiveness of the various methods without defining this preference. Later, we will analyze the extent of this tradeoff in the actual solutions found by each method.

For each scenario evaluated in the simulation, we enumerate all possible assignments of the 3 vehicles and 5 requests in that scenario, and generate all associated performance metrics from that assignment across the validation travel time sample. We then find the Pareto frontier of assignments in that scenario, and compare the assignments found by both the stochastic method and the benchmark method to this frontier, and to each other.

Figure 4-10 depicts an example of this analysis for a single scenario taken from the Polycentric(30,30) network with highly correlated travel times, in only two dimensions for ease of graphical interpretability - average delay and late arrival rate. Comparing the assignments found by the stochastic method in orange to the Pareto frontier in blue, we observe that many of the solutions are Pareto efficient, though there are a few that lie off of the frontier. The benchmark solution, shown in green, also lies off of the Pareto frontier, indicating that the benchmark method failed to generate a solution with optimal performance – it was possible to reduce late arrival frequency without increasing delay.

We further compare the stochastic solutions to the benchmark solution by dividing the feasible space into quadrants relative to the benchmark solution. Stochastic solutions falling in the lower-left quadrant Pareto dominate the benchmark solution, while those falling in the upper right quadrant are Pareto dominated by it. Only a small number of solutions in this scenario fall into either of these quadrants, with

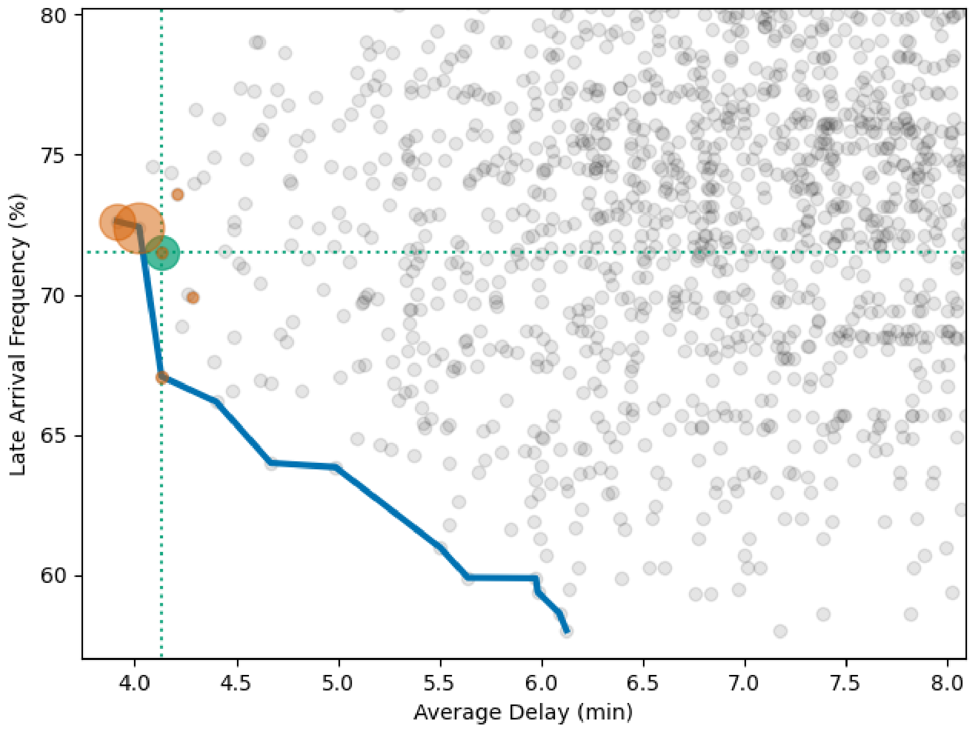


Figure 4-10: Example of Pareto efficiency analysis. Each grey dot is a single feasible solution to the assignment problem. The Pareto frontier is shown in blue. The radius of each orange circle indicates how many solutions found by the stochastic optimization method result in the combination of average delay and late arrival frequency corresponding to the center of that circle. The green circle shows the solution found by the benchmark method, with dotted lines dividing the feasible space into quadrants relative to the benchmark solution.

many more residing in the upper left and lower right quadrants, representing solutions which sacrifice one dimension to realize an improvement in the other. We note that because of the defined objective function for the stochastic optimization method, most solutions that are not Pareto dominant or Pareto dominated will be located within the upper left quadrant, minimizing average delay at the expense of late arrival frequency.

Building off of this example, we conduct the same analysis for the entirety of the simulation results and using all four performance metrics of interest, using the solutions generated for each network configuration, degree of travel time correlation, objective weight parameter, and sample size. We calculate the rate at which the benchmark method generates a Pareto efficient solution, the rate at which the stochastic method generates a Pareto efficient solution, and how frequently a solution to the stochastic method either Pareto dominates or is Pareto dominated by the benchmark method's solution when given the same scenario.

Table 4.8 displays the frequency with which each assignment method produces Pareto efficient solutions within each problem context. The benchmark method, despite its use of deterministic travel times which do not capture the variability of travel times, is still able to produce Pareto efficient solutions in most circumstances, varying between 60 and 90% frequency depending on the network structure and degree of travel time correlation. However, the stochastic method outperforms it in nearly all of the specific problem contexts simulated. With the exception of solutions generated using small sample sizes and $W = 0$ in medium and highly correlated scenarios in the Monocentric(10,60) network, the stochastic assignment method finds Pareto efficient solutions more frequently than the benchmark method in every problem context evaluated.

We also note the role that objective weight plays in the rate of Pareto efficiency. For nearly all problem contexts investigated, increasing the weight of total travel distance in the objective function increases the rate at which Pareto efficient solutions are found. Including a combination of travel distance and average delay clearly alters the solutions found by the method, resulting in more solutions which can only be improved in some performance metric at the cost of another. As total distance traveled is one of the metrics used in this analysis, the fact that incorporating that metric into the objective function results in more solutions that are on the optimal frontier of this set of performance metrics makes intuitive sense.

These findings suggest that, in aggregate, the stochastic optimization method better makes use of the information provided on travel time uncertainty to find higher quality solutions. However, as noted before, although more solutions are on the Pareto frontier they may involve tradeoffs between the four performance metrics used in this evaluation rather than providing unambiguous improvement over the less efficient solutions found by the benchmark method. To evaluate the extent to which these more efficient solutions provide reductions in both dimensions of interest, we also

Frequency of Pareto efficient solutions								
Sample Size	Monocentric(10,60)			Monocentric(30,30)		Polycentric(30,30)		
	Low Corr.	Medium Corr.	High Corr.	Low Corr.	High Corr.	Low Corr.	Medium Corr.	High Corr.
Benchmark Method								
N/A	88.0%	80.0%	75.0%	65.0%	85.0%	60.0%	60.0%	65.0%
Stochastic Method, $W = 0.0$								
N=10	94.8%	78.0%	77.5%	95.0%	93.0%	95.0%	91.5%	87.0%
N=25	96.5%	86.0%	86.0%	98.5%	97.5%	95.5%	92.5%	90.5%
N=50	97.1%	86.0%	92.5%	100.0%	100.0%	96.5%	95.5%	97.0%
N=100	98.3%	91.0%	98.0%	100.0%	100.0%	94.5%	96.0%	98.0%
N=150	97.7%	87.0%	95.5%	100.0%	100.0%	97.0%	96.5%	98.5%
N=200	97.7%	92.0%	96.5%	100.0%	100.0%	97.0%	94.0%	99.5%
N=250	99.4%	91.0%	98.0%	100.0%	100.0%	95.5%	94.5%	98.5%
Stochastic Method, $W = 0.1$								
N=10	97.7%	92.0%	90.0%	98.0%	97.0%	96.5%	93.5%	93.0%
N=25	97.1%	95.0%	93.0%	100.0%	99.5%	97.0%	96.0%	97.0%
N=50	97.7%	97.0%	98.0%	100.0%	100.0%	98.5%	98.0%	99.0%
N=100	98.3%	97.0%	100.0%	100.0%	100.0%	97.0%	97.5%	99.5%
N=150	97.7%	100.0%	99.5%	100.0%	100.0%	99.0%	98.5%	100.0%
N=200	97.7%	99.0%	99.0%	100.0%	100.0%	98.5%	97.0%	100.0%
N=250	99.4%	98.0%	99.5%	100.0%	100.0%	97.5%	97.5%	100.0%

Table 4.8: Frequency with which solutions to the assignment problem are Pareto efficient in various problem contexts across all four dimensions of analysis: average passenger delay, late arrival rate, average arrival time variability, and total vehicle travel distance.

investigated the frequency with which the stochastic method's assignments Pareto dominated those of the benchmark method (corresponding to the lower left quadrant in Figure 4-10). These results are displayed in Figure 4-11.

