Modeling and Design of Integrated Transit Systems with Strategic Passenger Behavior

A DISSERTATION

SUBMITTED TO THE FACULTY OF THE GRADUATE SCHOOL

OF THE UNIVERSITY OF MINNESOTA

BY

Pramesh Kumar

IN PARTIAL FULFILLMENT OF THE REQUIREMENTS

FOR THE DEGREE OF

DOCTOR OF PHILOSOPHY

September, 2022

Acknowledgements

The work presented in this dissertation is much richer because of many people who were part of my journey at the University of Minnesota, and they deserve a special mention.

To my advisor, Prof. Alireza Khani. You've been the best advisor, as kind and patient as you are excellent scholar and teacher. I would like to thank you for introducing me to the area of Transportation Networks and for countless discussions on various interesting problems in this area. Your immense knowledge of public transit systems has contributed a lot to my learning. Thank you for letting me freely explore interesting research ideas and supporting me through tough times, especially during the last year of this dissertation.

To my committee members, Prof. Gary Davis for helping me through my master's thesis and developing ideas for future research, Prof. Jean-Philippe Richard for valuable meetings that helped me solve design problems presented in this dissertation, Prof. Michael Levin for his valuable comments and feedback, and Prof. William Cooper for being my graduate advisor at ISyE and providing feedback on both my master's thesis and doctoral dissertation.

To the Department of Civil, Environmental, and Geo-Engineering, the courses, seminars, and student groups have helped me become a transportation engineer. I gratefully

acknowledge the department for supporting part of my doctoral studies through fellowship and travel grants. I was fortunate to be part of the Interdisciplinary Transportation Student Organization (ITSO), which taught me organizational skills and provided me with the great company of Jackie, Jack, Jhennyfer, Aaron, Alisha, Nina, and others. I remember we spent plenty of effort organizing the ITE Student Leadership Summit. I would also like to thank Tiffany for helping me through various departmental formalities.

To the Department of Industrial and Systems Engineering, I learned most of the mathematical techniques used in the current dissertation at ISyE. Discrete Optimization, taught by Prof. Jean-Philippe Richard, was my favorite course at ISyE. I would also like to thank the department for organizing the excellent seminar series that I enjoyed immensely.

To my friends at CEGE, I made some valuable friends at the Transit Lab - Jack, Jackie, Benj, Eugene, Dean, Kai, Yufeng, Ben, Alex, Kwangho, Behnam, and Ali, who made countless work in the lab far more enjoyable. I am also grateful to my friends outside the lab - Chris and Rongsheng for their wonderful company. The time we spent at TRB meetings, get-together at home, and while outings to Old Chicago was truly enjoyable.

To my friends outside CEGE, I was fortunate to have two amazing roommates - Tarun and Charlie. Tarun has been a great friend since my undergraduate. I enjoyed our daily cooking and wonderful discussions with tea. I still go to him whenever I need advice. Charlie has been a great friend who supported me during the final years of my studies. I enjoyed our movie nights and thoughtful discussions on various social issues. I would also like to thank Anuj, Alisha, Vivek, Shivam, Vineet, and Akhsit for their company.

Finally, I would like to acknowledge Metro Transit for sharing the data and funding

agencies - the National Science Foundation and the Minnesota Department of Transportation, for supporting my graduate studies.

Dedication

To my father, who died during the COVID-19 pandemic, and the rest of my family for their love, endless support, encouragement, & sacrifices.

Abstract

Experimental evidence shows that the uncertainty in travel time causes transit passengers to employ "strategies" when traveling between various origin-destination pairs. Such strategic behavior helps passengers adapt to the varying traffic conditions in the network. The current dissertation develops modeling frameworks to predict strategic passenger behavior in stochastic and time-dependent (STD) transit networks and use it for the design and long-term planning of integrated transit systems. It makes four principal contributions. First, it develops routing algorithms to describe the strategic behavior of transit and park-and-ride passengers using online information about road congestion and bus arrival at stops when traveling in a STD network. Second, to predict the average passenger flow on each link of the network, it develops schedule-based transit assignment models with online information for uncapacitated and capacitated transit networks. Third, it proposes an optimization model to design transit routes' alignment and corresponding frequencies incorporating the strategic passenger behavior. Fourth, it develops an optimization model to design an integrated Mobility-on-Demand (MoD) and transit systems to decide which transit routes to operate, the frequency of operating transit routes, and the MoD fleet size required to address the transit first-mile last-mile problem. Efficient algorithms are devised to solve the proposed models. Numerical experiments show that the park-and-ride passengers commuting from suburban regions to Downtown Minneapolis can save around 36 hours/year by employing strategic routing. The transit assignment results show complex passenger behavior as passengers consider alternative routes to avoid missing transfers and denied boarding due to congestion. Finally, the design results show a significant improvement in the congestion in the city center of the City of Sioux Falls with the introduction of the integrated system.

Contents

List of Tables

List of Figures

Chapter 1

Introduction

A public transit system is indispensable to the functioning of a big city. It transports workers to jobs, acts as a prominent engine of economic stability, promotes equity, and helps mitigate traffic congestion, accidents, and carbon emissions [6]. Providing a safe and efficient transit service is no easy task as its success (in terms of ridership) depends on the quality of mobility service it delivers to its passengers. Due to its fixed routes and schedules, limited network coverage, and uncertain waiting time, sometimes, it is less attractive to travelers in comparison to the auto mode. This attributes to the decline in its ridership and further loss of revenue. Therefore, constructing new solutions that help policymakers evaluate the current transit service and address challenges to improve it is of utmost importance.

The focus of the current dissertation is on transit network models, which describe the functioning of a transit service using nodes and links. The output of these models provides policymakers with information such as ridership of a route, cost of operation, and quality of service delivered to passengers in terms of travel time, wait time, walking time, and the number of transfers. Such information is crucial to assess and improve

the current service and propose solutions to address its network connectivity issues by designing new routes and integrating transit with other modes of transportation.

One of the important transit network models is the transit assignment model. *Transit assignment* is a network model that is conceptualized, designed, and calibrated to reflect the system-side and user-side behavior within a transit network. Specifically, it is the process of assigning passengers traveling between various geographic locations to a transit network in a way it predicts their route choice behavior. Urban planners use it to predict the impact on passenger behavior and system performance by design policies such as making alterations to transit frequencies, raising fares, or adding more transit routes so that better solutions can be tailored to address the needs of local transit users while avoiding unnecessary costs incurred from poorly informed solutions. There are two classes of transit assignment models, namely, *frequency-based* (FB) and *schedule-based* (SB) models. They are explained as follows:

1. *Frequency-based (FB) transit assignment*: FB models assume static representation of a transit network characterized using a directed graph with transit stops as nodes and consecutive segments of bus routes as links. These models are valuable for long-term planning operations such as network design, frequency design, etc.

2. *Schedule-based (SB) transit assignment*: SB models assume a dynamic representation of a transit network based on a published schedule characterized using a directed graph with instances of transit trips visiting stops as nodes and consecutive segments of trips as links. These models are valuable for short-term planning operations such as time-tabling, vehicle scheduling, etc.

In both types of assignment models, passenger behavior is described using the notion

of *strategy*. Spiess and Florian define a *strategy* as a set of rules that, when applied, allows a passenger to move from their origin to destination in a transit network [7]. In FB transit assignment, strategies are induced by the *uncertain* arrival of buses at stops that causes variable wait times to access the transit service. To minimize the average wait time, passengers select a subset of transit routes to travel between an origin-destination pair shared by several routes and board the first arriving route in that subset. In SB assignment, strategies are induced by the limited capacity of vehicles that causes *denied boarding* of some passengers. To minimize the average travel time, passengers select a subset of transit routes to travel between an origin-destination pair and board the first arriving vehicle that has available capacity.

As compared to FB assignment, there has been little research on SB assignment models incorporating the effect of uncertain arrival of buses at stops on passenger route choice. Moreover, advances in intelligent transportation systems such as Variable-message signs (VMS) on roads or Automatic Vehicle Location (AVL) technology installed in transit vehicles allow us to gather real-time information about the current state of the transportation network. Transportation agencies share online information about the congestion in the network through web-based and cellphone applications to help passengers make better choices and improve the overall travel time by re-evaluating their route choice. The research presented in this dissertation evaluates the effect of such real-time information on passenger route choice behavior in transit networks as well as integrated auto and transit networks using the park-and-ride mode. It is particularly important for the cities in the United States that provide transit service with a published schedule where such understanding of passenger behavior will help assess and improve the quality of service to attract more riders.

The FB transit assignment model can aid in the long-term planning of transit systems. The dissertation explores how to leverage the understanding of strategic passenger behavior obtained from the FB assignment model in designing new transit routes and their frequencies. Furthermore, it develops a FB assignment model to predict passenger behavior in transit systems integrated with emerging technologies such as Mobility-on-Demand (MoD) service to move passengers from low population density areas without transit service to high-density areas offering high-frequency services. The understanding of such behavior is important for the design of integrated transit systems.

1.1 Motivation

The general motivation behind pursuing the research in this dissertation is to understand the role of uncertainty due to transit vehicle arrival at stops on passenger route choice and use that understanding in designing integrated transit networks. The specific motivation behind the research presented in this dissertation is outlined below:

1. Current SB assignment models assume the timely arrival of buses at stops. However, in reality, bus travel time is subject to uncertainty due to road congestion (since buses use the same right of way as cars), traffic signals, inclement weather, varying dwell times, and maintenance disruptions. This uncertainty causes early/late arrival of buses at stops, which results in the possibility of missing transfers by passengers flowing in the network. Moreover, to avoid extra waiting time caused by missing transfers, it is common for passengers to use online information about the bus arrival time at different stops to make adaptive decisions *en-route* in this stochastic and dynamic system. The current SB assignment models fail to capture the adaptive passenger response to unreliable service, which causes an inaccurate estimation of passenger wait time and passenger loads

on various transit routes.

2. In the United States, several freeways with and without High-Occupancy Toll (HOT) lane connect the residential suburbs to Central Business District (CBD), where many people commute for work and education. Multiple express bus routes connect the suburban areas to the CBD along these facilities, with relatively high speed but low frequency. These routes are mostly supplemented by park-and-ride facilities alongside the freeways and give commuters the choice to transfer to transit mode at different points. On less congested days/times, one may continue driving to the destination, while on more congested days/times, parking at a park-and-ride location at a midway point and taking transit for the remaining part of the trip can lead to shorter travel time or cost. When real-time information on congestion level and waiting time of transit routes are provided to users using variable message signs along the freeway corridor and routing applications respectively, users can change their decisions *en route* considering current congestion level and short term traffic predictions. Under these conditions, understanding and modeling users' park-and-ride decisions become a difficult and important research question to assess the effectiveness of park-and-ride mode. Existing studies lack an understanding of adaptive passenger behavior in such a multimodal framework.

3. It is common to design transit routes and their corresponding frequencies based on modeling the transit operations and ignoring passenger behavior or not capturing it properly. For example, most studies on transit network design assume that passengers take the *a priori* shortest path and ignore the effect of strategic passenger behavior due to the uncertainty of transit vehicle arrivals at stops.

4. The limited network coverage makes it difficult or sometimes impossible to access transit service in some areas. This inaccessibility problem is also known as the *first mile/last mile (FMLM) problem for transit*. It is commonly faced by travelers commuting from low-density areas where transit service is not available or less frequent because of the economic in-viability of providing such service. Mobility-on-demand (MoD) service can provide fast and reliable mobility in low-density areas (i.e., by providing a first mile/last mile service) where it is difficult to provide fixed-route transit service due to economic inviability. This can help reduce congestion and carbon emissions in the network, improve the mobility of travelers, and reduce the overall cost of providing transit service. Existing studies lack understanding of passenger behavior in multimodal systems with uncertain arrival of MoD and transit service at a requested location. Further, the effect of passenger behavior on the design of integrated transit and MoD service has also not been explored.

1.2 Problem statement and research objectives

The previous section stated that the uncertain transit vehicle arrival at stops induces strategic behavior among transit passengers. Although such strategic behavior has been studied in the FB models extensively, it remains unexplored in the SB models that provide a more accurate representation of the functioning of time-dependent transit services. In the same line of research, there are no studies on modeling this behavior when multiple modes of travel, such as park-and-ride mode, are considered. Furthermore, the incorporation of transit assignment in designing integrated transit systems becomes a complex optimization problem, and there are no exact solution methods in the literature. With these research gaps, the fundamental problems to be addressed in

the current dissertation are as follows:

1. Given the stochastic travel time in a transit network, how do passengers select the time of departure from their origin, boarding stop, boarding route, transfer stop, transfer route, and alighting stop?

2. Given the stochastic travel time in transit and auto networks, how do park-and-ride users select the time of departure from their origin, auto route, park-and-ride stop, transfer route, and alighting stop?

3. Given the strategic passenger behavior in transit system through a FB assignment model, how should we design transit routes and their corresponding frequencies?

4. Given the strategic passenger behavior in an integrated MoD and transit system, how do we decide which transit routes to operate, the frequency of operating transit routes, and the MoD fleet size required to serve the FMLM of trips?

To this end, the dissertation has two major objectives. First, it develops route choice models to describe passenger behavior in stochastic networks. This involves the following objectives:

1. Propose a transit passenger route choice model that incorporates the uncertain arrival of buses at stops and the use of online bus arrival information in describing strategic passenger behavior in a stochastic and time-dependent (STD) transit network.

2. Propose a park-and-ride passenger route choice model that incorporates the online information about the congestion on the road and transit network in describing strategic passenger behavior in a stochastic and time-dependent (STD) multi-modal network.

3. Propose SB transit assignment models that incorporates the strategic behavior of passengers going between various origin-destination pairs in predicting the number of passengers on every segment of the transit network.

4. Propose a FB transit assignment model that predicts the travel behavior of passengers in an integrated transit and MoD network.

The second major objective of this research is to use the strategic behavior obtained from FB assignment models in designing routes, frequencies, and fleet of various modes of transit systems. This includes the following objectives:

1. Develop an optimization model to design transit routes' alignment and corresponding frequencies.

2. Develop an optimization model to design an integrated MoD and transit system that decide which transit routes to operate, the frequency of operating transit routes, and the MoD fleet size required to serve the FMLM of trips.

3. Develop efficient solution algorithms to solve the above models.

1.3 Thesis organization

The dissertation is organized around two major objectives. The focus of chapters 3-4 is to develop route choice and assignment models and chapters (5-6) on the development of design models. The organization of these chapters is described as follows:

The next chapter (Chapter 2) reviews the literature related to the research presented in this dissertation. This includes passenger routing models in stochastic networks, passenger assignment in transit networks, design of transit networks, and design

of integrated transit systems. The third chapter (Chapter 3) develops strategic passenger routing models for transit and park-and-ride modes. The fourth chapter (Chapter 4) develops SB transit assignment models for both uncapacitated and capacitated networks. It further presents solution algorithms to solve the corresponding models. The fifth chapter (Chapter 5) develops a bi-level transit network design optimization and proposes an efficient solution method to solve the corresponding design model. The sixth chapter (Chapter 6) develops an optimization model to design integrated transit and MoD system incorporating strategic passenger behavior through a multimodal FB assignment model. It further develops efficient solution methods to solve the corresponding design model. Finally, Chapter 7 summarizes the major conclusions and contributions of this research and also presents several ideas for future research in this and related areas.

Chapter 2

Literature Review

To place the research presented in this dissertation in an appropriate context, this chapter reviews the related literature. It can be broadly categorized into four categories. First, the literature on modeling passenger route choice in stochastic transit and auto networks is presented in Section 2.1. Second, the literature on transit assignment, which is further classified into Frequency-based (FB) and Schedule-based (SB) models, is described in Section 2.2. Third, the literature on designing transit networks is presented in Section 2.3. Finally, the literature on designing integrated transportation modes is presented in Section 2.4.

2.1 Passenger routing in stochastic networks

There has been a considerable amount of work on the shortest route planning in the literature. It consists of route planning for different modes of transport such as auto [8], transit [2, 9], and park-and-ride [10]. These route planning algorithms can be classified into two categories: deterministic and stochastic shortest path problems. In deterministic shortest path, using the historical average travel time on links, a path with minimum

travel time is sought. On the other hand, a stochastic shortest path problem considers link travel time as a random variable, and a path with minimum expected travel time between an origin and destination is sought. The stochastic shortest path problems can be further divided into two categories. The first category tries to find *a priori* solution that minimizes the expected cost [11] or expectation-variance cost [12, 13], while the second category finds an online optimal solution that allows decisions to be made at various stages (recourse, adaptive, or strategic routing problem) [14].

2.1.1 Adaptive routing in auto networks

The recourse problem is an opportunity for a decision-maker to re-evaluate their remaining path based on the information obtained *en route*. Croucher (1978) seems to be the first to study this type of problem [15]. Hall (1986) specified that the least expected travel time path between two nodes in a network with travel time both random and time-dependent cannot be found using standard shortest path algorithms (such as Dijkstra's algorithm), as the optimal route choice is not a simple path, but a strategy or hyperpath in which arcs are chosen based on an adaptive decision rule [16]. Andreatta and Romeo (1988) described this problem on a network with general dependency in which different possible realizations of the network are considered [17]. Psaraftis and Tsitsiklis (1993) presented the shortest path problem in acyclic networks in which the cost of an arc is a function of an environment variable at the head node of that arc [18]. Each of these environment variables is assumed to evolve according to an independent Markov process. Polychronopoulos and Tsitsiklis (1996) presented solution methods for two models of the shortest path with recourse [19]. The first model assumes a network with possible realizations of the arc costs whereas the second model assumes independent arc costs as a random variable. A label setting algorithm was developed by Miller-Hooks and

Mahmassani (2003) to evaluate *a priori* least expected travel time path in a stochastic time-varying (STV) network by assuming independence among arc costs as well as time periods [20]. They used a similar label setting algorithm to evaluate the least expected hyperpaths for adaptive route choice in STV networks [21]. Gao (2006) also developed an exact algorithm and several approximation algorithms such as certainty equivalence, no-online-information, and open-loop feedback algorithm to find the routing policy in STV networks [22].