Although the stochastic method is able to find Pareto efficient solutions more frequently than the benchmark, in many cases only a relatively small number of its solutions Pareto dominate the benchmark method's. The upper bound of how frequently Pareto dominant solutions could be found is equal to the frequency with which the benchmark method fails to find a Pareto efficient solution, which ranges between 10 and 40% depending on the problem context. We can therefore evaluate the capacity of the stochastic method to find such solutions when it is possible to do

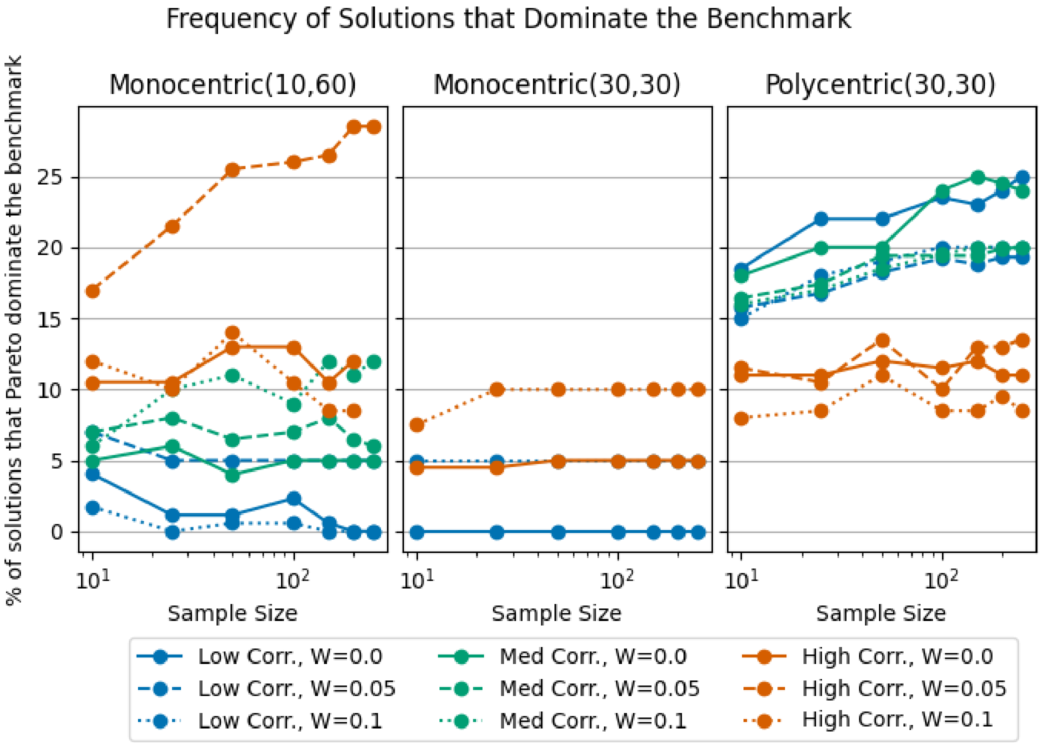


Figure 4-11: Frequency with which solutions to the stochastic assignment problem Pareto dominate the corresponding solutions produced by the benchmark method.

so by evaluating the proportion of non-Pareto optimal solutions to the benchmark method that are Pareto dominated by the stochastic methods implemented. The results of this analysis are shown in Figure 4-12.

The results shown in Figures 4-11 and 4-12 indicate significant variability in how effective the stochastic method is in finding solutions that Pareto dominate the benchmark depending on the operational environment and the optimization parameters used. For the Monocentric(10,60) network, we see that on low and medium correlation networks it is rare to find a stochastic assignment solution that Pareto dominates the benchmark solution, and many stochastic assignment methods do not result in improved performance across all metrics relative even to inefficient benchmark solutions. On the other hand, high correlation environments lead to much higher rates of Pareto domination, and with high sample sizes over 85% of suboptimal benchmark solutions are clearly improved upon by the stochastic method.

For the Polycentric(30,30) network, however, high correlation scenarios appear

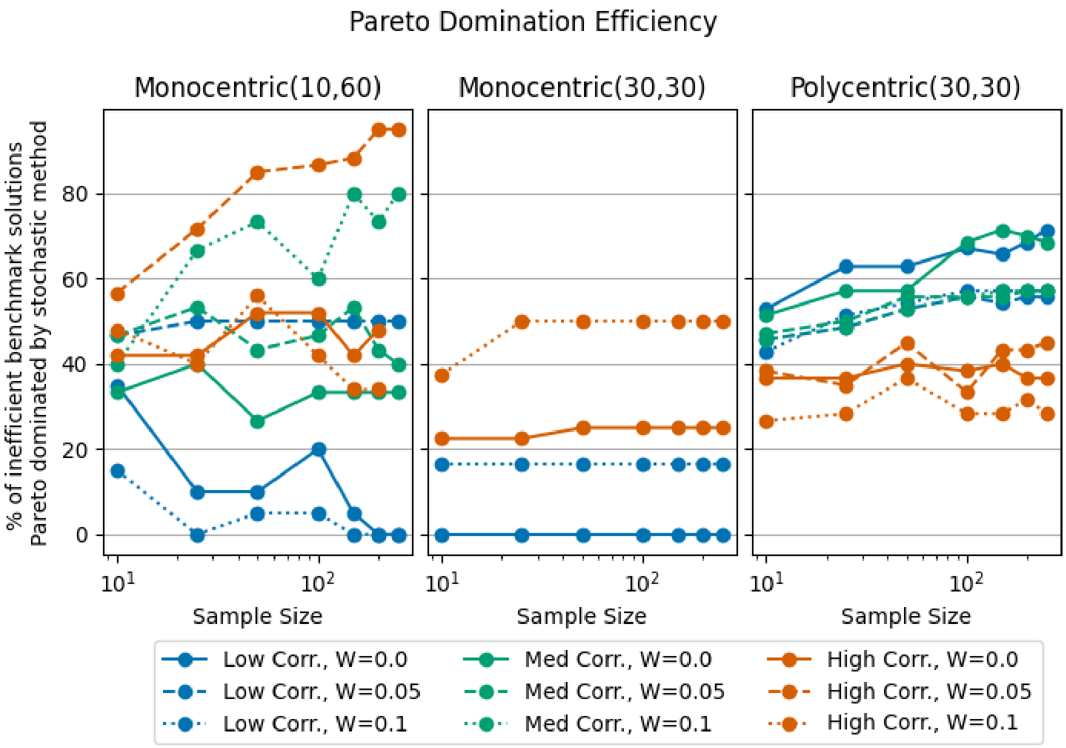


Figure 4-12: Frequency with which solutions to the stochastic assignment method Pareto dominate the suboptimal solutions provided by the benchmark assignment method.

to lead the stochastic assignment method to produce fewer solutions that Pareto dominate the benchmark than do low and medium correlation scenarios. This would appear to indicate that scenarios featuring highly correlated travel times result in greater tradeoffs between average delay and other performance metrics, meaning that minimizing this average delay is more likely to result in worse performance in other areas, corresponding to solutions in the upper left quadrant of Figure 4-10. An objective function which places greater weight on these other factors would likely result in a higher rate of finding Pareto dominant solutions over the benchmark. Indeed, in our results we find that incorporating total travel distance into the objective function increases the rate of Pareto dominant solutions for the Monocentric(30,30) network.

On this last point, however, we note that this behavior is not consistent across all scenarios investigated. For the Polycentric(30,30) network, we find that placing

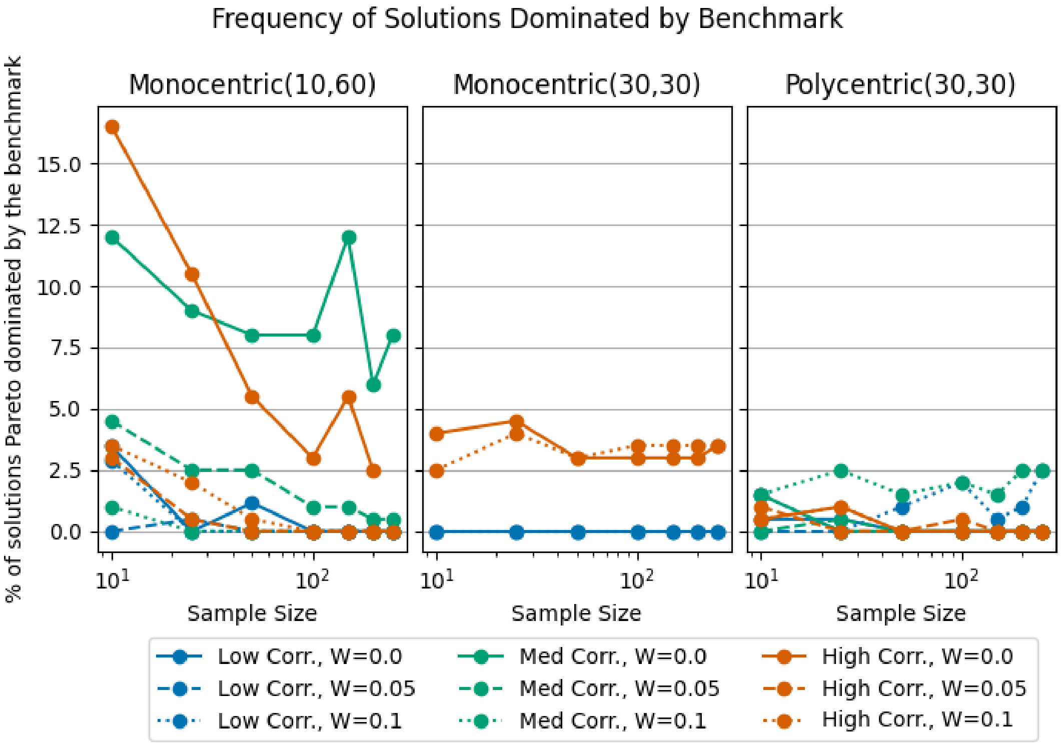


Figure 4-13: Frequency with which solutions to the stochastic assignment method are Pareto dominated by the corresponding solutions produced by the benchmark method.

higher weight on total travel distance in the objective function results in decreased effectiveness in finding Pareto dominant solutions compared to solely minimizing delay. On the Monocentric(10,60) network, the impact of this objective function parameter appears to be small with the exception of the highly correlated scenarios. This is surprising, as we would expect that jointly optimizing for two of the four performance metrics would result in more well-rounded solutions than optimizing for only one of the four, and that these well-rounded solutions would be more likely to perform at least as well as the benchmark across all four metrics. This may indicate that solutions that minimize average delay are more likely to be ones that minimize other metrics of interest compared to solutions that minimize total travel distance, or that the benchmark solutions are more effective in minimizing total travel distance than they are at minimizing average delay.

Finally, we are also interested in any downside of using the stochastic method that may result in solutions which are Pareto dominated by the benchmark method's. Figure 4-13 displays how frequently solutions of this type occur (corresponding to the upper right quadrant in Figure 4-10).

We observe that rates of these types of solutions are relatively low in most situations, with the exception of the Monocentric(10,60) network. As well, despite the greater rates of Pareto dominated solutions for the Monocentric(10,60) network, the only situations which result in higher rates of Pareto dominated solutions than Pareto dominant solutions for the stochastic method are medium and high levels of travel time correlation when $W{=}0$, although with high correlation this trend is reversed with larger sample sizes. This demonstrates strong performance of the stochastic method in avoiding poor solutions in many situations, indicating that even when Pareto efficient solutions are not found they still result in some improvement in at least one performance metric relative to the benchmark.

We also observe a strong trend in reducing frequency of Pareto dominated solutions as the sample size used by the stochastic method increases. This indicates that using a small sample size such as $N{=}10$ or $N{=}25$, although it may still frequently find Pareto efficient solutions, also has a substantial risk of finding very poor-quality solutions. Using a sample size of 100 or greater appears to minimize this risk, with diminishing returns past this point, though substantial variability still exists. The relatively poor performance of the stochastic method in the Monocentric(10,60) network relative to both of the square shaped networks may indicate that the stochastic method faces more substantial difficulty due to the elongated network shape.

Given that minimizing average delay is the main focus of our novel stochastic assignment method, the inclusion of this factor in the Pareto efficiency analysis may distort the findings because most assignments that minimize passenger delay will be on the Pareto frontier. To address this potential criticism, we further examine the solution efficiency of the stochastic method when evaluated using only the other 3 metrics of system performance: late arrival rate, average arrival time variability, and total travel distance. Table 4.9 presents the results of this analysis in the same format as 4.8. Only the results for the stochastic method using $W = 0$ are shown to isolate

Frequency of Pareto efficient solutions without average delay								
Sample Size	Monocentric(10,60)			Monocentric(30,30)		Polycentric(30,30)		
	Low Corr.	Medium Corr.	High Corr.	Low Corr.	High Corr.	Low Corr.	Medium Corr.	High Corr.
Benchmark								
N/A	82.0%	60.0%	75.0%	65.0%	80.0%	55.0%	55.0%	65.0%
Stochastic Method, $W = 0$								
N=10	72.8%	69.0%	59.5%	84.5%	86.5%	84.0%	80.5%	78.5%
N=25	72.8%	68.0%	68.5%	88.0%	91.5%	83.5%	80.0%	79.5%
N=50	73.4%	66.0%	74.0%	90.0%	94.0%	84.0%	82.5%	83.5%
N=100	71.7%	73.0%	78.5%	90.0%	94.5%	82.0%	81.5%	87.0%
N=150	72.7%	72.0%	73.5%	90.0%	95.0%	85.0%	83.0%	86.5%
N=200	71.5%	73.0%	75.5%	90.0%	95.0%	84.5%	80.0%	86.0%
N=250	72.7%	73.0%	79.0%	90.0%	95.0%	83.5%	81.5%	88.5%

Table 4.9: Frequency of Pareto efficient solutions when evaluated on performance outside of average passenger delay.

this analysis to the case when the stochastic method is solely optimizing for a metric not included in the Pareto analysis.

Overall, the frequency of Pareto efficient solutions found by both the benchmark method and the stochastic method are lower. The stochastic method sees a greater reduction in this frequency, as may be expected because its objective criterion has been removed from the analysis. However, we still see that the stochastic method is able to outperform the benchmark in the majority of cases considered, and often by significant margins. However, this does not hold true for the smaller Monocentric(10,60) network, where for both low correlation scenarios and several sample sizes in high correlation scenarios, the stochastic method finds fewer Pareto efficient solutions than the benchmark across the three metrics of interest besides average delay. These results indicate that the stochastic method is able to more effectively reach the efficient frontier of travel time variability, late arrival rate, and total travel distance even when none of these metrics are considered in the method's objective function.

As with the previous section, we are also interested in how the observed results may differ in regimes where there exist more passengers than vehicles. Using the results of the 8-request experiments with 6, 8, and 10 vehicles, we employ a similar analysis to

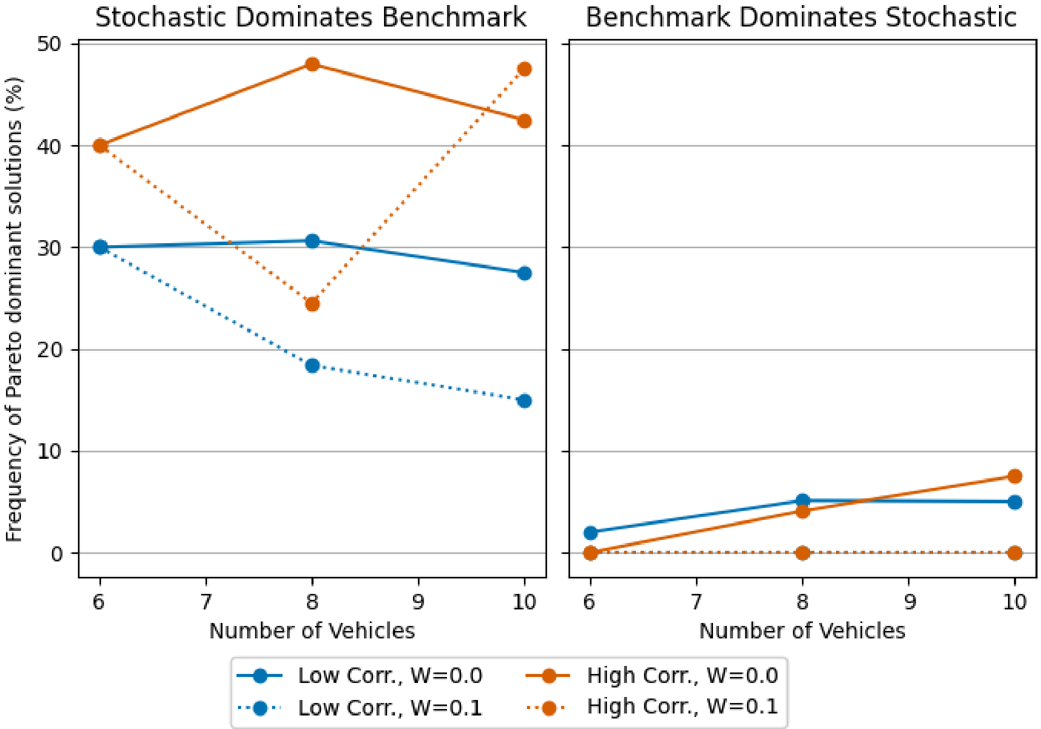


Figure 4-14: The frequency with which the stochastic assignment method produces solutions which Pareto dominate the benchmark method's solutions (left) and with which the benchmark method produces solutions which Pareto dominate the stochastic method's solutions (right) for experiments with 8 requests and 6, 8, or 10 vehicles.

understand whether the stochastic method produces solutions which are either Pareto dominant over the benchmark method's solutions or are Pareto dominated by those solutions. Note that with the larger number of passengers and vehicles, the number of potential routings becomes too large to feasibly evaluate the entire possibility space to construct a Pareto frontier, and therefore we restrict this analysis only to Pareto domination for one method's solutions of the other's.

Figure 4-14 displays the results of this analysis. We find that, with larger numbers of vehicles and passengers in the network, the stochastic assignment method is capable of delivering unambiguously better assignments in a large percentage of scenarios, while the benchmark method is only rarely able to conclusively outperform the stochastic method. In high correlation scenarios, the stochastic method obtains solutions that Pareto dominate the benchmark method over 40% of the time, while in low correlation scenarios it is somewhat lower at 15-30%. These results are much better than the corresponding results from the same Monocentric(30,30) network with only 3 vehicles and 5 requests, demonstrating the benefits of the stochastic assignment method at larger scales with a greater number of possible routes available to the operator.

Meanwhile, the benchmark method is able to obtain solutions that Pareto dominate the stochastic method's solutions less than 10% of the time in all cases. However, we observe that increasing the number of vehicles in the network from 6 (fewer than the number of requests) to 8 (at parity with the number of requests), and again from 8 to 10 (exceeding the number of requests) increases the frequency at which these types of solutions are found by the benchmark method. This supplements our findings from earlier that the advantages of the stochastic method, while still present, are reduced in cases where there is excess available supply.

4.4 Discussion

The results of the experiments described in the previous section provide several insights of relevance to dynamic ridesharing system assignment in both research and practice. At a high level, these results demonstrate the importance of consideration of stochastic travel times in the context of dynamic ridesharing system performance. Compared to the benchmark deterministic method, the stochastic formulation we propose is able to attain a higher rate of efficient solutions as well as consistent reductions in passenger delay, variability of arrival times, and total vehicle distance traveled.

As these indicators of performance are also considered by prospective passengers who might use on-demand ridehailing services, this indicates that adoption of stochastic assignment methods in DRS operation could positively impact the willingness of travelers to use these pooled services over their exclusive counterparts. This would have several further impacts in reducing negative externalities of on-demand ridehailing operations, such as reducing the number of vehicles in the system, reducing VMT, and reducing emissions. Further exploration of these methods could be helpful in validating the results found in this research using real-world case studies, and in understanding the extent to which the improved system- and request-level performance using these methods would translate to increased passenger uptake of these systems.

Our results also demonstrate the influence of several factors relating to the operational environment of the DRS system on performance. In terms of network structure, we examine both the topographical shape of the network as well as the spatial distribution of travel times. We also explore how the degree of link travel time correlation on each of these network types relate to these results. These variables correspond to environmental factors that may differ depending on the specific operational conditions of a DRS system, including location, time of day, and weather, among many others.

We also vary two parameters of the stochastic assignment method to understand their effects: the objective weight placed on total travel distance relative to average passenger delay, and the number of samples from the network travel time distribution used in the sample average approximation approach. These parameters correspond to decisions made by the network operator that reflect the priorities and computational capacity of the DRS system.