As several authors [17, 19] considered general spatial dependency, Waller and Ziliaskopoulos (2002) considered limited dependency in evaluating the shortest path with recourse [14]. They presented two types of dependencies in a stochastic network. The one-step spatial dependency describes the transition of an arc state to another depending on the state of adjacent arcs. On the other hand, the temporal dependency reveals the state of downstream arcs when a traveler reaches a particular node. In both cases, a labeling algorithm is presented to evaluate the online shortest path having a minimum expected length. Provan (2003) classified the shortest path with recourse problem into two cases: *Reset* and *No reset* [23]. In no reset case, if an arc is visited, then its cost becomes deterministically known upon further visits. On the other hand, in the reset case, each visit to a node is an independent stochastic trial. Provan showed that every instance of the No reset case is NP-Hard and provided a polynomial-time algorithm for the reset case. Boyles and Rambha (2016) formulated this problem as a total cost Markov Decision Process (MDP) to detect the unbounded instances in the presence of negative arc costs [24]. The stochastic version of adaptive route choice was developed by [25–27]. Gao et al. (2008) proposed an adaptive path model and an adaptive routing policy choice model for STV networks and showed that the adaptive routing policy choice model achieves lesser expected cost in comparison to the adaptive path model [25]. Gao

and Huang (2012) designed a heuristic algorithm for the adaptive routing problem for four different types of online information schemes, namely, perfect information, delayed global information, global pre-trip information, and up-to-date radio information on a subset of arcs [28].

2.1.2 Adaptive routing in transit networks

The adaptive routing by passengers in the case of transit networks is also a common phenomenon. Due to several bus options, passengers often adopt varying strategies to board the bus [29]. Spiess (1987) proposed that passengers take the first route among a set of attractive transit routes and developed a primal-dual method for finding optimal strategies [7]. Nguyen and Pallottino (1988) described this passenger behavior using *hyperpath*, which is a set of paths (a subnetwork) rather than just a single path. There are several studies such as [30, 31] which consider online routing based on the assumptions on the distribution of the headway. This problem is generally called the shortest path problem considering on-time arrival (SOTA) probability. Rambha (2016) presented a pioneering effort in the case of the adaptive transit routing problem in a time-dependent stochastic network [32]. They formulated a finite horizon MDP and presented several pre-processing ways to reduce the computational time of computing the least expected cost hyperpath. Khani (2019) developed a label setting algorithm to evaluate an online shortest path for reliable routing in the case of schedule-based transit networks considering transfer failure probability [33].

Although there has been significant effort in designing adaptive routing algorithms for auto and transit networks, none of the above studies developed a scalable adaptive routing algorithm suitable for the SB transit assignment. The literature also lacks an adaptive routing algorithm in a multimodal framework, such as park-and-ride mode,

which consists of routing in both auto and transit networks. In this research, we study the adaptive behavior of both transit and park-and-ride passengers.

2.2 Passenger assignment in transit networks

Transit assignment has attracted a lot of attention since the 1970s, and various models have emerged throughout the years. Chriqui and Robillard (1975) posed the decision problem, also known as *common lines problem*, faced by a passenger traveling between two stops served by several transit routes [29]. Spiess (1987) proposed that passengers adopt *strategies* when traveling, which is defined as a set of rules, when applied, allows a passenger to move from an origin to a destination in a transit network [7]. They proposed the first FB transit assignment model formulated as a linear program. Further, Nguyen and Pallottino (1988) formalized any strategy as a sub-network between two nodes in the transit network, known as a *hyperpath* and proposed a greedy algorithm to find the shortest hyperpaths in the network [34]. It was soon realized that current models could not predict passenger behavior in congested FB networks. Therefore, several approaches are proposed by the researchers to model congestion in the network. This includes asymmetric BPR-type function of waiting by [35] and [36], effective frequency function by [37] and [38], and failure-to-board probabilities by [39]. The FB models consider individual transit routes, which results in an approximation of accurate vehicle loads for a time-dependent transit service [40]. Therefore, *schedule-based or dynamic transit assignment models* emerged in the literature. Nguyen et al. (2001) presented a graph-theoretic framework for the SB transit network [41], Tong and Wong (1999) proposed a SB transit assignment model based on the schedule-based transit shortest path algorithm [42], and Poon et al. (2004) proposed a simulation-based assignment model with FIFO queuing discipline [43]. To model congestion into SB models, studies have

used a BPR-type discomfort function for in-vehicle links [44–46]. The main drawback of this approach is that discomfort is applied to all the passengers in a bus (both seating and standing), and the assignment results may not satisfy the strict capacity of transit vehicles. Hamdouch et al. (2004) and Hamdouch and Lawphongpanich (2008) proposed that passengers adopt strategies in a SB network when competing with other passengers for limited vehicle capacity. They proposed an assignment model based on the User Equilibrium principle [47, 48]. The logit-based strategy for capacitated SB transit assignment was proposed by [9, 40, 49]. Various studies have also used strategy-based models for capacitated traffic assignment [50, 51]. As seating and standing passenger have different comfort costs, SB assignment models have also incorporated the effect of discomfort on strategies [52–54].

Section 2.1.2 presented literature on adaptive passenger routing algorithms for navigating in the transit network. Due to several bus options, passengers often adopt varying strategies to board the bus [7, 29]. The strategies are affected by the online information, and various studies have proposed FB assignment models incorporating online information [55–58]. In the case of SB assignment, Hamdouch et al. (2014) developed an assignment model that incorporates the passengers' response to unreliable service by finding the strategies that minimize the sum of mean and variance of overall travel cost [59]. Zhang et al. (2010) model the risk-taking behavior of passengers in SB networks with random arc travel time using chance constraints [60]. Gardner et al. (2021) presented an estimation method for evaluating passenger travel time distributions in unreliable transit networks using phase-type distributed Markov chains [61]. They formulated a finite horizon MDP and presented several pre-processing ways to reduce the computational time of computing the least expected cost hyperpaths. Hickman and Wilson (1995) and Hickman and Bernstein (1995) proposed path choice models

for modeling passenger behavior of declining a bus route in favor of a faster bus route arriving at a stop based on online information [62,63]. Other approaches include model-free reinforcement learning-based SB assignment model by [64], and simulation-based models incorporating real-time information by [65] and [66].

A common approach for modeling passenger response to unreliable transit service is through evaluating strategies with the least mean-variance cost. However, this approach does not model the complexity associated with missing transfers. If buses are late, passengers miss transfers and take alternative bus routes. Therefore, the shortest path with recourse problem needs to be solved [16, 17]. The current literature lacks an analytical SB transit assignment model that incorporates both real-time information and capacity constraints in predicting the route choice behavior of passengers. There is also no efficient method that solves the current problem in a reasonable amount of time and is scalable for a large-scale SB transit network.

2.3 Design of transit networks

There is an extensive literature on modeling *Transit Network Design Problem* (TNDP) and developing solution algorithms for it. They can be grouped into two categories, namely, the studies with and without an explicit mathematical model. We focus primarily on the studies which present a mathematical optimization model for TNDP. For a comprehensive list of articles related to TNDP, we refer the interested readers to literature survey articles by [67–71], and [72].

TNDP is a complex variant of the network design problem, which was proved as an NP-Hard problem by [73]. From the standpoint of solution approaches to TNDP,

we can categorize the research articles mainly into two categories, namely, heuristic approaches [74] and exact approaches [75]. The computational complexity of exact approaches is unable to address the practical instances of the problem, whereas the heuristic approaches, unable to find an exact solution, have shown to be efficient in solving practical instances of the problem.

In this paragraph, we review some of the heuristic approaches. Mandl (1980) emphasizes the importance of optimizing a transit network to provide efficient service at lower costs [76]. They propose an iterative procedure to find the candidate routes and their associated frequencies and then improve the current set of routes based on the average transportation costs. Current et al. (1986) proposes to connect every pair of demand points using either a primary path or secondary path [77]. They present a 0-1 optimization model, which is solved using a heuristic involving both k-shortest path and minimum spanning tree algorithms. Ceder and Wilson (1986) and Baaj and Mahmassani (1991) propose optimization models that minimize the sum of the difference between the passenger route travel time and shortest travel time between every pair of demand nodes and total system travel time subject to maximum allowable load on the bus, the minimum and maximum frequency of routes, the maximum number of routes and maximum fleet available [78, 79]. They propose a route generation algorithm that improves the current set of routes based on node selection and insertion techniques. Bagloee et al. (2011) present an efficient heuristic to solving TNDP for large networks [80]. Unlike other studies, it considers various aspects such as categorization of stops, multiple classes of transit vehicles, hierarchy planning, and system capacity. They use a clustering approach to evaluate the set of stops, employ newton gravity theory to generate candidate transit routes, and then solve the TNDP using a genetic algorithm. Buba and Lee (2018) present an optimization model that minimizes the total

passenger cost and unmet demand [81]. They propose a differential evolution approach to solving their proposed model. To improve the solution of low-level heuristics, Ahmed et al. (2019) proposes a selection hyper-heuristic to solving the TNDP that minimizes the sum of passengers' and operator's cost [82].

The exact methods to solving TNDP have not received as much attention as the heuristic approaches. Borndörfer et al. (2007) presents a column generation approach to deciding transit routes and their corresponding frequencies and assigning passengers on the shortest path [75]. They could not apply their algorithm to large instances because the pricing problem corresponding to frequency variables turned out to be the longest path problem, which is NP-Hard. Goossens et al. (2004) and Steiner and Irnich (2018) propose a branch-and-cut approach with various valid inequalities to efficiently solve this problem [83, 84]. They presented real-life instances based on examples from Netherlands Railways. Marìn and Jaramillo (2009) and Mahèo et al. (2019) propose Benders decomposition approach to solving the TNDP and related multimodal network design problem respectively [85, 86].

In this paragraph, we review articles that formulates TNDP as a bi-level problem. Constantin and Florian (1995) is the first study that formulates transit frequency design as a bi-level optimization problem [87]. They include optimal strategy transit assignment in the lower level and minimize total expected travel cost of passengers in the upper level with constraints related to maximum fleet and minimum frequency of routes. They use the subgradient approach to find a local minimum. Gao (2004) presented a bi-level model for frequency optimization where they use a transit assignment model similar to the user equilibrium-based traffic assignment model [88]. They also use a subgradient approach to solve the problem after linearizing the upper-level objective function. Guan

et al. (2006) minimizes the total length of transit lines as a proxy to the operator cost in the upper level and ensures each road link is covered by at least one transit line [89]. The lower level assigns passengers to the shortest paths among the k-shortest paths. Uchida et al. (2005) and Uchida et al. (2007) present the frequency optimization model with probit-based stochastic transit user equilibrium [90, 91]. The papers employ sensitivity analysis to define linear approximation functions between the probit stochastic user equilibrium link flows and the design parameters, which are used as constraints in a single-level optimization program. Yu et al. (2010) optimizes the frequency of transit routes based on a bi-level model and solves it using the genetic algorithm [92]. Yu et al. (2015) propose a bi-level optimization model for designing a bus lane distribution plan in a multimodal network [93]. At the upper level, they minimize the average travel time of passengers, and the lower level is a multimodal assignment model. The branch-and-bound, column generation, and method of successive averages are used sequentially to solve the optimization programs at different levels. Szeto and Jiang (2014) formulate TNDP as a bi-level optimization model [1]. The upper-level problem is formulated as a mixed-integer non-linear program that minimizes the number of transferring passengers, and the lower-level problem is the optimal strategy transit assignment problem with capacity constraints. A hybrid artificial bee colony (ABC) algorithm is developed to solve the bi-level problem. Cancela (2015) also presents a bi-level model for TNDP, where both upper and lower level problems minimize the expected travel cost of passengers [3]. The lower level is the optimal strategy transit assignment model. They did not present any tailored solution procedure for the model and only presented results for the single-level formulation.

There have been many significant contributions in modeling and solving TNDP. However, several difficulties still need to be addressed. First, there is a need to assess

the effect of transfers on network design. This can be done through two approaches, namely soft transfer and hard transfer approach. The soft transfer approach minimizes or constrains the total number of passengers on transfer links, whereas the hard transfer approach ensures the existence of a path with a limited number of transfers for each origin-destination pair. Second, most studies on TNDP either ignore passenger route choice or route passengers on the shortest path in the network using multi-commodity flow constraints. However, due to the presence of common lines between various origin-destination pairs, passengers adopt strategies to travel in a frequency-based transit system [34]. Moreover, due to limited vehicle capacity, passengers experience denied boarding, which leads to increased waiting time, travel time, and discomfort. This can be addressed by solving a bi-level design problem. The capacity constraints are included in the upper level, and a frequency-based assignment model to capture the passenger behavior is included at the lower level. Doing so would indirectly avoid requiring a congested frequency-based transit assignment model. Third, currently, there is no exact solution method in the literature to solve the bi-level TNDP. This is because it is difficult to solve the current problem due to the presence of non-linearities and non-convexities in the model, along with combinatorial explosions arising from the discrete design decision variables.

2.4 Design of integrated transit systems

The passenger journeys that consist of auto, as well as transit mode, create a new mode of transportation known as *intermodal* or *multimodal* transportation. The research on modeling multimodal transportation has been an active area of research for several decades [94]. Many of these studies are focused on solving the transit FMLM problem by designing a multimodal transportation system. This includes designing a demand

responsive transit feeder service [95–102], using park-and-ride facilities [103–105], and integrating ridesharing and transit [106–111].

Recently, the studies are being focused on modeling the integration of MoD and transit service for future mobility. They can be divided into two categories: simulation-based and optimization-based approaches. Under a set of assumptions on vehicle operations and dispatching strategies, the simulation-based studies simulate the passenger flow to assess the service quality of providing such mobility service [112]. By using a four-step travel demand simulation model, Levin (2015) predicted that the transit ridership will decrease and the number of personal vehicles will sharply increase as a result of the repositioning of vehicles resulting in congestion on the network [113]. Vakayil et al. (2017) developed a simulation model that accounts for transit frequency, transfer costs, and MoD fleet re-balancing to use MoD as the FMLM solution to the transit mode [114]. Their results show that such an integrated system can reduce VMT in the network by up to 50%. Mendes (2017) developed an event-based simulation model to compare the performance of the MoD system with the light rail system under the same demand patterns, alignment, and operating speed [115]. They found that 150 vehicles with 12 passenger capacity would be needed to match the 39-vehicle light rail system if operated as a demand responsive system. Similar findings were also shown by the simulation model developed by [116]. They showed that the introduction of MoD will act as the competitor of mass transit, however, to reduce congestion and maintain a sustainable urban transportation system, it cannot replace mass transit. Shen et al. (2018) also simulate an integrated autonomous vehicle and public transportation system based on the fixed modal split assumption [117]. Using Singapore's organizational structure and demand characteristics, they propose to preserve high-demand bus routes while re-purposing low-demand bus routes and using shared MoD as an alternative. They

found that the integrated system has the potential of serving the trips with less congestion, less passenger discomfort, and economically viable service. Wen et al. (2018) included mode choice and various vehicle capacities and hailing strategies in an agent-based model to provide insights into fleet sizing and frequency of transit routes for the integrated system [118]. A few studies have used an optimization-based approach to developing an integrated passenger flow model. Salazar et al. (2018) developed a network flow model for intermodal MoD that couples the interaction between MoD and transit by maximizing social welfare [119]. Using this model, they proposed a tolling scheme for this intermodal system that helps reduce the travel time, costs, and emissions compared to the standalone vehicle mode. Liu et al. (2019) used Bayesian optimization to predict the mode choice of passengers in such a multimodal transportation system [120].

The above-cited studies show that an integrated MoD and transit system can provide an efficient mode of transportation that is sustainable, fast, eco-friendly, and economically viable. The design of such a system requires solving a *multimodal transportation network design problem* that can decide various design aspects of MoD and transit modes. Some aspects of the multimodal network design problem have been explored in a related research problem known as *hub and arc location problem* [121–123]. However, the hub and arc location problem have a major limitation of not being able to capture passenger behavior in the transit network. Recently, a couple of studies have proposed models for the transit network design in the context of integrated MoD and transit system [124–126]. Pinto et al. (2020) develops a bi-level optimization model to design a transit network integrated with MoD service [125]. The upper-level optimization problem modifies the frequency of the transit routes and determines the fleet size of MoD service and the lower-level model simulates the passenger trajectories based on a simulation-based traveler assignment model. Due to the complexity of the model,

they presented a heuristic approach to solving the current problem. Steiner and Irnich (2020) present various aspects of this problem and develop a path-based mixed-integer programming model to decide which sections of the transit routes to operate and locate the transfer stops to allow for intermodal trips in the network [126]. Due to an enormous number of possible paths in the network, they solve the current model using a branch-and-price approach.

The design of an integrated MoD and transit system is an important problem that can influence the future mobility of travelers. Recent studies have made important contributions to this complex problem but have several limitations. First, before designing the integrated system, we should understand how passengers would behave in an integrated system. It is common for studies to use the classic multi-commodity flow model to predict the behavior of travelers in the network design. This may be true if passenger trajectories are completely influenced by the mobility provider. However, this is certainly not applicable in the case of transit systems when passengers try to reduce the expected travel time based on waiting time, travel time, and fare. Second, there is a need to develop a mixed-integer optimization model that incorporates the multimodal passenger assignment and evaluates various aspects of an integrated system. We also need efficient solution methods to solve this difficult problem.