At the lowest level, our results show that the spatial arrangement of supply and demand had the greatest influence on the absolute performance of any assignment method. Given our use of a fixed wait time window, which also influenced the time bound at requests' destination locations, the average distance between passengers and the nearest vehicle in was the strongest influence on the average delay and late arrival rate observed in each experimental scenario. As this distance increased, both average delay and late arrival rate increased as well, and this trend held across each network structure, level of travel time correlation, and optimization parameter setting tested. This trend explained a significant portion of the variance observed in system performance across the individual experimental scenarios.

Aggregating across the individual scenarios, however, our results demonstrate that each of the five factors investigated had significant impacts on the performance of the stochastic method relative to the deterministic benchmark. The greatest of these impacts came from the network structure itself, with large differences in relative performance across the three synthetic networks.

The Monocentric(10,60) network was smaller and more elongated than the others, with much of the travel oriented down its longer axis with relatively constrained variety in movements. In the real world, this may correspond to environments with geographical barriers that constrain traffic, or with only two main areas of demand for DRS service and limited routes between these areas. We find that this network structure results in smaller relative differences in performance between the stochastic and benchmark methods. In some scenarios on this network, the stochastic method appears to perform very poorly compared to the benchmark method, increasing average passenger delay and finding many solutions which are unambiguously inferior to the benchmark method's, which are very rare occurrences on other networks.

The Polycentric(30,30) network, reflective of an urban region with several spatially distributed regions of high activity, also resulted in smaller differences between the stochastic and benchmark method than the corresponding Monocentric(30,30) network with a highly concentrated activity at the network's center. The stochastic method resulted in smaller reductions in average delay, variability of wait and journey times, and total distance traveled on the network with greater spatial distribution of activity, but also resulted in smaller increases in late arrival rate.

However, the deterministic benchmark method performed the worst at finding Pareto efficient solutions on the polycentric network, and the stochastic method had the highest rates of obtaining solutions that were unambiguous improvements over benchmark solutions on this network. Although the changes in system performance are not as great, these findings suggest that the greater dispersion of activity throughout the network leads to more cases in which the stochastic method is able to improve upon all metrics simultaneously.

The correlation between link travel times on each network type also had a large effect on the relative improvement in solution quality. In scenarios with greater travel time correlation, the variance of path travel times between any two points in the network is higher, meaning that the median travel times used by the deterministic method are likely to be further away from the actual travel times observed. Although the stochastic method reduced arrival time variability in scenarios with greater link travel time correlation, we also found perhaps counterintuitively that it also creates smaller reductions in average delay and smaller increases in late arrival rate in these scenarios.

Our results also demonstrate the importance of incorporating information on the correlation between links in developing methods for ridesharing assignment in stochastic regimes. Many methods in the literature for dynamic ridesharing assume that link travel times are independent of one another, and this assumption is incorporated directly into the solution methods [16]. Given the empirical demonstrations of travel time correlations across time and space [59, 60] and the findings from these experimental results that it can significantly alter the impact of stochastic assignment methods, we recommend that future research in this area either relax those assumptions or demonstrate that they are still able to achieve effective performance in networks with correlated travel times even if the methodology assumes otherwise.

The optimization parameters under the control of the DRS service provider were also influential in the performance of the stochastic method relative to the benchmark. Increasing the objective weight placed on total travel distance predictably resulted in a decrease in this metric. On two of the networks, when the stochastic method didn't incorporate distance into its objective function it increased total distance traveled compared to the benchmark method, while this trend was reversed by all methods that placed some emphasis on this minimization. Increasing this weight resulted in worse performance in terms of average delay, however.

Using a larger number of travel time samples for the stochastic method generally, though not always, resulted in greater decreases in average delay, late arrival rates, and arrival time variability compared to the benchmark method. Additionally, higher sample sizes increased the frequency of Pareto efficient solutions until reaching 50 to 100, and often dramatically decreased the number of solutions obtained that were unambiguously worse than those found by the benchmark method. This validates the importance of a large sample size in a stochastic approach to DRS assignment in minimizing the occurrence of bad outcomes, though generally 50 to 100 samples was sufficient to obtain most of the benefits. This is promising as it limits the computational burden needed to implement these systems, although the number of vehicles and requests in the system play a much larger role on computational performance than sample size does.

Overall, we found that despite the improvements made by the stochastic method in terms of average delay, arrival time variability, and total distance traveled, it resulted frequently in a higher frequency of late arrivals for the passengers of the system in these experiments. Despite this tradeoff, the stochastic method more frequently found solutions on the optimal frontier than the benchmark method. It was also clearly capable of finding solutions which unambiguously improve upon the benchmark solution in many cases, particularly those which appear to offer a greater flexibility of choice in the assignment problem (the polycentric model network, and experiments with greater numbers of vehicles and passengers).

Despite these improvements, an operator who prioritized the minimization of late arrival rate over these other metrics may still prefer the benchmark method to the stochastic method in many situations. In deciding to use the average passenger delay rather than the frequency of delays as the primary objective of this methodology in the previous chapter, we made a hypothesis that larger magnitude delays would be more impactful in reducing passengers' trust in the service's ability to deliver them on time and therefore their willingness to use the service in the future. In the absence of real-world evidence demonstrating the dynamics of what aspects of reliability and performance most influence passengers' trust in and willingness to use DRS it remains unclear what the best choice of objective for this class of problem should be.

However, we note that of the three reliability metrics used to evaluate these experimental results, the only one independent of the constraints placed on pickup and drop-off times was the variability of arrival times. In this metric, the stochastic assignment method offered consistent improvement over the deterministic benchmark in every case evaluated. Given the capacity for operators to change these constraints artificially by making less strict promises to passengers, achieving a reduction in this reliability metric that is otherwise difficult to alter demonstrates a large benefit of the stochastic method. However, we do note that reducing late arrival rate through changing the constraints on the system is also likely to decrease passengers' willingness to use the system.

In addition, we note that the poor performance of the benchmark method in terms of obtaining Pareto efficient solutions suggest that there is potential for alternative stochastic methods developed by future research to improve performance across all four of these metrics compared to the benchmark method. Existing work most relevant to this vein of research include [16], which optimizes dynamic ridesharing assignment to minimize the frequency of late arrivals, and [46], which presents a potential objective formulation which balances this tradeoff, although it did not satisfy all of the assumptions we made in formulating the optimization method in the previous chapter. Despite the negative aspects of the specific stochastic assignment method implemented in these experiments, these results highlight the potential improvements to DRS operations that may be made through incorporating travel time uncertainty into the decision-making process.

4.5 Conclusions

In this chapter, we furthered the understanding of dynamic ridesharing assignment in stochastic travel time environments through the implementation of a thorough experimental framework which uses a synthetic network cable of generating scenarios that reflect a wide variety of urban forms and travel time distributions. We use this framework to evaluate the proposed stochastic assignment method used in the previous chapter and compare its performance a benchmark method which assumes deterministic travel times. We find that incorporating knowledge of the stochastic travel times into the assignment method yields significant benefits in average passenger delay and arrival time reliability when operating in networks with stochastic travel times, and that the degree of these benefits depends strongly on the distribution of demand in the network and the distributional forms of those stochastic travel times.

Although our experiments cover a greater set of factors in terms of network configuration and travel time distributions in comparison with previous works, these results still face several limitations. First, due to computational limitations, our experiments examine only a single iteration of the assignment method to a single snapshot of system conditions. Future work would be useful to apply this framework and these methods to an online reoptimization implementation of the method, so that the performance over a longer timeframe of operation can be examined. Although it makes logical sense that the demonstration of improved performance on a single iteration will hold when the assignment is solved repeatedly over several timesteps with new arrivals to the system, empirical verification of this hypothesis is necessary to truly understand the degree of improvement that would be attained by real-world networks.

Secondly, the network we use to evaluate the varied urban network forms and travel time distributions is a synthetic network with synthetically generated demand and travel times. Implementation of this method on a real-world network with travel time distributions fitted to empirical observations of network speeds would be helpful to validate that the findings here do indeed apply to real-world situations. In particular, it would be interesting to take examples of cities that fall more towards the monocentric side and other examples of cities that fall more towards the polycentric side and observe how the differences observed performance of this method on each of those cities correspond to the differences between the monocentric and polycentric model networks we use in this study. We also note that the implementation of this method on real world case study networks would require detailed attention to the modeling of travel times, preferably using historical travel times which can be used to generate longitudinal distributions which capture the correlations across the entire road network. We observe that speed distributions from the Uber Movement database would be a useful tool for research exploring this aspect [70].

Finally, we acknowledge that the hypothetical operator of the DRS system in these simulations may have had an unrealistically small amount of information regarding the travel time distributions on the road network. The travel time distributions sampled from allowed a full spectrum of travel time variance on each link, from heavily congested to free-flow traffic. In reality, online real-time travel time predictions are frequently able to make much more specific predictions on link travel time conditions, which likely reduces the variability in link travel times compared to the simulations presented in this chapter. However, we note that urban roadways especially can have highly variable travel times even within freeflow or congested regimes due to highly unpredictable factors such as traffic signals or interactions with pedestrians, cyclists, or curbside vehicles. Furthermore, in many real world situations there exists a large potential for network conditions to change unpredictably while a vehicle is en route to a request's pickup or drop-off location, such as a vehicle collision creating a traffic jam or an event (such as a concert or sporting event) ending and generating large amounts of traffic that would be difficult or impossible to project ahead of time. Inclusion of these factors would decrease the predictability of travel times, more closely aligning with the simulation experiments conducted in this research.

Despite these limitations, this research makes several important contributions to the research literature on dynamic ridesharing assignment methods. These experiments evaluate the performance of these methods and quantify the benefits of using a stochastic assignment method across a set of scenarios which thoroughly reflect empirical understanding of real-world travel time distributions in urban networks. These experiments include correlation between travel times on nearby links and use asymmetric distributions with heavy right tails, matching insights from real-world data. In addition, we vary network configuration and correlation parameters to understand how the benefits of implementing a stochastic method varies across different realistic urban environments.

Our findings on the impact of the various network configurations implemented illustrates the role of the capacity for meaningful choice in attaining the benefits of assignment methods that incorporate travel time uncertainty. We find that in situations where the assignment problem has a greater number of viable solutions with meaningfully different performance, the differences between the stochastic and benchmark method are greater.

This is further supported by our results on networks with greater numbers of vehicles and passengers, where we see larger changes in system performance arising as the number of actors in the system increases. Increasing the problem scale creates exponentially more possible routings and leads to greater achievable savings in average delay in particular for the stochastic method. As the number of vehicles increases to parity with the number of passengers and then exceeds that number, however, the relative performances differences decrease. In these situations, most scenarios will result in an optimal solution that assigns a sole passenger to every vehicle, reducing the number of viable solutions and once again limiting meaningful choice.

Although this result is promising in that it indicates that the stochastic assignment method's benefits increase as problem scale increases, this is counterbalanced by the computational limitations of this method, as discussed in the prior chapter. While the benchmark method took less than 1 second to find an optimal solution to the majority of the scenarios presented in this chapter, our stochastic assignment method took upwards of 20 seconds at minimum to do the same, with this time increasing for larger numbers of travel time samples, and this disparity only increases as problem dimension increases. Although the results of this chapter demonstrate that there is significant room to improve from deterministic methods that currently exist in the literature, there remains significant barriers to implementing stochastic assignment in practice using the method we describe. This research serves instead as an indicator that continued investment in improving and developing these methods has promise in delivering substantial improvements to operators and to users of these systems relative to current methods.

Our results indicate that stochastic methods provide the opportunity to achieve significant improvements in average passenger delay and arrival time reliability, which are important factors in travelers' decisions to use these modes, as well as reducing total vehicle distance traveled. However, we also note that these benefits often come at the expense of increasing the number of people who arrive late to their destination. Our results identify the scenarios in which these tradeoffs are most severe and under what conditions these disbenefits can be mitigated. Furthermore, we find great capacity for unambiguous improvements in performance over the deterministic benchmark methods found in the literature. These results give much more depth to the discussion on assignment under stochastic travel time by enriching our understanding of how stochastic methods' benefits depend upon the urban environment it operates within and operator choices in implementation.

Chapter 5

Conclusion

The limitations of exclusive ridehailing services as a sustainable urban transportation option have become increasingly clear as their use has increased over the past decade. Dynamic ridesharing (DRS) services offer an alternative mode which mitigates many of these downsides while preserving some unique features of ridehailing which appeal to travelers: convenience, accessibility, and flexibility. However, DRS has failed to attain the high levels of ridership needed to realize much of this potential, due in part to traveler's lack of willingness to use these services at a price point that can be profitable for operators.

In this dissertation, we identify travel time uncertainty as a key contributor to these shortcomings and investigate its impact on both the supply of and demand for DRS. This holistic, multidisciplinary research approach allowed us to identify many interactions between the two components and address a variety of connected research questions. The findings and contributions from this research highlight future areas for research to enhance our understanding of many aspects of this area. We further use this research to guide recommendations for both policymakers and DRS operators that can be used to shift ridership from exclusive ridehailing to DRS.

5.1 Contributions and Key Findings

To understand how travel time uncertainty impacts DRS demand, we conducted a survey of 1,600 Singapore residents in April 2021. In this survey, we collected data on relevant perceptions and attitudes regarding travel time uncertainty and ridehailing, and identified psychological factors to describe the variation of these attitudes across respondents. We further modeled traveler choice between exclusive ridehailing service and DRS using results from a choice experiment with several distinct presentations of travel time uncertainty information, allowing to understand how choice differed depending upon the information shown to respondents.

The results of these models indicate that passengers' attitudes regarding time uncertainty in DRS are significant influences in their decision of whether to choose exclusive or pooled ridehailing trips. In particular, trust in the time estimates provided by DRS services was one of the most significant factors in determining this choice, and yet we also found that survey respondents overall reported quite low levels of this trust. This finding clearly demonstrates the importance of acknowledging travel time uncertainty as a key component of traveler willingness to pool when using ridehailing.

Our results regarding the influence of travel time variability information on traveler choice were also interesting. When using a traditional choice task used to investigate this factor, our results indicated that greater uncertainty in arrival time led to significantly less willingness to use exclusive ridehailing. However, using a more realistic choice task which presented travelers with a range of travel times, as seen in many contemporary ridehailing booking interfaces, our model indicated that increasing the width of this time range, thereby increasing variability, was likely to increase traveler willingness to use the exclusive mode. This counterintuitive finding demonstrates the importance of information presentation in traveler choice, and potentially indicates an optimistic or risk-seeking disposition on the part of ridehailing users. Additionally, neither form of travel time variability information appeared to significantly impact traveler willingness to use DRS, potentially belying that their lack of trust in these time estimates led them to discount or disregard the provided information.

This dissertation explored two main research areas regarding the supply of DRS in the presence of travel time uncertainty. First, we examined previous methods used for DRS assignment and found a gap in the literature regarding methods that could be used for online ridehailing assignment in contexts of travel time uncertainty. The few existing methods that addressed this topic made strong simplifying assumptions regarding the form of travel time uncertainty. Addressing this gap, we formulated a stochastic optimization method to minimize average passenger delay using a sample of travel times, allowing for generalizable implementation without making simplifying assumptions on specific distributional forms of travel time uncertainty. We further investigated the computational performance of this method and developed several approaches to improve the problem's tractability and scalability. These approaches should provide helpful insight for future research which continues to push the boundary on these methods for real-time application.

We further explored the impact of travel time uncertainty on DRS supply by using simulation experiments to evaluate the performance of this new formulation relative to an existing deterministic benchmark method. We used several synthetic networks to evaluate these methods, providing control over the spatial configuration of the network and its travel demand as well as several aspects of the travel time distributions, key among them the degree of travel time correlation.

Our comparative analysis of the two methods demonstrated several significant improvements offered by the stochastic method. In small-scale simulation experiments with only 3 vehicles and 5 requests, our stochastic assignment method proved capable of reducing average passenger delay, variability of wait and journey times, and total vehicle distance traveled across a wide variety of operational conditions. Although this method increased the frequency of passenger delays by up to 12%, it was much more frequently able to find Pareto efficient solutions and solutions which provided unambiguous improvement over the benchmark method's.

The experimental findings demonstrate that meaningful degrees of choice are important to obtain large benefits from stochastic assignment methods. Improvement in performance from the stochastic method is lowest in networks that are constrained by geographical factors, in situations where all possible routes are highly uncertain and at risk of delay, and in cases with smaller numbers of vehicles and passengers.

In the real world, DRS operations in large, complex operating environments are likely to see more improvement from the implementation of stochastic assignment than those in smaller regions with less diversity of potential routes. For instance, a region where most trips flow along a single major highway where speeds are much higher than smaller local streets is likely to see less improvement than a network with major travel flows along a large number of corridors.

Overall, these results clearly demonstrate the potential for improved DRS system performance through the use of assignment methods, and likely other operational approaches, that thoroughly account for the presence of travel time uncertainty.

5.2 Recommendations

The multidisciplinary approach used in this dissertation to approach the broad topic of travel time uncertainty in dynamic ridesharing sheds light on many facets of research, policy, and operation surrounding DRS systems. In this section, we synthesize many of the insights gained from this research to make recommendations targeted towards researchers, policymakers, and operators interested in this topic.

5.2.1 Recommendations for Researchers

The investigation of how travelers make decisions in the presence of uncertainty is a difficult task, especially when using stated preferences, as there exists a great variety of ways in which information about this uncertainty may be presented to survey respondents. Furthermore, developed understanding and analytical capabilities regarding probability, uncertainty, and risk are uncommon among the general populace, exacerbating the risk of placing excessive cognitive burden on research participants or otherwise altering their responses in ways which distort their reported preferences.

Our research findings suggest that the manner in which information on the variability of ridehailing trip journey time is presented to our survey respondents had a significant influence on their reported decisions to use pooled or exclusive ridehailing. Presenting users with a choice task which presented this information in a manner more closely matching the user interfaces of real-world ridehailing services resulted in significantly different findings compared to choice tasks which used a more traditional approach to presenting this information. These findings demonstrate the importance of information presentation on the preferences expressed by survey respondents, and also reveal the limitations associated with applying traditional survey techniques which do not closely resemble real-world situations and overtax the analytical capabilities of many respondents.

With the insights gathered from these findings, we therefore recommend any researchers conducting stated preference experiments on traveler decision-making to carefully consider the selection of choice tasks given to respondents. Prioritizing realism and ease of interpretability for respondents allows for clearer analogies to be drawn between research findings and any analogous real-world decisions made by like-minded travelers. At the same time, utilizing traditional approaches alongside new, more realistic methods allows for the comparison between results and insights into the shortcomings of these traditional approaches. When possible, we recommend that both approaches be conducted in parallel, as we have done in the research presented in Chapter 2, to provide the greatest scope of insights.

In addition, this research highlights the importance of incorporating travel time uncertainty in evaluating methods for DRS or other transportation modes, even for those methods that do not explicitly incorporate uncertainty information. Uncertainty in travel times is a fundamental property of traveling by road in urban systems due to the abundance of random factors such as traffic signals, pedestrians, vehicles using the curb, and so on. Without accounting for the impacts that this inherent uncertainty has on the supply of and demand for transportation, incomplete or incorrect conclusions may be reached.

For instance, in the research that formulates the assignment method that we use as the deterministic benchmark, the simulation environment used to evaluate its performance uses fixed travel times [7]. Given our findings in comparing this method to the stochastic method, this evaluation likely overlooked several issues that would arise in practical application of that method in an environment with stochastic travel times. Even without altering the assignment method to consider travel time uncertainty, evaluating its performance in a simulation environment with stochastic travel times would help understand and reveal the magnitude of these issues.

Beyond just the inclusion of travel time uncertainty, simulation approaches to evaluate urban transportation system performance should also be deliberate regarding any choices made in constructing the scenarios used. Our findings demonstrate the importance of many factors to the outcomes of DRS system performance: the physical layout of the operating region, the degree of correlation between travel times on nearby links, and the spatial distribution of supply and demand. In the absence of detailed knowledge on link travel time distributions in a network of interest and their correlation with nearby links, any research aiming to evaluate DRS performance should strive to model performance across a variety of network conditions informed by theoretical understanding of urban transportation systems.