Chapter 3

Strategic passenger routing with online information

Experimental studies show that the uncertainty in transit travel time causes passengers to employ strategies when traveling between various origin-destination pairs [127–131]. A strategy help passenger adapts to the varying traffic conditions in the network. Further, the strategies are aided by the use of online information provided in form of variable message signs, radio messages, or through cellphone applications. At every node in the network, the online information provides an estimate of the travel cost on the downstream arcs in form of travel time or wait time for a transit route. It helps passengers in making better choices and improving the overall travel cost by re-evaluating their choices. For transit passengers, this includes re-evaluating their departure time, boarding stop, boarding route, transfer stop, transfer route, and alighting stop. For park-and-ride users, this includes re-evaluating departure time, auto route, park-and-ride stop, transfer route, and alighting stop. The current chapter develops models to predict the adaptive routing behavior of transit and park-and-ride users in presence of online information.

3.1 Transit passenger routing with online information

In this section, we present mathematical models of passenger routing in a SB transit network with online information. We start by describing the notations used in the current and next chapter, the process of creating a SB transit network, and a motivating example in Section 3.1.1. This is followed by the development of mathematical models of passenger routing in uncapacitated and capacitated transit networks in Section 3.1.2 and Section 3.1.3 respectively. Finally, numerical experiment results are presented in Section 3.1.4.

3.1.1 Preliminaries

Let us consider a geographical area divided into traffic analysis zones given in set Z. The passenger demand is assumed to originate from the set of origins $O \subseteq Z$ and end at the set of destinations $D \subseteq Z$. It is distributed among various groups in set G. Each group of passengers $g \in G$ is characterized by an origin $o_g \in O$, a destination $d_g \in D$, the earliest departure time from the origin t_g^{ED}, the earliest arrival time at the destination t_g^{EA}, and the latest arrival time at the destination t_g^{LA}. Let $\{d_g^{od}\}_{(o,d)\in O\times D, g\in G}$ be the demand matrix of passenger groups traveling between various origin-destination pairs in the network. For the dynamic representation of the network, let us consider T as the set of integer-valued time intervals during a day.

A SB transit network is characterized using a directed graph $G(N, A)$, where N denotes the set of nodes and A denotes the set of links in the transit network. We use the trip-based representation of a transit network [132], which is created using the General Transit Feed Specification (GTFS) data [133], released publicly by various transit agencies around the world. The probability distributions of link travel times are

calibrated using Automatic Vehicle Location (AVL) data, which provides historical bus arrival times at various stops recorded using GPS devices installed in transit vehicles [134]. In transit schedule data, we denote the set of transit stops/stations[1] as $\mathfrak{B}$, set of bus routes[2] as R, and set of transit trips[3] as K. Each trip $k \in K$ is characterized by a bus route $r_k \in R$, set of nodes[4] $B_k \subset \mathfrak{B} \times K$, sequence $\gamma_k : B_k \mapsto \mathbb{N}$ in which various stops are visited, scheduled arrival/departure time[5] $\hat{t}_k : B_k \mapsto T$ at various stops in the itinerary, and a set of possible (actual) arrival time at those stops $\tilde{t}_k : B_k \mapsto 2^T$, which is obtained from the AVL data. The probability of a bus associated to trip k arriving at node $i \in B_k$ at time $t \in \tilde{t}_k(i)$ is denoted by $\tilde{p}_i(t)$. For a well-defined probability distribution, we must have $\tilde{p}_i(t) \geq 0, \forall t \in \tilde{t}_k(i), \forall i \in B_k, \forall k \in K$ and $\sum_{t \in \tilde{t}_k(i)} \tilde{p}_i(t) = 1, \forall i \in B_k, \forall k \in K$. Let $k(i)$ and $r(i)$ be the trip and route resp. associated to transit node $i \in B$ and $w : N \times N \mapsto \mathbb{R}$ be the walking time between two nodes in the network. There are three types of links $A = A_e \cup A_v \cup A_t$ in the network, namely, access/egress links A_e, in-vehicle links A_v, and walking/waiting transfer links A_t. The access/egress links are used to access/egress transit nodes in the network, i.e., $A_e = \{(i,j) \in O \times B \mid w(i,j) \leq \delta_0\} \cup \{(i,j) \in B \times D \mid w(i,j) \leq \delta_1\}$, where δ_0 and δ_1 are the acceptable walking times to access and egress a transit stop. The in-vehicle links are transit vehicle links created using the itinerary of a transit trip, i.e., $A_v = \{(i,j) \in B \times B \mid k(i) = k(j), \gamma_{k(j)}(j) = \gamma_{k(i)}(i) + 1\}$.

Creation of transfer links

The creation of waiting and walking transfer links is more involved than other types of links. This is because we cannot consider every link between a pair of nodes in the

[1] A transit stop is a geographic location where passengers can board a bus.
[2] A transit route is defined as a set of stops with two ends between which buses run back and forth.
[3] A trip is a travel itinerary of a bus with arrival and departure times specified at different stops.
[4] Here, a node is characterized by a stop and bus trip serving it.
[5] We assume that the bus departs from the stop as soon as it arrives.

network as a transfer link. A transfer link is created between two nodes i, j if they satisfy the following conditions:

1. Routes associated to both nodes are different, i.e., $r_{k(i)} \neq r_{k(j)}$.
2. The stop associated to node i is not the first stop in $k(i)$'s itinerary, i.e., $\gamma_{k(i)}(i) \neq 1$ and the stop associated to node j is not the last stop in $k(j)$'s itinerary, i.e, $\gamma_{k(j)}(j) \neq \max_{l \in B_{k(j)}} \gamma_{k(j)}(l)$.
3. Walking time between i and j is less than or equal to an acceptable walking time limit δ_2, i.e, $w(i, j) \leq \delta_2$.

Let us denote $A_t^{'}$ as the set of links that satisfy the above conditions. The number of links in this set can be quite large. To further reduce the number of transfer links, we assume that passengers have an acceptable waiting time limit δ_3 for transfers. δ_3 is the maximum waiting time that a passenger is willing to spend to access transit service. Moreover, we assume that if the waiting time exceeds δ_3 for a passenger, then the optimal choice for that passenger is to walk to her destination. The value of δ_3 should not be too low to exclude any reasonable choice and should not be too high to create a large number of transfer links. The passenger survey data can help us calibrate this value.

In chapter 4, we propose two types of SB transit assignment models, namely, uncapacitated and capacitated assignment. The uncapacitated assignment assumes unlimited capacity of transit vehicles, whereas the capacitated assignment assumes limited capacity of them. In both assignment models, we reduce the transfer links based on an acceptable waiting time limit. Moreover, in the uncapacitated assignment, the transfers can be further reduced based on the probability of making a transfer. For example, if there are multiple transfer trips of the same transit route available and all of them

provide transfer w.p. 1, then we should only keep one trip of the route that provides the least waiting time. This is because a passenger would not likely wait for a different trip of the same transit route. However, in the case of capacitated assignment, for a passenger, we cannot evaluate the probability of making a transfer as it depends on the availability of space which further depends on the strategies of other passengers.

The steps for creating transfer links are summarized in Algorithm 1. It takes transfer links A_t' created using the above criteria and the type of assignment as inputs and outputs the final transfer links. The algorithm starts by initializing the set of final transfer links A_t as an empty set and collecting all the transfer nodes in the network. Then, for each transfer node i, we find all the transit routes that can be transferred from it. For each transferring route, we find the set of nodes associated with it (*connecting_nodes*) and sort them in the increasing order of scheduled time. After that, for uncapacitated assignment, we start creating transfer links from node i to other nodes in *connecting_nodes* starting from the one for which there exists at least one arrival time instance so that transfer can be made successfully (i.e., with positive probability) to the one for which all its arrival time instances can be successfully transferred from any arrival time instance of node i (i.e., the transfer is made w.p. 1). If we cannot find a node that can be transferred w.p. 1, then we create a walking link from i to all destinations. This is done to finish the journey of travelers who find themselves in a situation, where there is no outgoing link to move forward. In practice, if a passenger finds that there is no bus available at the stop, then they either walk or use another mode of transportation to get to their destination. We only assume walking in our routing algorithm, although, one can consider other modes of transportation. In case of capacitated assignment, we create transfer links from node i to other nodes in *connecting_nodes* for which there exists at least one arrival time instance so that transfer can be made successfully (i.e.,

with positive probability) and provide an acceptable waiting time δ_3. In case of capacitated assignment, we compulsorily create walking links to various destinations as there may not be sufficient capacity in the considered transfer options.

Algorithm 1 Creation of transfer links

procedure CREATETRANSFERSUNCAPACITATED
  Input: $A_t^{'}$, *assignment_type*
  Output: A_t ▷ Set of final transfer nodes
  (Initialize) $A_t \leftarrow \phi$
  $transfer\_nodes \leftarrow \{i \in N : \exists j \in FS(i) \text{ s.t. } (i,j) \in A_t^{'}\}$
  for $i \in transfer\_nodes$ **do**
    $connecting\_routes \leftarrow \{r(j) : (i,j) \in A_t^{'}\}$
    for $\hat{r} \in connecting\_routes$ **do**
      $connecting\_nodes \leftarrow \{j \in FS(i) : (i,j) \in A_t^{'}, r(j) == \hat{r}\}$
      Sort nodes in *connecting_nodes* in the increasing order of scheduled arrival time
      Find the first node m in *connecting_nodes* for which $\exists t^{'} \in \tilde{t}_{k(i)}(i), t^{''} \in \tilde{t}_{k(m)}(m)$, s.t. $t^{'} + w_{ij} \leq t^{''}, \tilde{p}_i(t^{'}) > 0, \tilde{p}_m(t^{''}) > 0$.
      if *assignment_type* == "uncapacitated" **then**
        Find the first node n in *connecting_nodes* for which $\forall t^{'} \in \tilde{t}_{k(i)}(i), t^{''} \in \tilde{t}_{k(n)}(n)$, s.t. $t^{'} + w_{ij} \leq t^{''}, \tilde{p}_i(t^{'}) > 0, \tilde{p}_n(t^{''}) > 0$, and $\sum_{t' \in \tilde{t}_{k(i)}(i)} \sum_{t'' \in \tilde{t}_{k(n)}(n)} \tilde{p}_i(t^{'}) \tilde{p}_n(t^{''}) = 1$.
        if there is no such n **then**
          n is the last node in *connecting_nodes*
          Append all the links from (i,m) to (i,n) to A_t
          Create a walking link from node i to all $d \in D$ if it does not exists.
        else
          Append all the links from (i,m) to (i,n) to A_t
      else if *assignment_type* == "capacitated" **then**
        Find first node n in *connecting_nodes* for which $\forall t^{'} \in \tilde{t}_{k(i)}(i), t^{''} \in \tilde{t}_{k(n)}(n)$, s.t. $t^{'} + w_{ij} \leq t^{''}, \tilde{p}_i(t^{'}) > 0, \tilde{p}_n(t^{''}) > 0$, and $t^{''} - t^{'} - w_{ij} \leq \delta_3$
        if there is no such n **then**
          n is the last node in *connecting_nodes*
          Append all the links from (i,m) to (i,n) to A_t
        else
          Append all the links from (i,m) to (i,n) to A_t
        Create a walking link from node i to all $d \in D$ if it does not exists.
      else
        Raise error

Assumptions

We make the following assumptions to define the mathematical model for the current problem.

1. The transit vehicle arrival and departure time at stops are assumed to be the same, i.e., no dwell time is assumed.
2. The walking time on access, egress, and walking transfer links is integer-valued and constant.
3. The travel time on auto and in-vehicle transit links are modeled as time-dependent discrete random variables with finite support.
4. The wait time of transfer links is modeled as a time-dependent discrete random variable with finite support.
5. There are no time-dependent correlations in travel time or wait time on links. Further, the travel time is assumed to be independent across transit trips and routes, i.e., bus bunching is ignored.
6. The online information about the cost of downstream links is available at each node of the network.
7. At a transfer node, passengers use online information about only those bus routes which are accessible from that node by an acceptable walking distance.
8. The online information about the bus arrivals provided to any passenger is one of the realizations obtained from the historical data. The proposed stochastic shortest path computes an optimal policy/strategy for every such realization.
9. Passengers decide which bus route to board as soon as they arrive at a particular node.

10. Passengers are assumed to be expected-cost minimizers. The word "optimal" policy or strategy used in this dissertation minimizes the expected cost of travel (comprising of walking, waiting, travel time, parking cost, and transit fare), and it does not consider other attributes affecting passengers' utility.

11. The walking, waiting, and in-vehicle travel times are equally onerous to passengers.

12. Passengers can coordinate the departure time from origin based on the online information about the bus arrival at various stops to avoid waiting time. Therefore, no latest departure time penalty is assumed.

13. The transit network is connected.

Some of the assumptions are non-restrictive and can be easily relaxed. Assumption 1 can be relaxed by including dwell time in the travel time of in-vehicle links. Assumption 6 is also not a strict requirement as nodes with no online information can just have deterministic costs for adjacent downstream links (with probability 1). Assumptions 5 and 7 are needed to avoid enormous state space in the proposed stochastic shortest path problem. The correlations in travel time can be considered by assuming realizations of the travel time of each link in the network. However, the corresponding stochastic shortest path problem is NP-hard (e.g., see [19, 23]). One can relax assumption 7 by including online information about all the bus routes in the state space and developing a solution algorithm that evaluates more intelligent routing policies in the network at the expense of computational time (e.g., see [32]). However, such algorithms are more suited for providing routing policies to passengers through cellphone applications. For assignment purposes, we believe this is a reasonable assumption and can aid in developing faster algorithms. The relaxation of assumption 9 would require formulating a dynamic path choice problem similar to [63], which we leave for us to explore in future

work. Assumptions 2-4, 8, and 10-13 are required for modeling purposes.

The random transit arrival time at various nodes induces a node-dependent stochasticity as when a passenger arrives at node $i \in N$ at time $t \in T$, an online information vector θ is revealed to them. This information vector consists of travel cost $\{c_{ij}^{\theta}\}_{j \in FS(i)}$ of outgoing links from node i [24, 28]. Let $\Theta_i(t)$ be the set of possible information sets at node i and time t. For an information vector $\theta \in \Theta_i(t)$ associated to the head node i of a transfer or access link, the travel cost of that link (i, j) for a possible arrival of bus at node $j \in FS(i)$ at $t_j \in \tilde{t}_{k(j)}(j)$ is given as:

$$c_{ij}^{\theta} = \begin{cases} t_j - t, & \text{if } t + w_{ij} \leq t_j \\ \infty, & \text{otherwise} \end{cases} \tag{3.1}$$

The probability of observing $\theta \in \Theta_i(t)$ is denoted by p^{θ}. For this probability distribution, we must have, $p^{\theta} \geq 0, \forall \theta \in \Theta_i(t), \forall t \in \tilde{t}_{k(i)}(i), \forall i \in N$ and $\sum_{\theta \in \Theta_i(t)} p^{\theta} = 1, \forall t \in \tilde{t}_{k(i)}(i), \forall i \in N$. Let us denote $\Theta = \cup_{i \in N} \cup_{t \in T} \Theta_i(t)$ as the union of all node-time information sets.

Example

To better understand the notations, we consider an illustrative example provided in Figure 3.1. It shows two trips $K = \{1, 2\}$ of different transit routes going from stop A to C and E to C respectively. There are three in-vehicle nodes $B = \{A_1, B_1, C_1, E_2, D_2, C_2\}$, one origin node $O = \{o\}$, and one destination node $D = \{d\}$. There are four in-vehicle links $A_v = \{(A_1, B_1), (B_1, C_1), (E_2, D_2), (D_2, C_2)\}$, four access/egress links $A_e = \{(o, A_1), (o, E_2), (C_1, s), (C_2, s)\}$, and one transfer link $A_t = \{(B_1, D_2)\}$. The random link travel time of in-vehicle links or walk time of access/egress/transfer links is shown by the links in the figure. Assume that buses of trips 1 and 2 begin their

journey at their commencing stop at time $t = 0$. Then, the possible arrival times of different trips at different nodes with their probabilities are given as: $\tilde{t}_1(A_1) = 0$ *w.p.* 1.0, $\tilde{t}_2(E_2) = 0$ *w.p.* 1.0, $\tilde{t}_1(B_1) = \begin{Bmatrix} 2, & w.p.\ 0.6 \\ 8, & w.p.\ 0.4 \end{Bmatrix}$, $\tilde{t}_2(D_2) = \begin{Bmatrix} 3, & w.p.\ 0.2 \\ 5, & w.p.\ 0.3 \\ 10, & w.p.\ 0.5 \end{Bmatrix}$, $\tilde{t}_1(C_1) = \begin{Bmatrix} 17, & w.p.\ 0.6 \\ 23, & w.p.\ 0.4 \end{Bmatrix}$, $\tilde{t}_2(C_2) = \begin{Bmatrix} 16, & w.p.\ 0.2 \\ 18, & w.p.\ 0.3 \\ 23, & w.p.\ 0.5 \end{Bmatrix}$.

Using the arrival time information, the set of online information sets at various nodes can be written as: $\Theta_o(0) = \Big[\{0,0\}, \quad w.p.\ 1.0\Big]$, $\Theta_{A_1}(0) = \begin{bmatrix} \{2\}, & w.p.\ 0.6 \\ \{8\}, & w.p.\ 0.4 \end{bmatrix}$, $\Theta_{E_2}(0) = \begin{bmatrix} \{3\}, & w.p.\ 0.2 \\ \{5\}, & w.p.\ 0.3 \\ \{10\}, & w.p.\ 0.5 \end{bmatrix}$, $\Theta_{B_1}(2) = \begin{bmatrix} \{15,1\}, & w.p.\ 0.2 \\ \{15,3\}, & w.p.\ 0.3 \\ \{15,8\}, & w.p.\ 0.5 \end{bmatrix}$, $\Theta_{B_1}(8) = \begin{bmatrix} \{15,\infty\}, & w.p.\ 0.5 \\ \{15,2\}, & w.p.\ 0.5 \end{bmatrix}$, $\Theta_{D_2}(3) = \Big[\{13\}, \quad w.p.\ 1.0\Big]$, $\Theta_{D_2}(5) = \Big[\{13\}, \quad w.p.\ 1.0\Big]$, and $\Theta_{D_2}(10) = \Big[\{13\}, \quad w.p.\ 1.0\Big]$.