Furthermore, the cross-cutting nature of this research highlights the capacity for improvement in the connections made between methodology for DRS operations and understanding of passenger preference in using these systems. Studies which aim to model the performance of DRS systems or develop new methods for assigning vehicles to passengers, routing vehicles within the network, or other such problems often make simplistic assumptions regarding how passengers respond to system performance – for instance, establishing a deterministic threshold for allowable delay. While clearly appealing from the perspective of problem definition and tractability, these assumptions overly simplify the complex dynamics underlying travelers' decisions in whether to use these services, thereby running the risk of leading to results which do not bear out in real world situations.

Future Research on Supply-Demand Integration

This dissertation considers thoroughly the impacts of travel time uncertainty on both the supply and demand of DRS service using a multidisciplinary approach. Despite several theoretical connections being drawn across the two fields, however, the integration of the understanding gained is left as a topic for future research exploration. In this section, we outline a potential agenda for future research connecting travel time uncertainty in supply and demand of DRS, and describe methods that could be used to investigate the cross-cutting questions that arise from this dissertation.

Future research combining both supply and demand for DRS while incorporating travel time uncertainty into its methods can answer questions relating to interactions between the two that this dissertation is unable to provide concrete answers to. While we demonstrate the potential consequences of a stochastic assignment method on DRS system reliability and total vehicle distance traveled, this research is limited in the extent to which we can connect those effects back to a change in the potential willingness of travelers to use these systems.

Future research integrating these concepts can model the adoption of DRS, and the choices of ridehailing users between exclusive and pooled service specifically, under travel time uncertainty. These models can then be used to evaluate the consequences of operator decision-making processes, such as different assignment methods, or policy interventions, such as road pricing, on DRS system performance and/or the broader urban transportation network as a whole.

Extending the simulation environment used to evaluate the methods in Chapter 4 of this dissertation can enable this broader modeling approach. A diagram depicting a potential framework for this extended simulation method is shown in Figure 5-1. The event-based simulation environment as outlined consists of one or more ridehailing fleets, each offering exclusive and/or pooled rides, operating within a road network with a stochastic arrival process that generates new trip requests within the network.

Within the simulator, which would be used to simulate the dynamics of the ridehailing systems over the course of a time period separated into discrete time steps,

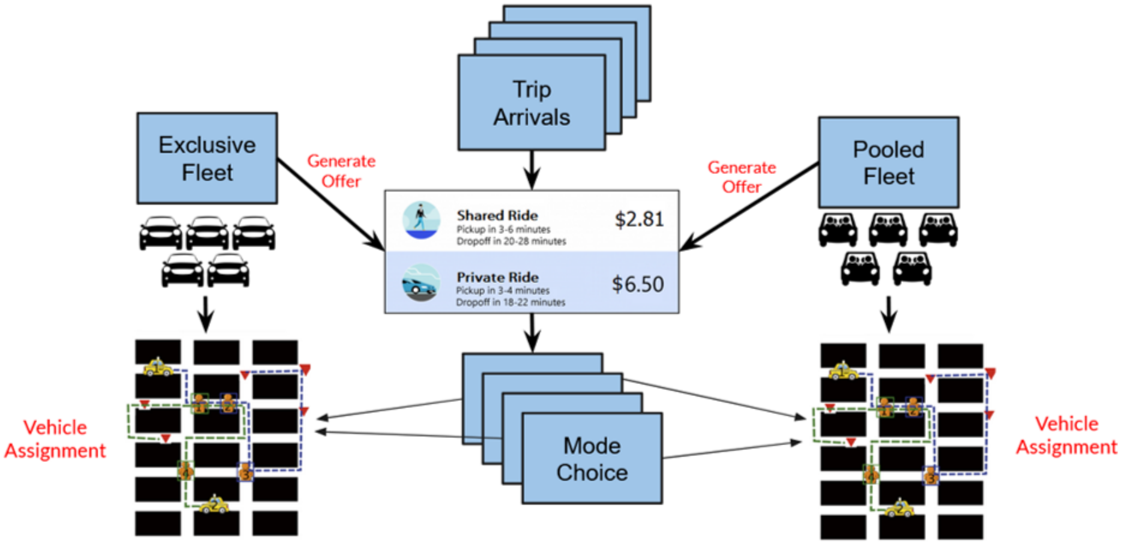


Figure 5-1: An overview of a simulation method framework that can be used to investigate system performance integrating both supply and demand for DRS under travel time uncertainty.

the following process would repeat each time step. First, arrivals enter the system according to a stochastic arrival process based on demand parameters for the system. As in our simulation, each request would be associated with an origin and destination location. In addition, the requests could feature other characteristics relevant to their preferences between competing trip offers, such as sociodemographic attributes, psychological attitudes, and/or trip context information such as preferred arrival time and the relative importance of on-time arrival.

For each of these requests, each operating fleet generates an offer for that request based on its current system state and the information provided by the request to the system (likely to be only the origin and destination locations). The offer consists of the trip's fare as well as the estimated wait and arrival times, and potentially information on the variability of the trip. In their simplest form, these offers could reflect the fleet operator's best estimate as to when they could deliver the passenger to their destination along with a cost dependent only on the total distance traveled to serve them. However, many more complex behaviors could be included in this process as well, incorporating an understanding of passenger preference and intentionally adjusting the offer to take advantage of these preferences to achieve operator goals such as increased revenue.

After the offers are provided to the travelers, each traveler uses a mode choice function dependent on their individual characteristics and the trip attributes of the offers shown to them, similar to those we present in Chapter 2, to decide between them. Their decisions are communicated to the fleets, which then proceed to run passenger-vehicle assignment methods to determine the routing of each vehicle they operate. The system then proceeds to the next time step by moving vehicles through the networks according to realizations of the network travel time distribution and updating the system state accordingly.

After completion of the designated time period, system performance statistics can be recorded, such as the ridership and profit for each operating fleet, the experience of passengers on each fleet including delay, late arrival rate, and other such metrics, the total vehicle distance traveled for each fleet and the system as a whole, and so on. Optionally, passenger preferences for each system could be updated using these metrics, such as altering passengers' trust in each system based on the observed late arrival rate that system had, and the entire process could be repeated with these updated preference functions.

This simulation environment could be used to study the impact of many different processes and policies on an individual fleet's performance, or on system performance as a whole. At the individual fleet level, altering the passenger-vehicle assignment method or the offer generation method for any fleet could yield interesting results regarding the effect of these methods on ridership and performance. For the system as a whole, investigating how policies such as road pricing affected the balance of exclusive vs. shared rides and total distance traveled could yield further insights into what consequences public policy could have on these systems. Combining the two perspectives, this method could be used to identify combinations of policies and operator decision-making methods where public sector goals such as road use reduction are achieved at the same time as DRS services are able to remain profitable.

Future research combining these two segments has the potential to greatly enhance our understanding of the realistic potential for DRS adoption in urban areas. The integrated simulation approach outlined here not only serves as a potential goal for research in this field, but also generates other ideas for potential research topics. Even without creating an overarching simulation, methods for offer generation that incorporate passenger preference can be formulated for integration into this simulation structure separately. Research exploring the dynamics of how passenger preferences can be updated based on the performance of the system in a former iteration would also serve to further our understanding of these systems as a whole, with the additional benefit of allowing a more complete feedback loop between supply and demand in this integrated simulation environment.

5.2.2 Recommendations for Policymakers

Should the share of ridehailing trips made in urban areas around the world continue to grow, increasing the average occupancy of ridehailing vehicles and reducing the vehicle distance traveled per ridehailing passenger should be key policy goals for policymakers looking to improve the environmental and social sustainability of urban transportation systems. Encouraging travelers to shift from use of exclusive ridehailing services to DRS serves as a way to accomplish these objectives while allowing travelers to capture many of the benefits of ridehailing which have prompted them to adopt these services in the first place. While generating this mode shift has proven difficult in practice, this research sheds light on some important, previously overlooked aspects which may be important in accomplishing these goals. These recommendations are presented with the caveat that the legal capabilities and jurisdiction of specific transportation policymaking bodies play a large role in determining what actions they are capable of taking.

The major impact of passenger trust and confidence in DRS on-time performance on their willingness to use these services highlights the importance of actions which can be taken to improve this trust. An example of a regulation which could be implemented to achieve this goal would be to implement requirements for on-time performance across all ridehailing services, along with associated penalties for operators who fail to achieve this performance threshold. Operators would have the decision of altering the methods used to generate arrival time estimates, assign and route vehicles, or both to achieve the required on-time performance.

From the passenger perspective, a guarantee on the likelihood with which a passenger could expect to arrive within the time estimate they are shown upon booking a trip, with oversight from a regulatory agency, would likely serve to increase travelers' trust in these services, which our findings demonstrate would serve to substantially increase their likelihood of choosing to use DRS over exclusive ridehailing. Even without the threat of penalties, merely collecting and providing data on on-time performance to the public would likely create similar incentives among ridehailing operators and engender greater trust in ridehailing users.

While much of the optimism surrounding DRS suggests that it offers some of the benefits of exclusive ridehailing - flexibility, convenience, and door-to-door service - with several benefits of transit, such as greater space efficiency and lower fares. However, it appears that the reluctance to use dynamic ridesharing arises due to travelers' perceptions that it provides the downsides of both services instead - lower certainty of on-time arrival than exclusive ridehailing, with the additional cost of sharing the ride with strangers, but also a significantly higher fare than transit. For DRS to overtake exclusive ridehailing as an on-demand transportation platform preferred by urban travelers, countering these implicit assumptions of travelers may play a large role.