At C_1 and C_2, for any possible arrival time, the information set will have a single link with cost of 1 *w.p.* 1.0.

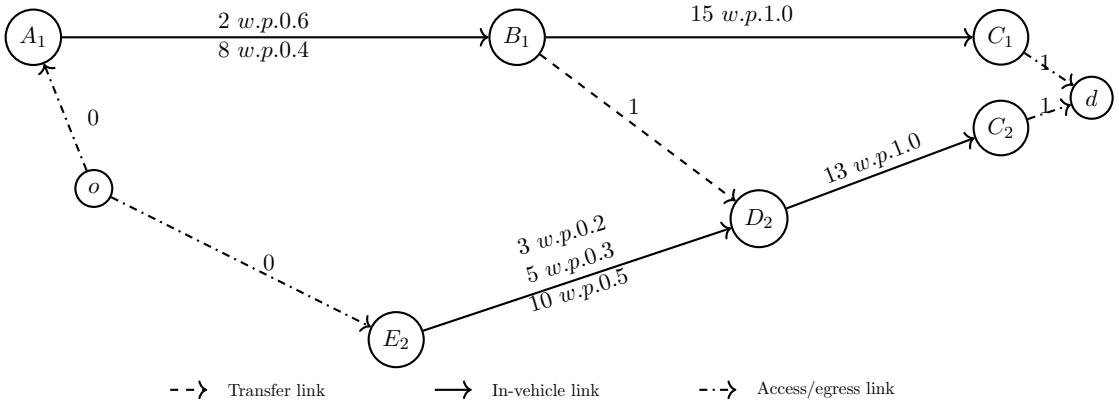


Figure 3.1: An illustrative example to show the stochastic transit network

In what follows, we present two different passenger routing algorithms, namely, for uncapacitated and capacitated networks.

3.1.2 Passenger routing in uncapacitated networks

In this case, we assume that transit vehicles have unlimited capacity. This routing algorithm is applicable for transit systems with low ridership but unreliable service, and for which denied boarding due to limited capacity is a rare phenomenon. The routing algorithm finds a strategy/policy in which arcs are chosen based on an adaptive rule. Such policy can be evaluated based on the online information about bus arrival time that helps passengers make a cleverer choice and improve their overall journey time. For example, in case of missed transfers, a passenger can consider an alternative route in their policy that provides the least expected cost to her destination. A policy induces a subgraph in the network known as *hyperpath*. A hyperpath, commonly used in FB models, is a collection of paths in the network that passenger travels on with positive probability. The current problem of finding a hyperpath exhibits sequential decision-making in a stochastic dynamic system, so it can be formulated as an infinite horizon Markov Decision Process (MDP)[6], which in this case, is also known as the stochastic shortest path (SSP) problem [135]. An SSP is characterized by a set of states, the set of actions available at each state, the probability of transitioning from one state to another by taking a particular action, a terminating state, and finally, the cost incurred by taking an action. These components of SSP for a specific destination $d \in D$ in case of the uncapacitated networks are described below:

1. *State space*: We use the state space description that has been used for static

[6]Note that the problem can be formulated as a finite horizon MDP where the time of arrival at various nodes is considered as different stages of the problem. However, we formulate it as an infinite horizon MDP by including the time of arrival at various nodes as part of the state of the problem. This is done to enjoy the nice properties and algorithms designed for infinite horizon problems.

and dynamic stochastic shortest path problems by various researchers in the past [17, 19, 24, 25]. The state space $S \subseteq N \times T \times \Theta$ defines the possible node states in which a traveler can be present. Each state $s \in S$ associated to transit node is characterized by a tuple $s = (i, t, \theta)$, where $i \in B$ represents the node in the network, $t \in \tilde{t}_{k(i)}(i)$ represents the possible arrival time at node i, and $\theta \in \Theta_i(t)$ represents the online information about the cost of links in $FS(i)$ obtained at node i and time t. A state s associated to origin node is characterized by $s = (o, t, \theta)$, where $o \in O$ is the origin node, $t \in T$ is the possible departure time from the origin, and $\theta \in \Theta_o(t)$ is the online information vector about the cost of outgoing links $FS(o)$ at time t. We also consider one special state known as the destination state d, which is an absorbing state.

2. *Action space*: Upon arrival at each node, the decision-maker considers the current travel cost and the availability of the information about the future travel time on downstream links to decide which arc to take next. For example, at every transfer node, a passenger receives information about the wait time of transferring nodes and whether a transfer is missed. Then, she has to decide which available action to take next. Therefore, the actions available at state (i, t, θ) are denoted by $u(i, t, \theta) = \{j \in FS(i) : c_{ij}^{\theta} \neq \infty\}$ i.e., the set of nodes in the forward star of node i that can be accessed from it. A stationary policy $\mu : S \mapsto N$ specifies the action to be taken at any state. Here, $\mu(i, t, \theta) \in u(i, t, \theta), \forall (i, t, \theta) \in S$.

3. *Transition Functions*: A traveler in state (i, t, θ), following policy μ, transitions to a new state $(\mu(i, t, \theta), t + c_{i\mu(i,t,\theta)}^{\theta}, \theta')$, by taking an action $\mu(i, t, \theta) \in u(i, t, \theta)$, the probability of which is denoted by $p^{\theta'}, \theta' \in \Theta_{\mu(i,t,\theta)}(t + c_{i\mu(i,t,\theta)})$. Note that by fixing the policy μ, one can construct a transition diagram with corresponding states and transition probabilities. The probability of transitioning from d to

itself, by taking any action $j \in u(d)$ is 1.

4. *One-step costs*: If the passenger chooses a link (action), a cost equal to the travel time, wait time, or walk time associated to the link is incurred. Let us denote the cost of choosing $\mu(i,t,\theta)$ at (i,t,θ) by $c^{\theta}_{i\mu(i,t,\theta)}$, where $\theta \in \Theta_i(t)$. The cost of transitioning from d to itself is zero.

5. *Value Function*: Let $J_\mu(i,t,\theta)$ at state (i,t,θ) denote the expected cost incurred by the passenger to reach her destination starting from node $i \in N$ at time t, receiving information $\theta \in \Theta_i(t)$ following policy μ. Mathematically, for a stationary policy μ, one can write

$$J_\mu(i,t,\theta) = \lim_{K\to\infty} \mathbb{E}\left\{ \sum_{k=0}^{K-1} c^{\theta^k}_{i^k,t^k,\mu(i^k,t^k,\theta^k)} | i^0 = i, \theta^0 = \theta, t^0 = t, \mu \right\} \tag{3.2}$$

where, (i^k, t^k, θ^k) represents the state at k^{th} stage, which represents the decision points of the passenger. The task is to determine the least expected cost $J^*(i,t,\theta)$ as well as an optimal routing policy $\mu^*(i,t,\theta)$.

Definition 1. *(Bellman Operator) The Bellman optimality operator $\mathcal{T} : S \cup \{d\} \to \mathbb{R}$ is defined as follows:*

$$(\mathcal{T}J)(i,t,\theta) = \min_{j\in u(i,t,\theta)} \{ c^{\theta}_{ij} + \sum_{\theta' \in \Theta_j(t+c^{\theta}_{ij})} p^{\theta'} J(j, t + c^{\theta}_{ij}, \theta') \} \tag{3.3}$$

In Definition 1, $\mathcal{T}J(i,t,\theta)$ is the optimal cost function for one-stage problem that has stage cost c^{θ}_{ij} and average terminal cost $\sum_{\theta' \in \Theta_j(t+c^{\theta}_{ij})} p^{\theta'} J(j, t + c^{\theta}_{ij}, \theta')$, where j achieves the minimum in (3.3). We next show the result that for any stationary policy μ, the resulting transition graph is acyclic.

Proposition 1. *For any stationary policy μ, the associated transition graph is acyclic, assuming that there do not exist self-transition probabilities associated with any state except the destination.*

Proof. Without the loss of generality, let us assume that there exists a cyclic path of minimum length 2. Let $\mathcal{P}\{[(i,t,\theta),(j,t^{'},\theta^{'})],[(j,t^{'},\theta^{'}),(i,t,\theta)]\}$ be that path. Since, $t^{'} = c_{ij}^{\theta} + t,\ \forall i,j \neq d$, implying that $t^{'} + c_{ji}^{\theta^{'}} = t$, which is a contradiction unless $c_{ij}^{\theta} = c_{ji}^{\theta^{'}} = 0$, in which case there does not exist such transition. Hence, there does not exist a cyclic path of length 2. One can extend this argument for a cyclic path of any length. □

The acyclicity property will be useful in developing efficient solution algorithms for finding the optimal policy. [135] showed that one can evaluate the optimal cost function vector $\mathbf{J}^*$ by solving the Bellman equation (3.4). Since the solution of this high-dimensional system of equations can be time-consuming, we can also repeatedly apply mapping $\mathcal{T}$ on some initial guess of vector $\mathbf{J}_0$ to evaluate $\mathbf{J}^*$ (see equation (3.5)). This is the basis of the value iteration method, which is a popular method to solve the Bellman equation (3.4).

$$\mathbf{J}^* = \mathcal{T}\mathbf{J}^* \tag{3.4}$$

$$\lim_{a\to\infty} \mathcal{T}^a J(i,t,\theta) = J^*(i,t,\theta), \forall (i,t,\theta) \in S \tag{3.5}$$

Finding the optimal policy

The previous section described that the optimal cost functions and a stationary optimal policy can be evaluated by solving the Bellman equation (3.4). However, it becomes difficult to solve it when the cardinality of the state space is large. This problem is commonly referred to as the "curse of dimensionality" in the MDP literature. It

turns out that the state space of the current problem can be reduced by averaging the information vector $\theta \in \Theta$. This is because θ is an uncontrollable component of the state. It depends on node $i \in N$ at which the traveler is currently located and the time of arrival $t \in T$ at that node but not on the control $u(i,t,\theta)$. In that case, we can formulate the problem only based on the controllable components (i.e., (i,t)) of the state space with the dependency on the uncontrollable component θ being "averaged out." To do that, let us consider another cost function $\hat{J}(i,t)$ defined on the reduced state space $\hat{S} = (N \times T) \cup \{d\}$. We have,

$$\hat{J}^*(i,t) = \sum_{\theta \in \Theta_i(t)} p^\theta J^*(i,t,\theta) \tag{3.6}$$

$$= \sum_{\theta \in \Theta_i(t)} p^\theta \min_{j \in u(i,t,\theta)} \{c_{ij}^\theta + \sum_{\theta' \in \Theta_j(t+c_{ij}^\theta)} p^{\theta'} J^*(j, t + c_{ij}^\theta, \theta')\} \tag{3.7}$$

$$\hat{J}^*(i,t) = \sum_{\theta \in \Theta_i(t)} p^\theta \min_{j \in u(i,t,\theta)} \{c_{ij}^\theta + \hat{J}^*(j, t + c_{ij}^\theta)\} \tag{3.8}$$

The equation (3.8) is the Bellman equation in the reduced state space $\hat{S}$. The cost functions $\hat{J}^*(i,t)$ can be viewed as the optimal expected cost-to-go from node i and time t before realizing the online information. The above stochastic shortest path problem can be solved using the methods used for solving the infinite-horizon MDP. It includes value iteration (VI), policy iteration, linear programming, etc.

The value iteration method works as follows. We start by intializing $\hat{J}_0(i,t) = 0, \forall (i,t) \in \hat{S}$. Then, for $a = 1, 2, ...$, we calcualte $\hat{J}_{a+1}(i,t) = \sum_{\theta \in \Theta_i(t)} p^\theta \min_{j \in u(i,t,\theta)} \{c_{ij}^\theta + \hat{J}_a^*(j, t + c_{ij}^\theta), \forall (i,t) \in \hat{S}$. Generally, VI requires an infinite number of iterations to converge to an optimal solution. However, for the current problem, it converges to an optimal solution in finite number of iterations (Proposition 2).

Proposition 2. *Given a directed acyclic transition graph corresponding to an optimal*

stationary policy μ^*, the value iteration will yield the optimal cost vector $\hat{J}^*$ in at most $|N|$ iterations when initialized as below:

$$\hat{J}^*(i,t) = \infty, \ \forall (i,t) \in \hat{S} \backslash \{d\} \tag{3.9}$$
$$\hat{J}^*(d) = 0$$

Proof. To show this, let us consider various subsets of the state space created as below:

$$\hat{S}_0 = \{d\} \tag{3.10}$$

$$\hat{S}_{a+1} = \left\{ (i,t) : \sum_{\substack{\theta:\mu^*(i,t,\theta)=j, \\ \theta' \in \Theta_j(t+c_{ij}^{\theta})}} \mathbb{P}_{\mu} \left[(i,t,\theta), (j, t + c_{ij}^{\theta}, \theta') \right] = 0, \forall j \notin \cup_{m=0}^{a} \hat{S}_m \right\}, a = 0, 1, ... \tag{3.11}$$

The above construction of sets adds various states in the backward direction of the destination node. For example, $\hat{S}_1$ will contain d and all the states associated to the nodes in $BS(d)$. Let $\hat{S}_{\bar{a}}$ be the last of these sets that is non-empty. In view of the acyclicity and proper stationary optimal policy assumptions, we have $\bar{a} \leq |N|$ and $\cup_{m=0}^{\bar{a}} \hat{S}_m = \hat{S}$. After this, one can show using induction that

$$(\hat{T}^a \hat{J})(i,t) = \hat{J}^*(i,t), \ \forall (i,t) \in \cup_{m=0}^{a} S_m, a = 1, ..., \bar{a} \tag{3.12}$$

The mathematical induction part is same as the proof given in [135]. □

Note that the result proved in Proposition 2 is a stronger result than the one proved in [135], which shows the convergence of VI in $|\hat{S}|$ iterations rather than $|N|$ iterations.

Proposition 3. *The worst case computational complexity of the value iteration algorithm is $\mathcal{O}(|\hat{S}||N||A|)$.*

Proof. Initialization of J can be done in $\mathcal{O}(1)$ time. The control computation can performed in $\mathcal{O}(|A|)$. The previous operation is performed for every $(i,t) \in \hat{S}\backslash\{d\}$, repeatedly $|N|$ number of times. Therefore, the overall complexity of the value iteration algorithm is equal to $\mathcal{O}(|\hat{S}||N||A|)$. □

Due to the structure of the current problem, we can also solve the Bellman equation using the label correcting algorithm proposed by [136]. The algorithm starts by initializing a scan eligible list SE containing the neighbors of the destination node. Then, it scans elements in the backward direction updating the label of every node for every time interval. Unlike the value iteration algorithm, it updates the labels of sets $\hat{S}_0$, $\hat{S}_1$, etc. described in the proof of the proposition 2 sequentially in various iterations rather than attempting to update labels of all the elements of set $\hat{S}$ in each iteration. The overall steps of finding the optimal policy using the label correcting algorithm are summarized in Algorithm 2.

Algorithm 2 Label correcting algorithm for finding optimal policy in uncapacitated networks

procedure ULC(d)
  (Initialize) $\hat{J}(i,t) \leftarrow \infty, \forall (i,t) \in \hat{S}\backslash\{d\}$ and $\hat{J}(d) \leftarrow 0$
  $SE \leftarrow BS(d)$
  while $SE \neq \phi$ **do**
    Remove an element i from SE
    for $t \in \tilde{t}_{k(i)}(i)$ **do**
      $tempJ \leftarrow 0$
      for $\theta \in \Theta_i(t)$ **do**
        $tempJ \mathrel{+}= p^{\theta} \min_{j \in u(i,t,\theta)} \{c_{ij}^{\theta} + \hat{J}(j, t + c_{ij}^{\theta})\}$
      if $tempJ < \hat{J}(i,t)$ **then**
        $\hat{J}(i,t) \leftarrow tempJ$; $SE \leftarrow SE \cup BS(i)$
  $\mu^*(i,t,\theta) \leftarrow \underset{j \in u(i,t,\theta)}{\operatorname{argmin}} \{c_{ij}^{\theta} + \hat{J}^*(j, t + c_{ij}^{\theta})\}, \forall (i,t,\theta) \in S\backslash\{d\}$ **return** $\hat{J}^*, \mu^*$

Proposition 4. *The worst case computational complexity of Algorithm 2 is* $\mathcal{O}(|\hat{S}||A|)$.

Proof. Assuming that the elements of SE are removed according to the FIFO rule, lines 2-11 are standard Bellman-Ford algorithm and can be performed in $\mathcal{O}(|\hat{S}||A|)$ time. The computational complexity of finding the optimal policy (line 12) can be performed in $\mathcal{O}(|\hat{S}||A|)$. Therefore, the overall complexity of the Algorithm 2 is equal to $\mathcal{O}(|\hat{S}||A|)$. □

3.1.3 Passenger routing in capacitated networks

In this case, the transit vehicles are assumed to have limited capacity. This can cause some arcs to become saturated and thus inaccessible depending on the route choice of other passengers. To model such behavior, previous studies have proposed to use *failure-to-board* probabilities or *access* probabilities. They evaluate the probability with which a passenger waiting at a bus stop can access an outgoing link. Such access probabilities result in multiple paths that a traveler can take with positive probability. The collection of such paths is known as "hyperpath." In the capacitated networks, the hyperpaths/strategies are induced by both risks of denied boarding due to limited capacity and missing transfers due to unreliable service. A strategy helps passengers minimize their expected costs under various types of uncertainties.

To incorporate the access/availability probabilities and find a strategy that minimizes the expected cost of travel in a capacitated network, we require augmenting the state space. For that purpose, let us define $X_i^\theta(t)$ as the random variable supported on $\{0,1\}^{|u(i,t,\theta)|}$ indicating the availability of arcs in $FS(i)$, when arriving at node $i \in N\backslash\{d\}$ at time $t \in T$, and receiving information θ. To be more precise, the component j of vector $x \in X_i^\theta(t)$ will indicate whether link $(i,j) \in A$ is available to board or not. Let π^x be the probability of observing the availability vector $x \in X_i^\theta(t)$ and $X = \cup_{(i,t,\theta)\in S} X_i^\theta(t)$ be the collection of such availability vectors. The use of π^x is akin to "access" probabilities

in the previous literature, as the former describes the node-based availability of outgoing links and later describes the availability of any link. It is assumed that passengers do not know about the availability vector x and information vector θ in advance and realize them when reaching a particular node at a particular time. To find an optimal strategy/policy in this case, we need to solve the corresponding stochastic shortest path problem, whose components are described below:

1. *State space*: The state space $S_C \subseteq N \times T \times \Theta \times X$ defines the possible node states in which a traveler can be present. Each state $s \in S_C$ is characterized by a tuple $s = (i, t, \theta, x)$, where $i \in N$ represents the node in the network, t represents the possible arrival time at node i, $\theta \in \Theta_i(t)$ represents the online information about the cost of links in $FS(i)$ obtained at node i and time t, and $x \in X_i^{\theta}(t)$ represents the availability of outgoing links. Similar to the uncapacitated case, destination $d \in D$ is considered as an absorbing state.

2. *Action space*: Upon arrival at each node at a particular time, the decision-maker considers the current travel cost, the future travel time information, and the availability of downstream links to decide which arc to take next. For example, at every transfer node, the passenger receives information about the wait time of transferring nodes and whether a link is available or not. A link may be unavailable due to missed transfer or the vehicle associated with it being full. Then, she has to decide which available action to take next. Therefore, the actions available at state (i, t, θ, x) are denoted by $u_C(i, t, \theta, x) = \{j \in u(i, t, \theta) : x[j] \neq 0\}$. Note that due to assumption 13 and the presence of walking links from transfer nodes, there is no state $s = (i, t, \theta, x)$ such that $u_C(i, t, \theta, x) = \phi$.

3. *Policy*: A policy/strategy specifies the subset of actions that can be taken at a state. To be precise, $\mu_C : S \mapsto 2^{|\sum_{s \in S} u_C(s)|}$ maps every state to a subset of

controls that provide equal expected cost to destination.

4. *Transition Functions*: A passenger in state $s = (i, t, \theta, x)$ following policy μ transitions to a new state $(j, t + c_{ij}^{\theta}, \theta^{'}, x^{'})$ by taking an action $j \in \mu_C(i, t, \theta, x)$, the probability of which is denoted by $p^{\theta^{'}} \pi^{x^{'}}$. The probability of transitioning from d to itself, by taking any action $j \in u(d)$ is 1. The value of π^x depends on the route choice of other passengers, and it is obtained from the network loading procedure described in Chapter 4.

Using the components defined above, we can formulate the Bellman equation for finding the optimal strategy as below:

$$J_C^*(i,t,\theta,x) = \min_{j \in u_C(i,t,\theta,x)} \left\{ c_{ij}^{\theta} + \sum_{\theta' \in \Theta_j(t+c_{ij}^{\theta})} \sum_{x' \in X_j^{\theta}(t+c_{ij}^{\theta})} p^{\theta^{'}} \pi^{x^{'}} J_C^*(j, t + c_{ij}^{\theta}, \theta^{'}, x^{'}) \right\}, \forall (i,t,\theta,x) \in S_C \tag{3.13}$$

where, $J_C^*(i, t, \theta, x)$ denotes the optimal cost-to-go from state (i, t, θ, x) to the destination in case of capacitated networks. Similar to uncapacitated networks, one can reduce the state space and define the Bellman equation only on controllable components by averaging the uncontrollable components.

$$\hat{J}_C^*(i,t) = \sum_{\theta' \in \Theta_j(t+c_{ij}^{\theta})} \sum_{x' \in X_j^{\theta}(t+c_{ij}^{\theta})} p^{\theta} \pi^{x} \min_{j \in u_C(i,t,\theta,x)} \left\{ c_{ij}^{\theta} + \hat{J}_C^*(j, t + c_{ij}^{\theta}) \right\}, \forall (i,t) \in \hat{S} \tag{3.14}$$

The above Bellman equation can be solved using a label correcting algorithm similar to Algorithm 2.

3.1.4 Numerical experiments

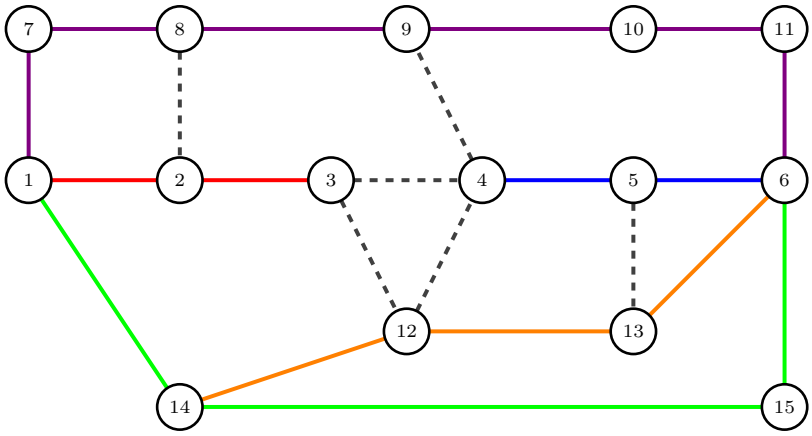
In this section, we show the application of the proposed routing algorithms based on the transit network given in [2]. The network and schedule are provided in Figure 3.2. There are 15 stops and 5 color-coded transit routes in the network. The original network

has only three walking transfer links, namely, 3-4, 3-12, and 12-4. To better understand transfer behavior in the presence of online information, we created more walking transfer links in the network. They are given as 2-8, 9-4, and 13-5. Stop 14 is the only stop that provides waiting transfer from one route to another in the network. There are 4 trips of each route whose complete schedule is shown in Figure 3.2(b). There are two destinations, namely 5 and 6, and three origins, namely, 1, 7, and 14 in the network.

The support of random travel times of in-vehicle links is given as $\{0.9\bar{c}_{ij}, \bar{c}_{ij}, 1.2\bar{c}_{ij}, 1.5\bar{c}_{ij}\}$, where $\bar{c}_{ij}$ is the scheduled travel time of link $(i,j) \in A_v$. All trips are assumed to have a capacity of 20 passengers. The early and late arrival penalties are assumed to be $\eta_1 = \eta_2 = 0.5$. The acceptable waiting and walking times are assumed as $\delta_0 = \delta_1 = \delta_2 = \delta_3 = 15$ minutes. Overall, there are 89 nodes and 173 links in the schedule-based transit network, including 24 access, 20 egress, and 64 in-vehicle links. In what follows, we present the results of expected cost of travel between every origin-destination pair for the uncapacitated and capciatated transit networks in Section 3.1.4 and 3.1.4 respectively. All implementations were coded in Pyhon 3.8 and tests were executed on Intel(R) CPU running at 2.2 GHz with 16 GB RAM under a Windows operating system.

Uncapacitated network

We start by creating the transfer links using Algorithm 1. For the uncapacitated network, it creates only 4 waiting transfer links and 25 walking transfer links in the current schedule-based network. The number of generated states is 4,720. After this, we solve the Bellman equation (3.4) for individual destinations. It took a fraction of a second

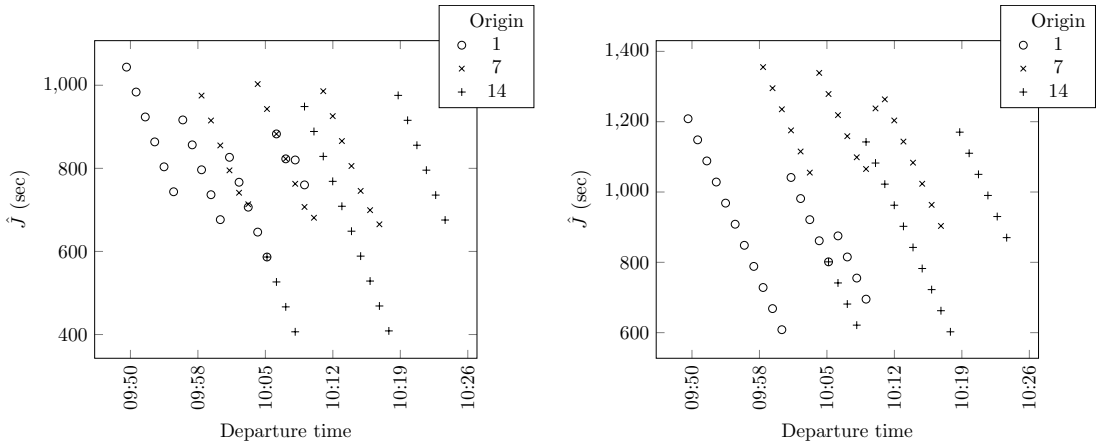


(a) Network

RouteID	TripID	Schedule
Red	1001	10:00, 10:02, 10:04
	1002	10:10, 10:12, 10:14
	1003	10:20, 10:22, 10:24
	1004	10:30, 10:32, 10:34
Blue	2001	10:06, 10:08, 10:10
	2002	10:16, 10:18, 10:20
	2003	10:26, 10:28, 10:30
	2004	10:36, 10:38, 10:40
Violet	3001	10:00, 10:02, 10:04, 10:06, 10:08, 10:10, 10:12
	3002	10:07, 10:09, 10:11, 10:13, 10:15, 10:17, 10:19
	3003	10:14, 10:16, 10:18, 10:20, 10:22, 10:24, 10:26
	3004	10:21, 10:23, 10:25, 10:27, 10:29, 10:31, 10:33
Orange	4001	10:08, 10:10, 10:12, 10:14
	4002	10:18, 10:20, 10:22, 10:24
	4003	10:28, 10:30, 10:32, 10:34
	4004	10:38, 10:40, 10:42, 10:44
Green	5001	9:55, 9:57, 10:04, 10:06
	5002	10:05, 10:07, 10:14, 10:16
	5003	10:15, 10:17, 10:24, 10:26
	5004	10:25, 10:27, 10:34, 10:36

(b) Schedule

Figure 3.2: Network and schedule [2]



(a) Destination 6 (b) Destination 5

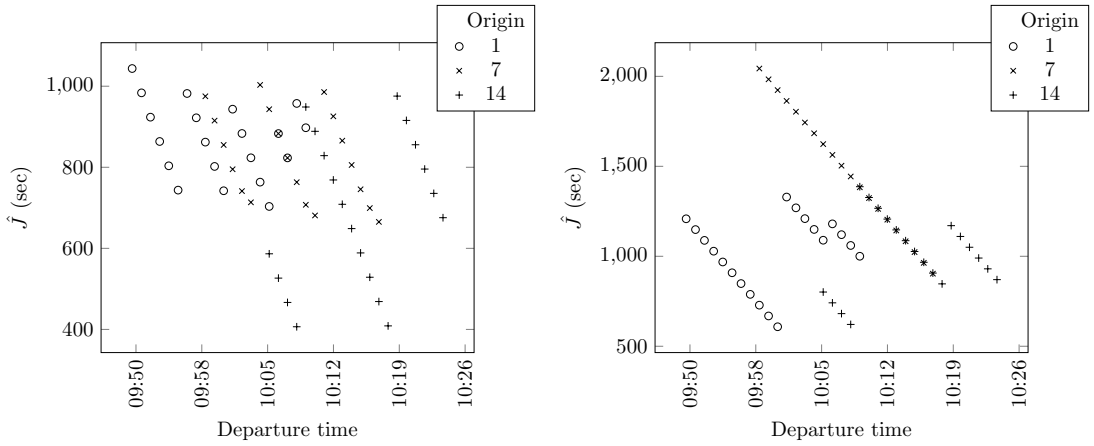
Figure 3.3: Expected cost between various origin-desination pairs for varying departure times in case of uncapacitated networks

to solve the current problem using both value iteration as well as label correcting algorithm. Figure 3.3 shows the expected cost of travel between various origin-destination pairs for varying departure times. We can observe that for various origins, the expected cost to the respective destination decreases with time until we reach the time when a bus trip departs from that origin. Then, similar behavior is observed for the passengers waiting for the next trip to arrive. Further, we see that the average cost to destination 6 is lower than the average cost to destination 5. This is because destination 6 could be reached from various origins without transferring to a different route. On the other hand, to reach destination 5, one must take a transfer to a different route, which sometimes causes a longer expected cost.

Capacitated network

In this case, Algorithm 1 creates 7 waiting transfer links, 24 walking transfer links, and 34 walking links for failed transfers. The access probabilities are calculated using the

assignment procedure described in Chapter 4. The number of generated states is 29,832, which is six times higher as compared to the uncapacitated case. The final values of the expected cost of traveling between various origin-destination pairs are plotted in Figure 3.4. The average cost of travel to destination 5 is more as compared to destination 6. This is because of the presence of paths without transfer between various origins and destination 6. On the other hand, one has to take at least one transfer to get to destination 5. Due to limited capacity, passengers miss transfers which leads to higher expected travel times. If we compare the expected cost from various origins to destination 6 in both uncapacitated and capacitated cases, we find that the expected cost of travel between 1-6 is higher in the case of the capacitated case as compared to the uncapacitated case. This is because passengers who do not get the preferred option of the red route due to limited capacity would either have to take the blue route or the green route resulting in higher expected cost. For destination 5, the expected cost of travel between 7-5 in the case of the capacitated case has risen considerably as compared to the uncapacitated case. This is because passengers who want to take transfer 8-2 coming from 7 on the violet route do not get priority over passengers who are continuing their journey on link 2-3 of the red route.



(a) Destination 6 (b) Destination 5

Figure 3.4: Expected cost between various origin-desination pairs for varying departure times in case of capacitated networks

3.2 Park-and-ride passenger routing with online information

In this section, we consider the park-and-ride mode in which travelers drive up to a certain stretch of the highway, and then take transit to reach their destination. The routing of park-and-ride users is different from transit and auto users as travelers need to decide where to park and take transit. These decisions are often made adaptively considering the realized state of traffic and transit availability. We model the adaptive routing problem for park-and-ride users, which is defined as follows: Given the uncertainty in travel time on a freeway and bus arrival time, the problem is to navigate a traveler commuting from a suburban region to a downtown location in minimum expected travel cost by deciding:

1. Whether to continue driving towards the destination or use a park-and-ride location to take transit?

2. At which park-and-ride location should the traveler park their car?

3. At what times of the day, taking transit from a park-and-ride location minimizes the overall expected travel cost?

To better understand the problem, let us consider an illustrative example provided in Figure 3.5. The network given in this example consists of 4 nodes and 4 links, with travel time shown on the links. The path $(r^{'}, p^{'}, d^{'})$ represents a freeway with $p^{'}$ as the park-and-ride node from where it is possible to take transit. The travel time and wait time on links $(p^{'}, d^{'})$ and $(p^{'}, t^{'})$ respectively are random variables (but not time-varying for simplicity) whose distribution is shown on the links. When a traveler arrives at node $p^{'}$, one of the possible realizations of cost on the outgoing links is revealed to them. Depending on the information and time of arrival, we need to compute an optimal policy that decides whether to take transit or continue driving on the freeway to minimize the expected cost to reach the destination. In this case, it is optimal to take freeway from $p^{'}$ if the travel time of 10 units is revealed to the traveler on it, otherwise, it is optimal to take transit from there. The expected cost of using the adaptive routing policy is 42.5 units in comparison to an expected cost of 56 units by always taking the freeway and 45 units by always taking the park-and-ride option.

3.2.1 Problem formulation

In this section, we formulate the adaptive park-and-ride location problem. For this purpose, we extend the ideas from the previous section on transit to the park-and-ride case. Along with assumptions 1-13, we also assume an uncapacitated network, where there is sufficient capacity at park-and-ride facilities to park and sufficient capacity to accommodate passengers in transit vehicles. Let us consider the multimodal network with auto as well as transit networks connected by park-and-ride locations. The auto

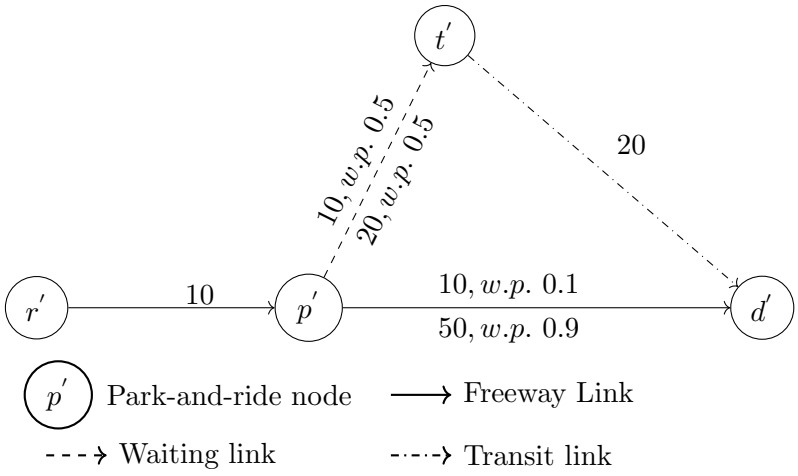


Figure 3.5: An illustrative example of a park-and-ride trip in a network with random travel time

network is represented by a directed graph $G_R(N_R, P_R, A_R)$, where N_R represents the set of nodes, A_R represents the set of links, and $P_R \subset N_R$ represents the set of nodes where park-and-ride facilities are located in the road network. Let $\mathcal{Z} : P_R \mapsto 2^N$ be the set-valued map that assigns a park-and-ride node to a set of transit nodes. Note that there can be no park-and-ride node connected to a transit node if the walking distance to the nearest park-and-ride node is more than a certain threshold (say 0.75mi). Let $M = \{(n_R, n_t) \in P_R \times N : n_t \in \mathcal{Z}(n_R)\}$ be the set of links created between park-and-ride nodes and transit nodes and vice-versa known as *Mode transfer links*. These links involve walking time to access the transit stop and waiting time before the arrival of the bus. Let $Z = \cup_{n \in P_R} \mathcal{Z}(n)$ be the collection of all transit network nodes which are connected to various park-and-ride facilities. As the decision of taking transit at a park-and-ride node depends on the time the traveler arrives at that node, we need to expand the static road network into a time-dependent network. This is done by replicating the nodes $i \in N_R$ at different time intervals $t \in T$ and connecting them with the corresponding cost of travel.