This could involve shifting the default form of on-demand ridehailing from exclusive to pooled, viewing exclusive service as a premium form of travel where you pay to have the exclusive right to the vehicle while you are in it. This possibility stands in contrast to much of the current perception and marketing around DRS, offering it as a cost-conscious method of travel where one must share the vehicle with others in order to save money. Alternatively, if the price differential between the two is large enough, travelers will at some point be more willing to use pooled services despite its perceived downsides as the cost benefit to them outweighs those downsides.

Because of this, another potential approach that transportation policymakers could use to provide further incentives for passengers to use DRS as opposed to exclusive ridehailing would be implementing measures that accurately reflect the external costs of the exclusive mode. Methods such as road pricing, fees for exclusive ridehailing, or subsidies for pooled ridehailing are among the possible ways to achieve this goal.

The rationale behind these measures is that reductions in road use offered by DRS provide positive externalities to the transportation system as a whole, freeing up road capacity for other travelers to arrive to their destinations more quickly. Based on the lack of DRS adoption to date, the current economic benefits from fare savings on DRS are insufficient for many passengers, while the system is not efficient enough to justify reducing fares further. Although not all trips are substitutable for pooling (time sensitive trips, errands requiring transporting bulky objects, etc.), there likely exist many exclusive trips that would have been substituted for pooled if the price difference were marginally higher to reflect the negative externalities of exclusive ridehailing.

An important caveat to this discussion is that, while DRS service provides definite sustainability benefits when compared to exclusive ridehailing, the impacts of these services on multimodal urban transport systems is much less clear. Convincing as many on-demand rides as possible to use pooling may achieve significant benefits in terms of increasing transportation system efficiency and reducing emissions and energy use. However, for many dense urban regions a transportation system consisting entirely of pooled ridehailing will not be feasible as vehicles that seat only 4 to 6 passengers are still a relatively inefficient use of space relative to modes such as walking, biking, bus, or train.

Our research indicates that many travelers are hesitant to adopt DRS and that it is very difficult to improve its performance to meet travelers' needs. Furthermore, should these improvements occur, it is still relatively inefficient compared to these other modes. While the accessibility and convenience benefits of exclusive ridehailing are clear, DRS may have the worst of both worlds: less convenience and worse reliability as an on-demand door-to-door transportation option, without the space efficiency and other positive externalities of active modes and mass transit. For many policymakers, this may make the cost-benefit analysis for shifting travelers from exclusive ridehailing to DRS look less appealing than investing those same resources in improving transit service or bicycle infrastructure, for instance.

5.2.3 Recommendations for Dynamic Ridesharing Operators

It is unclear to what extent the incentives of DRS operators align with policymaker desires to improve the sustainability of transportation systems and reduce traffic, which corresponds with increasing vehicle occupancy, reducing travel distance per passenger, and shifting ridership from exclusive ridehailing to DRS. Many DRS operators offer a competing exclusive service alongside any pooled offerings. This means that a passenger's choice between exclusive and pooled service doesn't impact the operator's ridership numbers, although they may have disparate effects on revenues, costs, and profits. Furthermore, existing research indicates that pooled ridehailing trips more frequently act as substitutes to other modes such as public transit than to exclusive ridehailing [21], meaning that ridehailing operators may view DRS services as a way to attract ridership to their products from other modes, rather than as a means to reduce the negative externalities of the exclusive ridehailing services they also operate.

Nevertheless, increasing the reliability and effectiveness of DRS service is likely important to most DRS operators, whether as a response to actions such as those in the previous section taken by regulators to create better aligned incentives for operators, a genuine interest in creating more sustainable transportation services and attracting riders to use pooling over exclusive trips, or, more cynically, to improve these systems' ability to attract travelers away from modes such as transit. This research should prompt DRS operators to consider the development of methods that improve the reliability of these services and increase their appeal to travelers, especially in the aftermath of the COVID-19 pandemic which led to widespread suspensions of service in many markets.

The simulation experiments in Chapter 4 provide clear evidence regarding the potential benefits of assignment methods which incorporate travel time uncertainty in reducing passenger delay. The specific formulation for this problem described in this dissertation has computational challenges which limit its applicability in practical situations, and results in solutions which increase the frequency of passengers arriving late despite their other advantages. However, our findings demonstrate the shortcomings of existing deterministic methods in obtaining robust solutions, and a large opportunity for improvement exists for any operators that can implement scalable, tractable, stochastic assignment methods.

In addition, better understanding of travel time variability can be incorporated into DRS and ridehailing operations in many other ways. Knowledge of passengers' valuation of travel time variability for ridehailing can be incorporated into existing services or even help define new ones. By better informing passengers on the range of wait and travel times that they may experience, DRS operators may be able to increase passenger trust in DRS, and therefore their willingness to pool, without altering the assignment method itself at all. This could be done through providing more accurate time ranges as well as information validating this accuracy and communicating what level of confidence a passenger can have.

For a more ambitious example, a DRS operator could explicitly use travel time variability as a service parameter that the passenger has control over. Passengers could indicate the amount of travel time variability and/or the extent of delay that they are willing to tolerate to the operator when booking the ride. For instance, the operator could facilitate this by providing multiple trip options differentiated by arrival time range width, with the greater variability options offering cheaper fare.

Under this system, travelers with less time sensitive trips or otherwise lower values of travel time variability could save fare by taking trips that give the operator a greater degree of flexibility in creating assignments, likely resulting in a greater possibility of pooling that trip with others. Meanwhile, travelers who are willing to tolerate little delay may have an increased degree of confidence that the operator is prioritizing their quick arrival to their destination. From the operational side, however, it would preserve the possibility of pooling in convenient cases that hardly impact delay and variability, such as a vehicle with a passenger already inside being very close to the time sensitive customer's pickup location.

In this manner, multiple goals are accomplished that improve the willingness of travelers to use DRS and increasing system performance. The operator, having better knowledge of customer preference, can accurately assess the costs associated with travel time uncertainty and delays resulting from any given pooling assignment. Meanwhile, by communicating their preferences to the operator when booking, passenger trust in system on-time performance may increase while at the same time providing discounts due to the increased economies of scale from pooling rides. Whether radically redefining service provision, modifying decision-making during passenger-vehicle assignment, or simply altering the information given to passengers to better establish trust, incorporating travel time uncertainty into DRS operation has the potential to yield clear benefits to the service operator.

Bibliography

1. Wu, X. & MacKenzie, D. The evolution, usage and trip patterns of taxis & ridesourcing services: evidence from 2001, 2009 & 2017 US National Household Travel Survey. *Transportation* **49,** 293–311 (2022).
2. Rayle, L., Shaheen, S., Chan, N., Dai, D. & Cervero, R. *App-Based, On-Demand Ride Services: Comparing Taxi and Ridesourcing Trips and User Characteristics in San Francisco* Working Paper, 2014.
3. Schwieterman, J. & Smith, C. Sharing the ride: A paired-trip analysis of UberPool and Chicago Transit Authority services in Chicago, Illinois. *Research in Transportation Economics* **71,** 9–16 (2018).
4. Henao, A. & Marshall, W. E. The impact of ride-hailing on vehicle miles traveled. *Transportation* **46,** 2173–2194 (2019).
5. Tirachini, A. & Gomez-Lobo, A. Does ride-hailing increase or decrease vehicle kilometers traveled (VKT)? A simulation approach for Santiago de Chile. *International Journal of Sustainable Transportation* **14,** 187–204 (2020).
6. Liu, X., Li, W., Li, Y., Fan, J. & Shen, Z. Quantifying Environmental Benefits of Ridesplitting based on Observed Data from Ridesourcing Services. *Transportation Research Record* **2675,** 335–368 (2021).
7. Alonso-Mora, J., Samaranayake, S., Wallar, A., Frazzoli, E. & Rus, D. On-demand high-capacity ride-sharing via dynamic trip-vehicle assignment. *Proceedings of the National Academy of Sciences* **114,** 462–467 (2017).
8. Santi, P. *et al.* Quantifying the benefits of vehicle pooling with shareability networks. *Proceedings of the National Academy of Sciences* **111,** 13290–13294 (2014).
9. Tachet, R. *et al.* Scaling law of urban ride sharing. *Scientific reports* **7,** 1–6 (2017).
10. Lunden, I. Uber says that 20% of its rides globally are now on UberPool. *TechCrunch. May* **10.** Accessed February 23, 2022. `https://techcrunch.com/2016/05/10/uber-says-that-20-of-its-rides-globally-are-now-on-uber-pool/`. (2016).
11. Zheng, F., Zuylen, H. & Liu, X. A Methodological Framework of Travel Time Distribution Estimation for Urban Signalized Arterial Roads. *Transportation Science* **51,** 893–917 (2020).

12. Young, M., Farber, S. & Palm, M. The true cost of sharing: A detour penalty analysis between UberPool and UberX trips in Toronto. *Transportation Research Part D* **87,** 102540 (2020).

13. Naumov, S. & Keith, D. *Hailing rides using on-demand mobility platforms: What motivates consumers to choose pooling?* in *Proc., Academy of Management Annual Meeting* (Briarcliff Manor, NY, 2019).

14. Alonso-González, M. J., Oort, N., Cats, O., Hoogendorn-Lanser, S. & Hoogendoorn, S. Value of time and reliability for urban pooled on-demand services. *Transportation Research Part C* **115** (2020).

15. Bansal, P., Liu, Y., Daziano, R. & Samaranayake, S. Impact of discerning reliability preferences of riders on the demand for mobility-on-demand services. *Transportation Letters* **12,** 677–681 (2020).

16. Li, C., Parker, D. & Hao, Q. *Vehicle dispatch in on-demand ride-sharing with stochastic travel times* in *IEEE/RSJ International Conference on Intelligent Robots and Systems* (Prague, Czech Republic, 2021).

17. Xiang, Z., Chu, C. & Chen, H. The study of a dynamic dial-a-ride problem under time-dependent and stochastic environments. *European Journal of Optimization Reesarch* **185,** 534–551 (2008).

18. Erhardt, G. D. *et al.* Do transportation network companies decrease or increase congestion? *Science Advances* **5** (2019).