We assume that the travel time and wait time on the auto network links A_R and mode transfer links M, respectively, are time-dependent discrete random variables with finite support. The transit network is assumed to have properties described in Section 3.1. Similar to the transit case, the random travel time on links $A_R \cup M$ results in a node-dependent stochasticity as when the traveler arrives at node $i \in N_R$, the information about the travel time on all the downstream links attached to it is revealed to them. Let $\Theta_i(t)$ denote the set of possible states at node $i \in N_R$, where probability of observing a particular state $\theta \in \Theta_i(t)$ at time t is p^θ. Each state $\theta \in \Theta_i(t)$ is a realization of travel time or wait time (if $i \in P_a$) on downstream links attached to node i. Let $\mathcal{S}_{ij}(t)$ be the set of possible realizations of travel time (or wait time) on link $(i,j) \in A_R \cup M$ at time t. Clearly, $\Theta_i(t) = \times_{(i,j)\in A_R \cup M} \mathcal{S}_{ij}(t)$ are the possible states at node i and time t, where $\times$ represents the cross product of the sets. A value of time parameter α is used to convert monetary costs into time units. The monetary costs may include parking cost τ_p at a particular node or transit fair τ_f to board a transit route.

The current problem exhibits sequential decision making in a stochastic dynamic system, so it can be formulated as a Markov Decision Process (MDP). The components of MDP are described below:

1. *State Space*: The state space $S_{PNR} \mapsto (N_R \cup M \cup N) \times T \times \Theta$ defines the possible states in which a traveler can be present. Each state $s \in S$ is defined by tuple $s = (i, t, \theta)$, where, i represent the auto, park-and-ride, or the transit node. Here, $t \in T$ is the time of arrival at node i, and $\theta \in \Theta_i(t)$ represents the information obtained at node i at time $t \in T$. Any state corresponding to the destination node is considered as an absorbing state (once a traveler reaches there will remain there forever).

2. *Action Space*: Upon arrival at each node with no park-and-ride facility, the decision-maker considers the current travel cost, and the availability of the information about the future travel time on downstream arcs and then decide which arc to take next. On the other hand, when the traveler arrives at a node with a park-and-ride facility, she has two options available: whether to park and wait for transit or take one of the downstream auto links. Therefore, the actions available at state (i, t, θ) are denoted by $u(i, t, \theta) = \{j \in (N_R \cup M \cup N) : j \in FS(i)\}$ i.e., the set of nodes in the forward star of node i. A policy μ defines a stationary policy that specifies the action to be taken at any state.

3. *One-step costs*: If the decision maker chose to take an auto link, a cost equal to the travel time on the forward link is incurred. Similarly, if she decides to take a waiting link to board a transit route, a cost equal to the wait time is incurred. Let us denote the cost of choosing $\mu(i, t, \theta)$ at (i, t, θ) by $c^{\theta}_{i\mu(i,t,\theta)}$, where $\theta \in \Theta_i(t)$. The cost of transitioning from (d, t, θ) to itself is zero, $\forall \theta \in \Theta_d(t), \forall t \in T$.

4. *Transition Functions*: A traveler at state (i, t, θ), following policy μ, transitions to a new state $(\mu(i, t, \theta), t + c^{\theta}_{i\mu(i,t,\theta)}, \theta')$, by taking an action $\mu(i, t, \theta) \in u(i, t, \theta)$, the probability of which is denoted as $p^{\theta'}, \theta' \in \Theta_{\mu(i,t,\theta)}(t + c^{\theta}_{i\mu(i,t,\theta)})$. The probability of transitioning from $(d, t, \theta), \theta \in \Theta_d(t)$ to itself, by taking any action $j \in u(d, t, \theta)$ is 1.0.

5. *Value Function*: Let $J_\mu(i, t)$ denote the cost incurred by a traveler to reach the destination starting from node $i \in (N_R \cup M \cup N)$ at time $t \in T$ following policy μ.

Similar to the transit case, the adaptive park-and-ride routing problem can be formulated using the following Bellman equation

$$\hat{J}^*(i,t) = \sum_{\theta \in \Theta_i(t)} p^\theta \min_{j \in u(i,t,\theta)} \{c_{ij}^\theta + \hat{J}^*(j, t + c_{ij}^\theta)\}, \forall (i,t) \in ((N_R \cup M \cup N) \times T) \quad (3.15)$$

The above equation can be solved using Algorithm 2.

3.2.2 Numerical experiment (Case study of I-394)

In this section, we present a case study of the freeway I-394 in Twin Cities, MN to show the application of the proposed model. I-394 is 9.8 miles long freeway (E-W) serving the Hennepin County of Minnesota. The major junctions along this freeway include I-494 and US-169 in Minnetonka, MN-100 in Golden Valley, and I-94 in Minneapolis. A High Occupancy Vehicle (HOV) lane was built on this freeway in May 2005 to maximize its capacity. The HOV lane can be used by buses and generally provide reliable service when general-purpose lanes are congested.

I-394 connects various sub-urban areas to Downtown Minneapolis. Due to the congestion during peak periods, five park-and-ride facilities are provided to the travelers (Table 3.1) along this freeway. These park-and-ride facilities are served by several bus routes, connecting them to various locations in Twin Cities such as Downtown Minneapolis, University of Minnesota campus, Downtown St. Paul, and so on. The park-and-rides facility locations are shown geographically in Figure 3.6, and a list of bus routes serving these park-and-ride facilities are provided in Table 3.1.

Table 3.1: Park-and-ride locations along I-394

Order	Name	Bus routes	Address
1	Plymouth Road Transit Center & Park & Ride	652, 672, 645, 677	13126 Wayzata Blvd, Minnetonka, MN 55305
2	I-394 & Co. Rd. 73 Park & Ride	615, 652, 673, 645, 679	1100 Hopkins Crossroads, Minnetonka, MN 55305
3	General Mills Blvd. & I-394	645, 652, 672, 756	8675 Wayzata Blvd.
4	Louisiana Avenue Transit Center Park & Ride	9, 604, 643, 645, 652, 663, 672, 705, 756	1300 Louisiana Avenue, St, Louis Park, MN 55426
5*	Park & Ride	645, 9	Wayzata Blvd, Minneapolis, MN 55416

* 5 being closest to Downtown Minneapolis

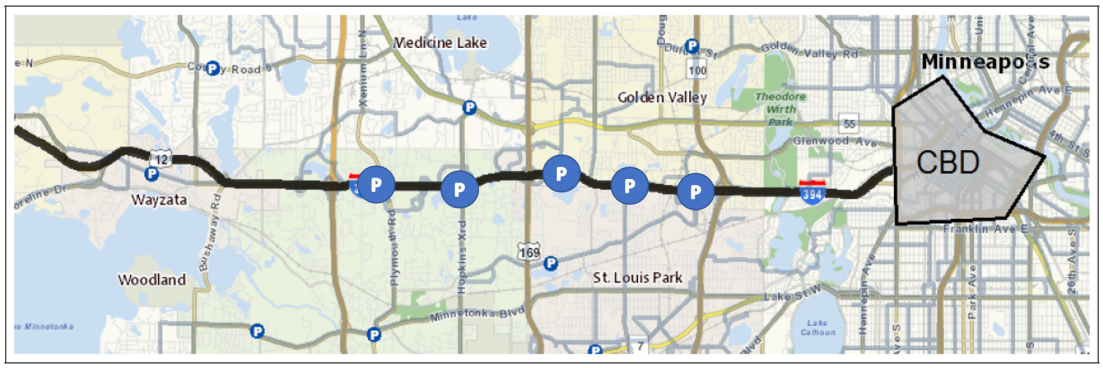


Figure 3.6: Minneapolis CBD and park-and-ride facilities along I-394 corridor

Network and calibration of distribution of travel time and wait time

The Minnesota Department of Transportation (MnDOT) has located loop detectors every 0.5mi along I-394. They collect data about the travel speed of cars, which in turn gives us the travel time on different sections of the freeway. We used Google My Maps to create the network. More details about the network topology are given in Table 3.2.

Table 3.2: I-394 network topology

# of nodes	32
# of auto links	23
# of access or egress links	16
Time horizon	6:00 A.M. - 10:00 A.M.
# of time steps	480
# of transit trips	68

It is assumed that the travel time recorded using a particular detector applies to half of the distance between the upstream detector and current location and similarly

half of the distance between the downstream detector and the current detector location. We consider only two possible states of the freeway links, namely, "congested" and "uncongested". Highway Capacity Manual defines the quality of any freeway on a scale of A-F known as the level of service (LOS). Any level of service below C is considered as a "congested" state of the freeway, and "uncongested", otherwise. For a typical highway, the travel speed below 60mph is considered as the level of service C. This value of speed is used to determine the probability of a freeway link being congested or uncongested using the loop detector data collected in April 2017 [137]. The probability distribution of the travel time is calibrated for every 30 seconds of the time horizon and the mean value of travel time in each time interval is used for a particular state of the link. To avoid inconsistency between the transition of states, each value of travel of time on every link was rounded to the nearest multiple of 30 seconds.

For this case study, the bus routes which serve the Downtown Minneapolis area are only considered. The destination node in our network is assumed to be the intersection of Hennepin Ave and 12th St, which is located in Downtown Minneapolis. The actual bus arrival time at various park-and-ride bus stops is obtained from the historical Automatic Vehicle Location (AVL) data collected by Metro Transit (transit agency in the Twin Cities region) over one year. Then, the difference between the actual and the scheduled arrival time at these bus stops is used for the calibration of the probability distribution of the random wait time of the mode transfer links. Note that we also considered the cases when the bus arrived early at the park-and-ride bus stops. The value of time (α) is assumed to be \$23/h as recommended by [138].

After creating the network, calibrating the required probability distributions and reducing the state space, we use both value iteration and label correcting methods to

solve the stochastic shortest path problem presented in previous section. We calculated the online shortest path during morning peak hours from 6:00 A.M. to 10:00 A.M., from different nodes of the freeway to the specified destination in Downtown Minneapolis. The number of states generated for this experiment after reducing the state space was 11,562. A typical value of transit fare $\tau_f = \$3$ and parking cost $\tau_p = \$12$ for one day is assumed for this experiment. Usually, the parking is free at park-and-ride locations, so we did not consider any cost associated with it. The value iteration method for this small network took 24 iterations and 18 minutes to converge to the optimal solution. On the other hand, the label correcting method outperformed the value iteration method and took only 1.42 minutes to converge.

Before delving into the results produced by the algorithm, let us first explore the congestion conditions on I-394. Figure 3.7 and 3.8 show heatmaps of the travel time during both congested and uncongested conditions. The links are shown on the horizontal axis in the E-W direction whereas the vertical axis shows the time of the day. The node numbers increase in the direction of travel. We can see that during congested conditions (Figure 3.8), the overcrowding happens after 7:00 A.M. along three different stretches of the freeway. The first stretch is upstream before the first park-and-ride location. The second and third stretch of overcrowding is between Node 273 and 277 and after 279 up to central downtown respectively. Intuitively, taking park-and-ride between node 273 and 278 seems to be a reasonable choice.

Expected cost

Figure 3.9 shows the results of the expected cost of travel (in seconds) from various nodes to node 291 (destination). The four-digit nodes represent park-and-ride nodes

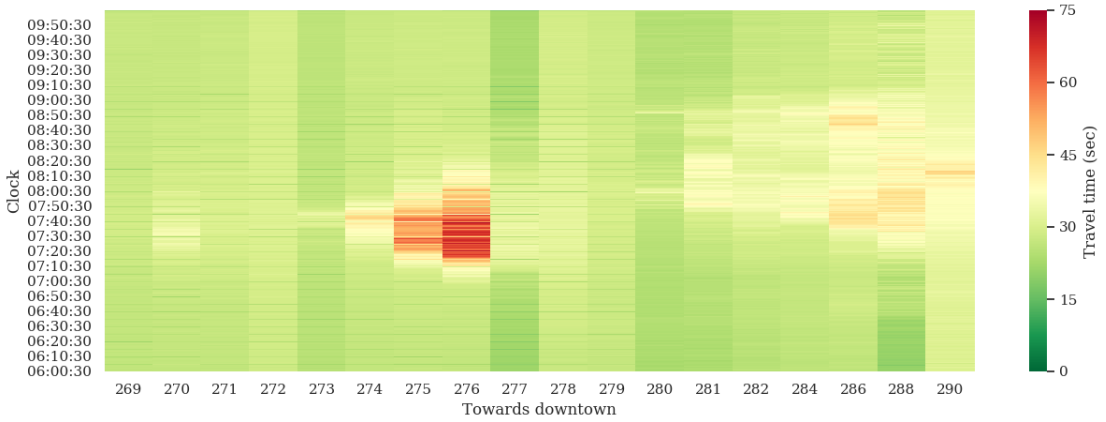


Figure 3.7: Travel time (sec) on links during uncongested conditions (For interpretation of colors in this figure, the reader is referred to the web version of this dissertation)

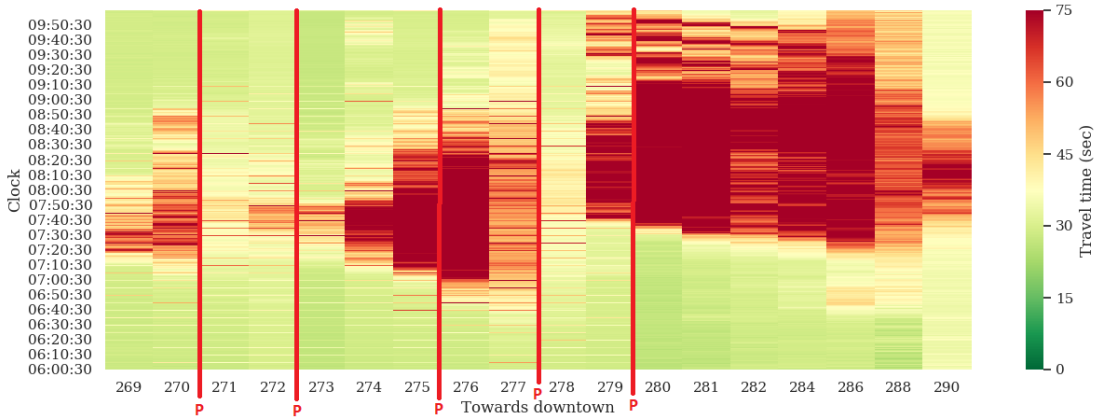


Figure 3.8: Travel time (sec) on links during congested conditions (For interpretation of colors in this figure, the reader is referred to the web version of this dissertation)

along the freeway, which are also marked with the letter P in the figure. The figure shows an increase in the congestion as the time increases on the time scale with severe congestion between 7:00 A.M.-9:45 A.M. However, the availability of the bus at the park-and-ride nodes provides a reduction in the expected travel time during several periods. This is evident from the light yellow-colored stripes appearing within the red-colored region. This reduction in travel time is due to the policy of taking transit at a park-and-ride location. The figure clearly shows the time of the day when park-and-ride

mode becomes more attractive in comparison to auto as this will provide faster access to the destination. We observe this significant reduction in the travel time when the buses are not frequent. In case if Metro Transit provides frequent service, we expect to see further improvement in the travel time of the commuters.

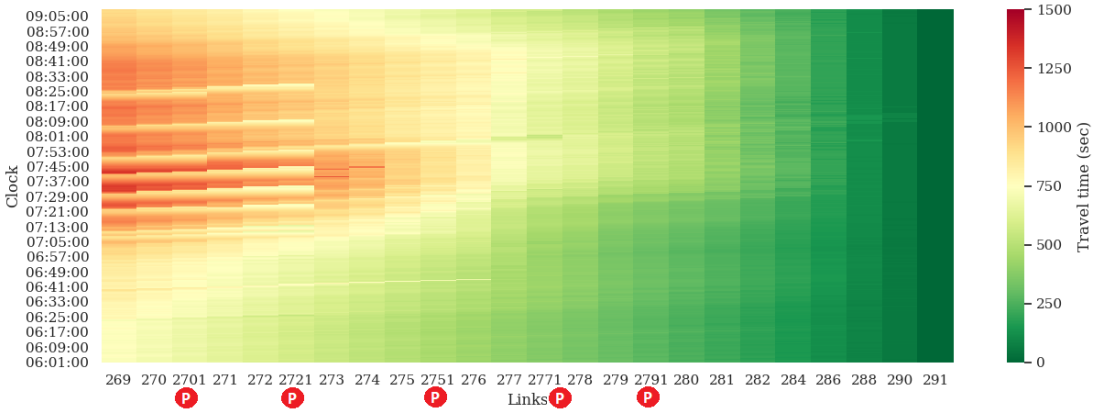
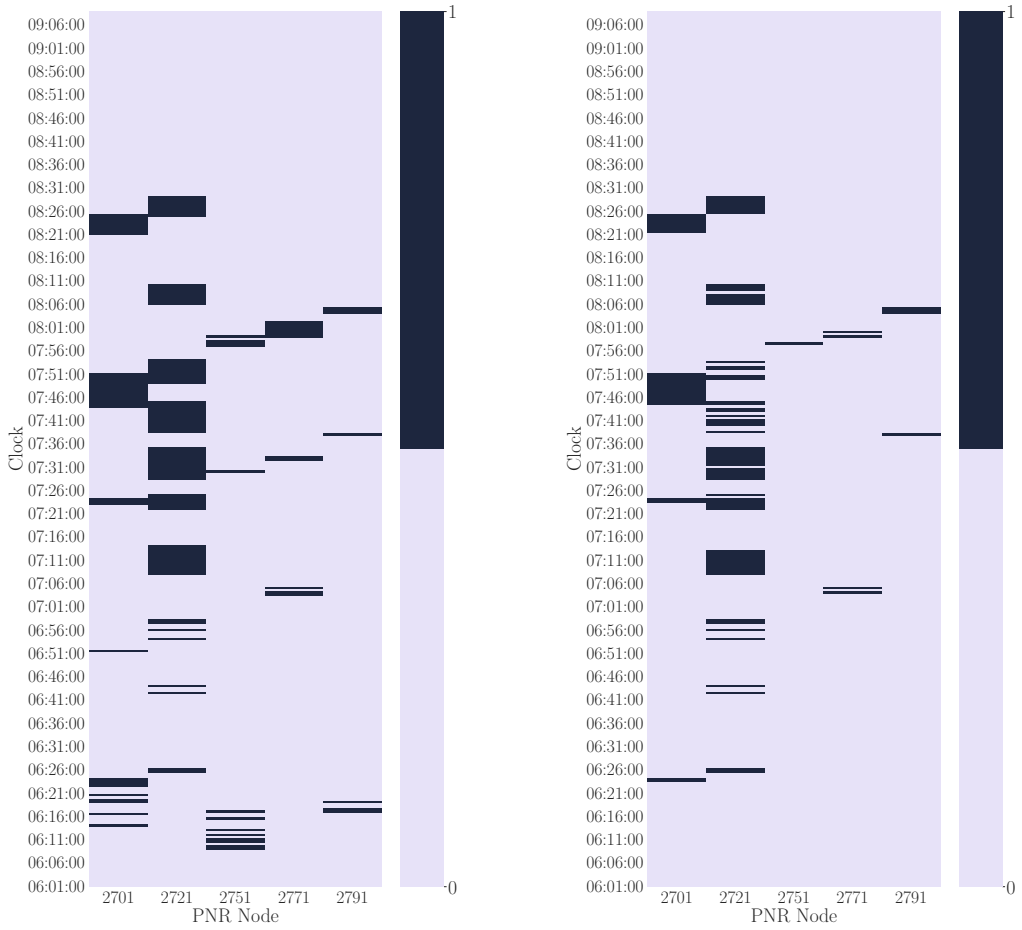


Figure 3.9: Travel cost (sec) of from different nodes to downtown (For interpretation of colors in this figure, the reader is referred to the web version of this dissertation)

Optimal Policy

The output of the Algorithm 2 also yields the optimal policy for different node states. The policy will help a traveler to decide which downstream arc to take when arriving at any node. The point of interest in this research is when does taking the transit from any park-and-ride location becomes optimal. To observe this behavior, we plotted the optimal policy for a traveler at park-and-ride exit nodes, which are plotted on the vertical axis in Figure 3.10, with the time of arrival at different nodes on the horizontal axis. Figure 3.10(a) and 3.10(b) depict the optimal policy when the road network is in the congested and the uncongested state respectively. The policy shows that the park-and-ride option becomes more attractive during the congested conditions than uncongested

conditions. Moreover, it is interesting to note that taking transit from I-394 & Co. Rd. 73 Park & Ride and Plymouth Road Transit Center & Park & Ride (which are located upstream) is more frequent than the other park-and-ride facilities.



(a) Congested conditions

(b) Uncongested conditions

Figure 3.10: Optimal policy for park-and-ride nodes

We use Monte-Carlo simulation to evaluate the variability of the optimal policy μ^*. For a given time interval, we generate 1000 random trajectories following policy μ^* starting from the farthest end (node '269' in Figure 3.9) and ending at the specified destination. In particular, for every sample and node-time (i, t) pair, we draw a random state $\theta \in \Theta_i(t)$ based on the discrete probability distribution $\{p^\theta\}_{\theta \in \Theta_i(t)}$. Figure 3.11 shows the results of the simulation that includes the distribution with mean and standard deviation of travel time at different times of arrival at node '269'. We can observe that the standard deviation ranges from 20 to 80 seconds for most time intervals except 7:30 A.M. at which it rises to 222 seconds. This is because of one particular trajectory with a travel time equal to 300 seconds.

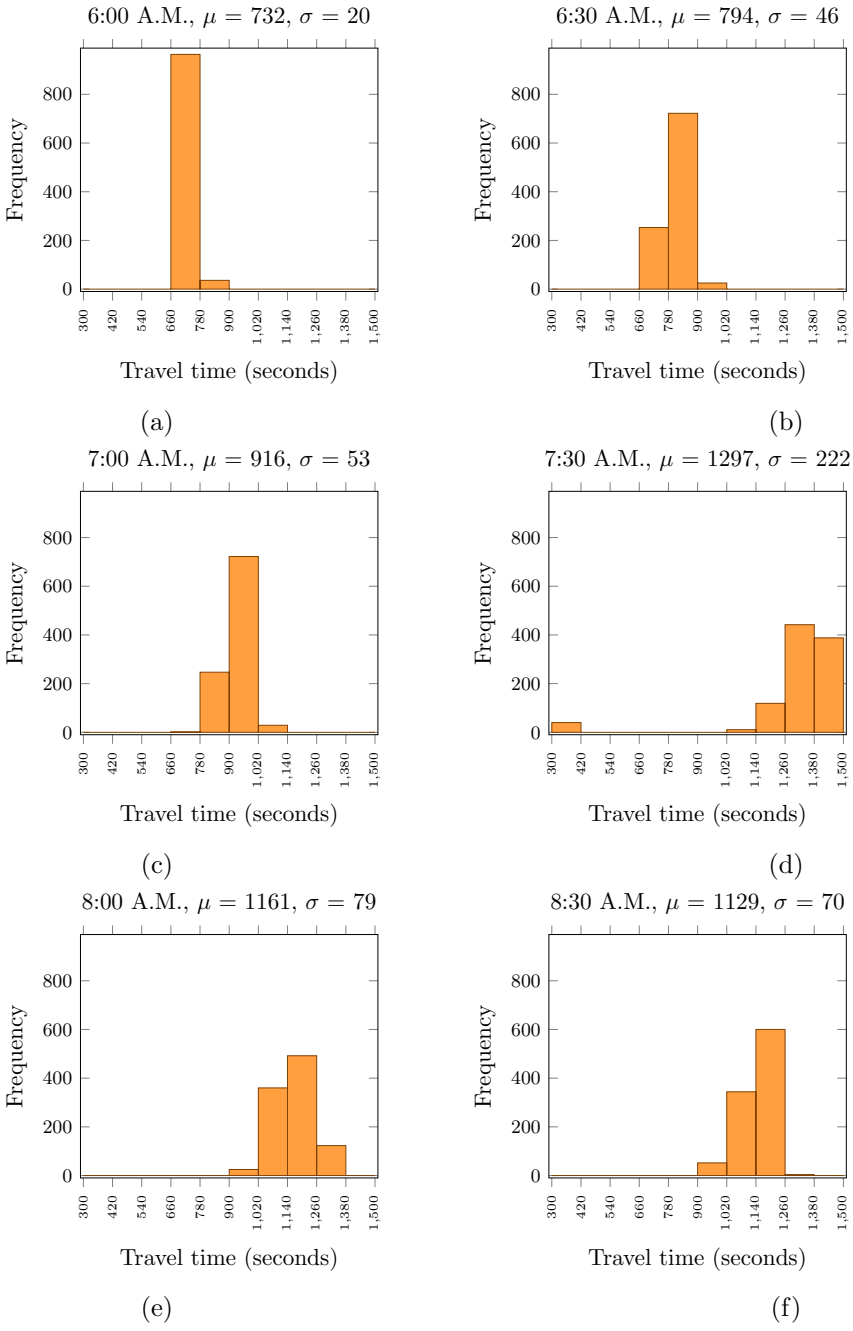


Figure 3.11: Mean and standard deviation of travel time (seconds) to the destination for different time of arrival at node '269' computed using 1000 sample trajectories

We further use same Monte-Carlo simulation to evaluate the attractiveness of various park-and-ride facilities. Figure 3.12 shows the frequency of use of different park-and-ride facilities at various times of arrival out of 10,000 sample trajectories. Through this analysis, we can make similar observations as made in Figure 3.10. The park-and-rides I-394 & Co. Rd. 73 Park & Ride and Plymouth Road Transit Center & Park & Ride (which are located upstream) are more attractive than the other park-and-ride facilities.

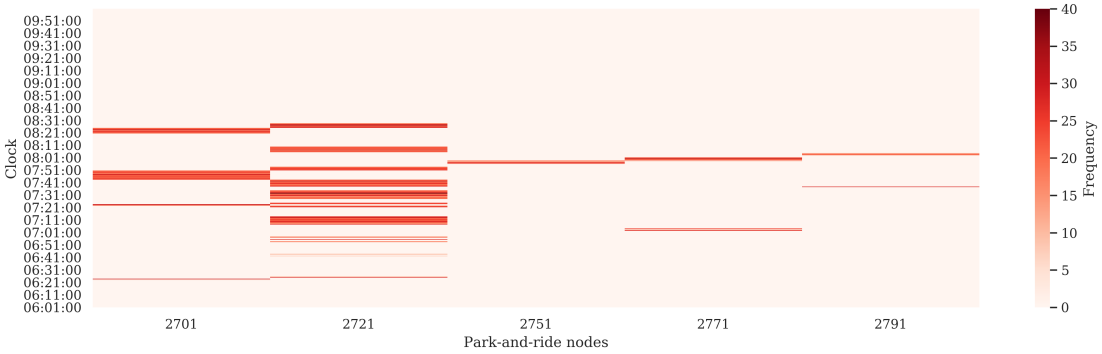
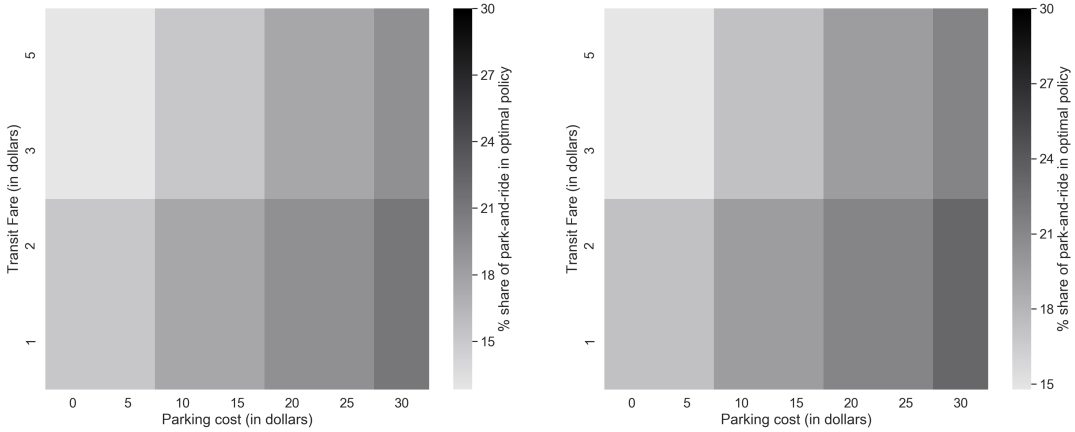


Figure 3.12: Frequency of use of different park-and-ride facilities out of 10,000 sample trajectories

Sensitivity analysis

To see how parking cost τ_p at the destination location and transit fare τ_f affects the optimal policy, we performed a sensitivity analysis on these two parameters. The value of τ_f and τ_p was varied from $\$1 - \$5$ and $\$0 - \$30$ respectively. The transit fare cost is added twice as we also consider the cost of taking transit for the reverse commute. We calculated the percentage of times when park-and-option opted as a mode of travel in the optimal policy. The figures 3.13(a) and (b) depicts that the park-and-ride mode is attractive when the transit fare is low and parking cost is high. With no parking fee and \$3-5 transit fare, we observe the lowest share of park-and-ride option in the optimal policy, i.e., 12.82% and 14.78% during uncongested and congested conditions

respectively. On the other hand, with $30 parking fee and $1-2 transit fare, we observe the highest share of park-and-ride option in the optimal policy, i.e., 21.10% and 23.11% during uncongested and congested conditions respectively. We also performed a sensitivity analysis on the individual use of park-and-ride facilities for varying parking costs and transit fares. However, we did not find any significant change in the behavior of park-and-ride location choice.



(a) Uncongested conditions (b) Congested conditions

Figure 3.13: Sensitivity analysis on transit fare and parking cost

Chapter 4

Schedule-based transit assignment with online information

This chapter considers the SB transit assignment with online information. For this purpose, it proposes models that compute average passenger flow on each link of the SB transit network rather than the flow for a particular realization of the network. For example, if we know the state of the network on a particular day, then we can just perform the assignment of passengers in a deterministic fashion using models described by previous studies [9, 53]. An average flow of passengers computed based on the historical states of the network will aid in evaluating the long-term congestion in the network.

The current chapter is divided into three sections. Section 4.1 and 4.2 describe the uncapacitated and capacitated assignment models respectively. Then, the results of numerical experiments based on proposed models are presented in Section 4.3. The

notations introduced in the previous chapter are also used in the current chapter.

4.1 Uncapacitated assignment

In the uncapacitated assignment, transit vehicles are assumed to have unlimited capacity. This assignment model is applicable for transit systems with low ridership but unreliable service, and for which denied boarding due to limited capacity is a rare phenomenon. The computation of optimal policy μ^* and expected costs $\hat{J}^*$ for the passenger routing in uncapacitated networks is performed for individual destinations using Algorithm 2 presented in Chapter 3. After this, for every group, we need to figure out the optimal departure time from their origins. We assume that passengers are rational and select the departure time, which provides them with the least expected cost to their destination. We present two different approaches to finding the optimal departure time:

1. *If early and late arrival penalties are included*: In this case, penalties are used to avoid arriving outside the desired travel time window. We consider two different types of penalties, i.e., early (η_1 / time units) and late arrival (η_2 / time units) penalties. Based on these penalties, the optimal departure t_g^* for group $g \in G$ is given as:

$$t_g^* \in \underset{t_g^{ED} \leq t \leq t_g^{ED}+\delta_3}{\operatorname{argmin}} \{\hat{J}^*(o_g,t) + \eta_1 * \max(0, t_g^{EA} - (t + \hat{J}^*(o_g,t))) + \eta_2 \max(0, (t + \hat{J}^*(o_g,t)) - t_g^{LA})\} \quad (4.1)$$

   In equation (4.1), t_g^* is searched within the time interval $\left[t_g^{ED}, t_g^{ED} + \delta_3\right]$ for the least expected cost with associated penalties, where δ_3 is the maximum acceptable wait time. The passenger assignment, in this case, can performed using Algorithm 3.

2. *If early and late arrival penalties are not included*: In this case, optimal departure

t_g^* for group $g \in G$ is given as:

$$t_g^* \in \underset{t_g^{ED} \le t \le t_g^{ED}+\delta_3}{\operatorname{argmin}} \{\hat{J}^*(o_g, t)\} \tag{4.2}$$

In equation (4.2), t_g^* is searched within the time interval $\left[t_g^{ED}, t_g^{ED} + \delta_3\right]$ for the least expected cost, where δ_3 is the maximum acceptable wait time. The passenger assignment, in this case, can be performed using Algorithm 3 or the linear program (4.3) developed in the following paragraph.

If we do not include penalties for passenger arrival outside the desired window, then we can derive a linear program for the assignment of passengers on optimal policies. To do so, let us denote $v^d(i, t, \theta, j)$ as the number of passengers arriving at state $(i, t, \theta) \in S$ and choosing control $j \in u(i, t, \theta)$. Furthermore, let V_{gt}^d be the number of passengers from group g, departing at time $t \in \left[t_g^{ED}, t_g^{ED} + \delta_3\right]$ from their origin o_g to destination $d \in D$. Then, we have the following assignment LP:

$$\min_{\mathbf{V},\mathbf{v}} \quad \sum_{d\in D}\sum_{(i,t,\theta)\in S}\sum_{j\in u(i,t,\theta)} v^d(i,t,\theta,j)c^{\theta}_{ij} \tag{4.3a}$$

$$\text{s.t.} \quad \sum_{j\in u(i,t,\theta)} v^d(i,t,\theta,j) - p^{\theta} \sum_{\substack{(k,t^{'},\theta^{'})\in S\setminus\{d\}:i\in u(k,t^{'},\theta^{'}) \\ \&\ t=t^{'}+c^{\theta^{'}}_{ki}}} v^d(k,t^{'},\theta^{'},i) = 0, \forall (i,t,\theta)\in S, \forall d\in D \tag{4.3b}$$

$$\sum_{j\in u(o,t,\theta)} v^d(o,t,\theta,j) - p^{\theta} \sum_{\substack{g\in G: o_g=o\ \& \\ t\in\left[t^{ED}_g, t^{ED}_g+\delta_3\right]}} V^d_{gt} = 0, \forall \theta\in\Theta_o(t), \forall t\in T, \forall o\in O, \forall d\in D \tag{4.3c}$$

$$\sum_{t\in\left[t^{ED}_g, t^{ED}_g+\delta_3\right]} V^d_{gt} = d^{o_g d}_g, \forall g\in G: d_g = d, \forall d\in D \tag{4.3d}$$

$$\sum_{(k,t',\theta')\in S\setminus\{d\}:d\in u(k,t',\theta')} v^d(k,t^{'},\theta^{'},d) = \sum_{g\in G: d_g=d} d^{o_g d}, \forall d\in D \tag{4.3e}$$

$$v^d(i,t,\theta,j)\geq 0, \forall j\in u(i,t,\theta), \forall (i,t,\theta)\in S, \forall d\in D \tag{4.3f}$$

$$V^d_{gt}\geq 0, \forall t\in\left[t^{ED}_g, t^{ED}_g+\delta_3\right], \forall g\in G, \forall d\in D \tag{4.3g}$$

In the above optimization program (4.3), we minimize the total expected travel time given by (4.3a). Constraints (4.3b)-(4.3e) describe the conservation of flow for every destination. For any state, (4.3b) shows that the sum of flow going out from state (i,t,θ) is equal to the fraction of the sum of flow coming into it from other states at time t and observing θ. (4.3c) describes the conservation constraint for states associated to origin nodes, i.e., the sum of flow going out from any origin state (o,t,θ) is equal to the sum of the flow from different groups that have the same origin o departing at time t and experiencing the real-time information θ. Equation (4.3d) describes that the total sum of flow from a group at different departure times to their destination d should be equal to the demand associated with that group. Then, for every destination $d\in D$, the total flow coming into the destination state should be equal to the total demand of groups going to d. Finally, (4.3f)-(4.3g) represent the non-negativity constraints for the

flow variables.