19. Hou, Y. *et al.* Factors influencing willingness to pool in ride-hailing trips. *Transportation Research Record* **2674,** 419–429 (2020).

20. Kang, S., Mondal, A., Bhat, A. C. & Bhat, C. Pooled versus private ride-hailing: A joint revealed and stated preference analysis recognizing psycho-social factors. *Transportation Research Part C: Emerging Technologies* **124,** 102906 (2021).

21. Moody, J. & Zhao, J. Adoption of Exclusive and Pooled TNC Services in Singapore and the US. *Journal of Transportation Engineering, Part A: Systems* **146,** 04020102 (2020).

22. Sarriera, M. *et al.* To Share or Not To Share: Investigating the Social Aspects of Dynamic Ridesharing. *Transportation Research Record* **2605,** 109–117 (2017).

23. Moody, J., Esparza-Villarreal, E. & Keith, D. Use of Exclusive and Pooled Ride-hailing Services in Three Mexican Cities. *Transportation Research Record* **2675,** 507–518 (2021).

24. Werth, O., Sonneberg, M.-O., Leyerer, M. & Breitner, M. H. Examining Customers' Critical Acceptance Factors toward Ridepooling Services. *Transportation Research Record* **2675,** 1310–1323 (2021).

25. Bansal, P., Liu, Y., Daziano, R. & Samaranayake, S. Impact of discerning reliability preferences of riders on the demand for mobility-on-demand services. *Transportation Letters* **12,** 677–681 (2020).

26. Gooze, A., Watkins, K. E. & Borning, A. Benefits of real-time transit information and impacts of data accuracy on rider experience. *Transportation Research Record* **2351,** 95–103 (2013).

27. Lu, H. *et al.* The impact of real-time information on passengers' value of bus waiting time. *Transportation Research Procedia* **31,** 18–34 (2018).

28. Tseng, Y.-Y., Verhoef, E., Jong, G., Kouwenhoven, M. & Hoorn, T. A pilot study into the perception of unreliability of travel times using in-depth interviews. *Journal of Choice Modelling* **2,** 8–28 (2008).

29. Alemi, F., Circella, G., Mokhtarian, P. & Handy, S. What drives the use of ridehailing in California? Ordered probit models of the usage frequency of Uber and Lyft. *Transportation Research Part C* **102,** 233–248 (2019).

30. Li, Z., Hensher, D. A. & Rose, J. M. Willingness to pay for travel time reliability in passenger transport: a review and some new empirical evidence. *Transportation Research Part E* **46,** 384–403 (2010).

31. Carrion, C. & Levinson, D. Value of travel time reliability: A review of current evidence. *Transportation Research Part A* **46,** 720–741 (2012).

32. Ajzen, I. The theory of planned behavior. *Organizational behavior and human decision processes* **50,** 179–211 (1991).

33. Moody, J. & Zhao, J. Car pride and its bidirectional relations with car ownership: Case studies in New York City and Houston. *Transportation Research Part A* **124,** 334–353 (2019).

34. Kline, R. B. *Principles and Practice of Structural Equation Modeling* 4th (Guilford Press, New York, 2016).

35. Cherchi, E. & Manca, F. Accounting for inertia in modal choices: some new evidence using a RP/SP dataset. *Transportation* **38,** 679–695 (2011).

36. Kim, J., Rasouli, S. & Timmermans, H. J. The effects of activity-travel context and individual attitudes on car-sharing decisions under travel time uncertainty: a hybrid choice-modeling approach. *Transportation Research Part D* (2017).

37. Feng, Y., Niazadeh, R. & Saberi, A. *Two-stage Stochastic Matching with Application to Ridehailing* in *Proceedings of the 2021 ACM-SIAM Symposium on Discrete Algorithms (SODA)* (2021), 2862–2877.

38. Lin, Q., Xu, W. & Chen, M. *A Probabilistic Approach for Demand-Aware Ride-Sharing Optimization* in *ACM International Symposium on Mobile Ad Hoc Networking and Computing (Mobihoc '19)* (Catania, Italy, 2019).

39. Aouad, A. & Saritaç, Ö. *Dynamic stochastic matching under limited time* in *Proceedings of the 21st ACM Conference on Economics and Computation* (2020), 789–790.

40. Lowalekar, M., Varakantham, P. & Jaillet, P. Online spatio-temporal matching in stochastic and dynamic domains. *Artificial Intelligence* **261,** 71–112 (2018).

41. Wang, W. & Xie, L. Dynamic Optimal Pricing of Ridesharing Platforms under Network Externalities with Stochastic Demand. *Complexity* **2021** (2021).

42. Simonetto, A., Monteil, J. & Gambella, C. Real-time city-scale ridesharing via linear assignment problems. *Transportation Research Part C* **101,** 208–232 (2019).

43. Bertsimas, D., Jaillet, P. & Martin, S. Online vehicle routing: The edge of optimization in large-scale applications. *Operations Research* **67,** 143–162 (2019).

44. Levin, M. Congestion-aware system optimal route choice for shared autonomous vehicles. *Transportation Research Part C* **82,** 229–247 (2017).

45. Febbraro, A. & Sacco, N. Optimization of Dynamic Ridesharing Systems. *Transportation Research Record: Journal of the Transportation Research Board,* 44–50 (2013).

46. Jaillet, P., Qi, J. & Sim, M. Routing Optimizaton Under Uncertainty. *Operations Research* **64,** 186–200 (2016).

47. Miranda, D. & Conceição, S. The vehicle routing problem with hard time windows and stochastic travel and service time. *Expert Systems With Applications* **64,** 104–116 (2016).

48. Rajabai-Bahabaadi, M., Shariat-Mohaymany, A., Babaei, M. & Vigo, D. Reliable vehicle routing problem in stochastic networks with correlated travel times. *Operational Resaerch* **21,** 299–330 (2021).

49. Çimen, M. & Soysal, M. Time-dependent green vehicle routing problem with stochastic vehicle speeds: An approximate dynamic programming algorithm. *Transportation Research Part D* **54,** 82–98 (2017).

50. Fu, L. Scheduling dial-a-ride paratransit under time-varying, stochastic congestion. *Transportation Research Part B* **36,** 485–506 (2002).

51. Long, J., Tan, W., Szeto, W. & Li, Y. Ride-sharing with travel time uncertainty. *Transportation Research Part B* **118,** 143–171 (2018).

52. Li, B., Krushinsky, D., Woensel, T. & Reijers, H. The Share-a-Ride problem with stochastic travel times and stochastic delivery locations. *Transportation Research Part C* **67,** 95–108 (2016).

53. Chen, C.-Y., Yan, S. & Wu, Y.-S. A model for taxi pooling with stochastic vehicle travel times. *International Journal of Sustainable Transportation* **13,** 582–596 (2019).

54. Schilde, M., Doerner, K. & Hartl, R. Integrating stochastic time-dependent travel speed in solution methods for the dynamic dial-a-ride problem. *European Journal of Operational Research* **238,** 18–30 (2014).

55. Vodopivec, N. & Miller-Hooks, E. An optimal stopping approach to managing travel-time uncertainty for time-sensitive customer pickup. *Transportation Research Part B* **102,** 22–37 (2017).

56. Lim, S., Balakrishnan, H., Gifford, D., Madden, S. & Rus, D. Stochastic motion planning and applications to traffic. *The International Journal of Robotics Research* **30,** 699–712 (2011).

57. Herman, R. & Lam, T. Trip time characteristics of journeys to and from work. *Transportation and Traffic Theory* **6,** 57–86 (1974).

58. Polus, A. A study of travel time and reliability on arterial routes. *Transportation* **8,** 141–151 (1979).

59. Gajewski, B. & Rilett, L. Estimating link travel time correlation: an application of Bayesian smoothing splines. *Journal of Transportation and Statistics* **7,** 53–70 (2005).

60. Esawey, M. & Sayed, T. Travel time estimation in urban networks using limited probes data. et. *Canadian Journal of Civil Engineering* **38,** 305–318 (2011).

61. Chen, B. *et al.* Most reliable path-finding algorithm for maximizing on-time arrival probability. it. *Transportmetrica B: Transport Dynamics* **5,** 248–264 (2016).

62. Codato, G. & Fischetti, M. Combinatorial Benders' Cuts for Mixed-Integer Linear Programming. *Operations Research* **54,** 756–766 (2006).

63. Susilawati, S., Taylor, M. A. & Somenahalli, S. V. Distributions of travel time variability on urban roads. *Journal of Advanced Transportation* **47,** 720–736 (2013).

64. Emam, E. & Al-Deek, H. Using real-life dual-loop detector data to develop new methodology for estimating freeway travel time reliability. *Transportation Research Board: Journal of the Transportation Research Board,* 140–150 (1959).

65. Srinivasan, K. K., Prakash, A. & Seshadri, R. Finding most reliable paths on networks with correlated and shifted log-normal travel times. *Transportation Research Part B* **66,** 110–128 (2014).

66. Kim, J. & Mahmassani, H. S. Compound Gamma representation for modeling travel time variability in a traffic network. *Transportation Research Part B* **80,** 40–63 (2015).

67. Castillo, E., Nogal, M., Menendez, J., Sanchez-Cambronero, S. & Jimenez, P. Stochastic demand dynamic traffic models using generalized beta-gaussian Bayesian networks. *IEEE Transactions on Intelligent Transportation Systems* **13,** 565–581 (2012).

68. Fosgerau, M. & Fukuda, D. Valuing travel time variability: Characteristics of the travel time distribution on an urban road. *Transportation Research Part C* **24,** 83–101 (2012).

69. Zockaie, A., Nie, Y. M. & Mahmassani, H. S. Simulation-Based Method for Finding Minimum Travel Time Budget Paths in Stochastic Networks with Correlated Link Times. *Transportation Research Record* **2467,** 140–148 (2014).

70. Technologies, U. & Inc. *Uber Movement* Accessed February 9, 2022. 2022. `https://movement.uber.com`.