Lemma 1. *The optimal solution of* (4.3) *assigns flow to the optimal policy corresponding to each destination.*

Proof. We will show this by deriving the KKT conditions of the assignment program (4.3). Let us associate dual variables $\{J^d(i,t,\theta)\}_{\forall\theta\in\Theta_i(t),\forall t\in\tilde{t}_{k(i)}(i),\forall i\in N,\forall d\in D}$, $\{J^d(o,t,\theta)\}_{\forall\theta\in\Theta_o(t),\forall t\in T,\forall o\in O,\forall d\in D}$, $\{J^d(g)\}_{\forall g\in G:d_g=d,\forall d\in D}$, $\{J^d(d)\}_{\forall d\in D}$, $\{\sigma^d(i,t,\theta,j)\}_{\forall j\in u(i,t,\theta),\forall(i,t,\theta)\in S,\forall d\in D}$, and $\{\lambda^d_{gt}\}_{\forall t\in\left[t^{ED}_g,t^{ED}_g+\delta_3\right],\forall g\in G,\forall d\in D}$ to the constraints (4.3b)-(4.3g) respectively. The, the Lagrangian of (4.3) can be written as:

$$\mathcal{L}(\mathbf{V},\mathbf{v},\mathbf{J},\sigma,\lambda)=\sum_{d\in D}\Bigg[\sum_{(i,t,\theta)\in S}\sum_{j\in u(i,t,\theta)}v^d(i,t,\theta,j)*c^{\theta}_{ij}+$$

$$\sum_{\substack{\forall\theta\in\Theta_i(t),\\ \forall t\in\tilde{t}_{k(i)}(i),\forall i\in N}}J^d(i,t,\theta)*\left(p^{\theta}\sum_{\substack{(k,t^{'},\theta^{'})\in S\setminus\{d\}:i\in u(k,t^{'},\theta^{'})\\ \&\ t=t^{'}+c^{\theta^{'}}_{ki}}}v^d(k,t^{'},\theta^{'},i)-\sum_{j\in u(i,t,\theta)}v^d(i,t,\theta,j)\right)+$$

$$\sum_{\forall\theta\in\Theta_o(t),\forall t\in T,\forall o\in O}J^d(o,t,\theta)*\left(p^{\theta}\sum_{\substack{g\in G:o_g=o\ \&\\ t\in\left[t^{ED}_g,t^{ED}_g+\delta_3\right]}}V^d_{gt}-\sum_{j\in u(o,t,\theta)}v^d(o,t,\theta,j)\right)+$$

$$\sum_{\forall g\in G:d_g=d}J^d(g)\left(d^{o_g d}_g-\sum_{t\in\left[t^{ED}_g,t^{ED}_g+\delta_3\right]}V^d_{gt}\right)+$$

$$J^d(d)*\left(\textstyle\sum_{g\in G:d_g=d}d^{o_g d}-\sum_{\substack{(k,t^{'},\theta^{'})\in S\setminus\{d\}:\\ d\in u(k,t^{'},\theta^{'})}}v^d(k,t^{'},\theta^{'},d)\right)-$$

$$\sum_{\forall j\in u(i,t,\theta),\forall(i,t,\theta)\in S}\sigma^d(i,t,\theta,j)*v^d(i,t,\theta,j)-\sum_{\forall t\in\left[t^{ED}_g,t^{ED}_g+\delta_3\right],\forall g\in G}\lambda^d_{gt}*V^d_{gt}\Bigg]$$

The KKT conditions are given below:

1. *Primal feasibility*: (4.3b)-(4.3g)

2. *Dual feasibility*:

$$\sigma^d_{i,t,\theta,j} \geq 0, \forall j \in u(i,t,\theta), \forall (i,t,\theta) \in S, \forall d \in D \tag{4.4}$$

$$\lambda^d_{gt} \geq 0, \forall t \in \left[t^{ED}_g, t^{ED}_g + \delta_3\right], \forall g \in G, \forall d \in D \tag{4.5}$$

3. *Complementary slackness*:

$$v^d(i,t,\theta,j) * \sigma^d(i,t,\theta,j) = 0, \forall j \in u(i,t,\theta), \forall (i,t,\theta) \in S \tag{4.6}$$

$$V^d_{gt} * \lambda^d_{gt} = 0, \forall t \in \left[t^{ED}_g, t^{ED}_g + \delta_3\right], \forall g \in G, \forall d \in D$$

4. *Gradient of the Lagrangian wrt primal variables vanishes*:

$$\frac{\partial \mathcal{L}(\mathbf{V}, \mathbf{v}, \mathbf{J}, \sigma, \lambda)}{\partial v^d(i,t,\theta,j)} = c^\theta_{ij} + \sum_{\theta' \in \Theta_j(t)} p^{\theta'} J^d(j, t + c^\theta_{ij}, \theta') - J^d(i,t,\theta) - \sigma^d(i,t,\theta,j) = 0, \forall j \in u(i,t,\theta), \forall (i,t,\theta) \in S, \forall d \in D$$

$$\frac{\partial \mathcal{L}(\mathbf{V}, \mathbf{v}, \mathbf{J}, \sigma, \lambda)}{\partial V^d_{gt}} = -\sum_{\theta \in \Theta_{o_g}(t)} p^\theta J^d(o,t,\theta) - J^d(g) - \lambda^d_{gt} = 0, \forall t \in \left[t^{ED}_g, t^{ED}_g + \delta_3\right], \forall g \in G, \forall d \in D$$

Using (4.4), we can write above two equations as:

$$c_{ij}^{\theta} + \sum_{\theta' \in \Theta_j(t)} p^{\theta'} J^d(j, t + c_{ij}^{\theta}, \theta') - J^d(i,t,\theta) \geq 0, \forall j \in u(i,t,\theta), \forall (i,t,\theta) \in S, \forall d \in D \tag{4.7a}$$

$$- \sum_{\theta \in \Theta_{o_g}(t)} p^{\theta} J^d(o,t,\theta) - J^d(g) \geq 0, \forall t \in \left[t_g^{ED}, t_g^{ED} + \delta_3\right], \forall g \in G, \forall d \in D \tag{4.7b}$$

(4.7a) and (4.7b) can further be written as:

$$J^d(i,t,\theta) = \min_{j \in u(i,t,\theta)} \left\{ c_{ij}^{\theta} + \sum_{\theta' \in \Theta_j(t)} p^{\theta'} J^d(j, t + c_{ij}^{\theta}, \theta') \right\}, \forall (i,t,\theta) \in S, \forall d \in D \tag{4.8a}$$

$$J^d(g) = \min_{t \in \left[t_g^{ED}, t_g^{ED} + \delta_3\right]} \left\{ \sum_{\theta \in \Theta_{o_g}(t)} p^{\theta} J^d(o,t,\theta) \right\}, \forall g \in G, \forall d \in D \tag{4.8b}$$

(4.8a) and (4.8b) are the Bellman equations for finding the optimal policies given in (3.4) and (4.2) respectively. This completes our proof. □

The dual variables $J^d(i,t,\theta)$ of (4.3b)-(4.3c) represent the optimal cost to go from state (i,t,θ) to d. Similarly, the dual variables $J^d(g)$ of (4.3d) represent the optimal cost incurred by group g to go from its origin to destination. In fact, the dual program of (4.3) is the linear programming formulation for solving the Bellman equations (3.4) and (4.2). One of the advantages of assignment LP (4.3) is that it is decomposable for every destination $d \in D$ and side constraints related to the flow of passengers can be used in this formulation. However, the number of variables can still be large due to the cardinality of state space. Therefore, it is much easier to use Algorithm 3 presented in the next paragraph for sole assignment purposes.

Algorithm 3 starts with the initialization of the transitioning flows $v^d(i,t,\theta,j)$ for individual destinations. Then, for every destination $d \in D$, it computes the optimal cost

functions $\hat{J}^{d*}$ and optimal policy μ^{d*}. After that, for every group, we find the optimal departure time(s) from their origin using (4.1) or (4.2). Then, for every possible real-time information vector they could receive for a given optimal departure time, we assign a fraction of the group demand to that transitioning flow variable. We repeat this for every group. After this, the total demand has already been originated in the transition graph corresponding to μ^{d*}. Using the topological order of nodes and processing the times in chronological order, we assign the transitioning flow using Line 18. Note that after assignment, we can calculate the average flow on a link $(i, j) \in A$ as below:

$$v_{ij} = \sum_{d \in D} \sum_{t \in \tilde{t}_{k(i)}(i)} \sum_{\theta \in \Theta_i(t)} v^d(i, t, \theta, j), \forall (i, j) \in A \tag{4.9}$$

Algorithm 3 Uncapacitated transit assignment

(Initialization) $v^d(i,t,\theta,j) \leftarrow 0, \forall j \in u(i,t,\theta), \forall (i,t,\theta) \in S, \forall d \in D$
for every $d \in D$ **do**
  $\hat{J}^{d*}, \mu^{d*} \leftarrow ULC(d)$
  if arrival penalties are included **then**
    $t_g^* = \underset{t_g^{ED} \le t \le t_g^{ED}+\delta_3}{\operatorname{argmin}} \{\hat{J}^*(o_g,t) + \eta_1 * \max(0, t_g^{EA} - (t + \hat{J}^*(o_g,t))) + \eta_2 \max(0, (t + \hat{J}^*(o_g,t)) - t_g^{LA})\}$
  else
    $t_g^* = \underset{t_g^{ED} \le t \le t_g^{ED}+\delta_3}{\operatorname{argmin}} \{\hat{J}^*(o_g,t)\}$
  for every $g \in G : d_g = d$ **do**
    for $t \in t_g^*$ **do**
      for $\theta \in \Theta_{o_g}(t)$ **do**
        $j \leftarrow \mu^{d*}(o_g,t,\theta)$
        $v^d(o_g,t,\theta,j)$ += $p^\theta * d_g^{o_g d} * \frac{1}{|t_g^*|}$
  Find the topological order of nodes
  for $i \in$ topological order **do**:
    for $t \in \tilde{t}_{k(i)}(i)$ **do**
      for $\theta \in \Theta_i(t)$ **do**
        $j \leftarrow \mu^{d*}(i,t,\theta)$
        $$v^d(i,t,\theta,j) \leftarrow v^d(i,t,\theta,j) + p^\theta \left(\sum_{\substack{(k,t',\theta') \in S \setminus \{d\} : \mu^{d*}(k,t',\theta') = i \\ \& \ t = t' + c_{ki}^{\theta'}}} v^d(k,t',\theta',i) \right)$$

To explain the calculations of the uncapciatated assignment problem, we present an example problem based on Figure 3.1 provided in Chapter 3. Let us compute the optimal cost functions and optimal policy for destination d. Clearly, $\hat{J}^*(d) = 0$. $\hat{J}^*(C_1, 17) = 1 * (1 + 0) = 1$, $\hat{J}^*(C_1, 23) = 1 * (1 + 0) = 1$, $\hat{J}^*(C_2, 16) = 1 * (1 + 0) = 1$, $\hat{J}^*(C_2, 18) = 1 * (1 + 0) = 1$, and $\hat{J}^*(C_2, 23) = 1 * (1 + 0) = 1$.

$$\hat{J}^*(D_2,3) = 1*(13+\hat{J}^*(C_2,16)) = 1*(13+1) = 14$$

$$\hat{J}^*(D_2,5) = 1*(13+\hat{J}^*(C_2,18)) = 1*(13+1) = 14$$

$$\hat{J}^*(D_2,10) = 1*(13+\hat{J}^*(C_2,23)) = 1*(13+1) = 14$$

$$\begin{aligned}\hat{J}^*(B_1,2) &= 0.2*\min\{15+\hat{J}^*(C_1,17), 1+\hat{J}^*(D_2,3)\} + 0.3*\min\{15+\hat{J}^*(C_1,17), 3+\hat{J}^*(D_2,5)\} \\ &\quad + 0.5*\min\{15+\hat{J}^*(C_1,17), 8+\hat{J}^*(D_2,10)\} \\ &= 0.2*15 + 0.3*16 + 0.5*16 = 15.8\end{aligned}$$

$$\begin{aligned}\hat{J}^*(B_1,8) &= 0.5*\min\{15+\hat{J}^*(C_1,15+8), \infty\} + 0.5*\min\{15+\hat{J}^*(C_1,15+8), 2+\hat{J}^*(D_2,8+2)\} \\ &= 0.5*16 + 0.5*16 = 16\end{aligned}$$

$$\hat{J}^*(A_1,0) = 0.6*(2+15.8) + 0.4*(8+16) = 20.28$$

$$\hat{J}^*(E_2,0) = 0.2*(3+14) + 0.3*(5+14) + 0.5*(10+14) = 21.1$$

$$\hat{J}^*(o,0) = \min\{20.28, 21.1\} = 20.28$$

After computing the expected cost to go from various nodes at various times, one can evaluate the optimal policy by comparing these optimal costs. These are evaluated below:

$$\mu^*(o,0,\{0,0\}) = \{A_1\}$$

$$\mu^*(A_1,0,\{2\}) = \{B_1\}, \quad \mu^*(A_1,0,\{8\}) = \{B_1\}$$

$$\mu^*(E_2,0,\{3\}) = \{D_2\}, \quad \mu^*(E_2,0,\{5\}) = \{D_2\} \quad \mu^*(E_2,0,\{10\}) = \{D_2\}$$

$$\mu^*(B_1,2,\{15,1\}) = \{D_2\}, \quad \mu^*(B_1,2,\{15,3\}) = \{C_1\} \quad \mu^*(B_1,2,\{15,8\}) = \{C_1\}$$

$$\mu^*(B_1,8,\{15,\infty\}) = \{C_1\}, \quad \mu^*(B_1,8,\{15,2\}) = \{D_2,C_1\}$$

$$\mu^*(D_2,3,\{13\}) = \{C_2\}, \quad \mu^*(D_2,5,\{13\}) = \{C_2\} \quad \mu^*(D_2,10,\{13\}) = \{C_2\}$$

$$\mu^*(C_1,17,\{1\}) = \{d\}, \quad \mu^*(C_1,23,\{1\}) = \{d\}$$

$$\mu^*(C_2,16,\{1\}) = \{d\}, \quad \mu^*(C_2,18,\{1\}) = \{d\} \quad \mu^*(C_2,23,\{1\}) = \{d\}$$

Let us assume only one group of 100 passengers moving from o to d. For the sake of simplicity, we do not consider any arrival time penalties. Obviously, $t^* = 0$. Further, we can evaluate the values of transitioning flow at various states as below:

$$
\begin{aligned}
v(o,0,\{0,0\},A_1) &= 100 & v(o,0,\{0,0\},E_2) &= 0 & & \\
v(A_1,0,\{2\},B_1) &= 100*0.6 = 60, & v(A_1,0,\{8\},B_1) &= 100*0.4 = 40 & & \\
v(E_2,0,\{3\},D_2) &= 0, & v(E_2,0,\{5\},D_2) &= 0 & v(E_2,0,\{10\},D_2) &= 0 \\
v(B_1,2,\{15,1\},C_1) &= 0, & v(B_1,2,\{15,1\},D_2) &= 0.2*60 = 12 & & \\
v(B_1,2,\{15,3\},C_1) &= 0.3*60 = 18, & v(B_1,2,\{15,3\},D_2) &= 0 & & \\
v(B_1,2,\{15,8\},C_1) &= 0.5*60 = 30 & v(B_1,2,\{15,8\},D_2) &= 0 & & \\
v(B_1,8,\{15,\infty\},C_1) &= 0.5*40 = 20, & v(B_1,8,\{15,\infty\},D_2) &= 0 & & \\
v(B_1,8,\{15,2\},C_1) &= 0.5*0.5*40 = 10, & v(B_1,8,\{15,2\},D_2) &= 0.5*0.5*40 = 10 & & \\
v(D_2,3,\{13\},C_2) &= 12, & v(D_2,5,\{13\},C_2) &= 0 & v(D_2,10,\{13\},C_2) &= 10 \\
v(C_1,17,\{1\},d) &= 48, & v(C_1,23,\{1\},d) &= 30 & & \\
v(C_2,16,\{1\},d) &= 12, & v(C_2,18,\{1\},d) &= 0 & v(C_2,23,\{1\},d) &= 10
\end{aligned}
$$

Computing the average link flow on various links using (4.9), we have, $v(o,A_1) = 100, v(o,E_2) = 0, v(A_1,B_1) = 100$, $v(E_2,D_2) = 0$, $v(B_1,D_2) = 12{+}20 = 22$, $v(B_1,C_1) = 78$, $v(D_2,C_2) = 22$, $v(C_1,d) = 78$, $v(C_2,d) = 22$.

4.2 Capacitated Assignment

The uncapacitated assignment model may produce unrealistic passenger flows on various transit routes. This is because of the limited capacity of vehicles, due to which some arcs may become saturated and thus inaccessible depending on the route choice of other passengers. If we assume that passengers mingle at nodes and have equal probability to access outgoing links, then one can include the following capacity constraint in the assignment program (4.3) to produce capacity-feasible flows:

$$\sum_{d\in D}\sum_{t\in\tilde{t}_{k(i)}(i)}\sum_{\theta\in\Theta_i(t)} v^d(i,t,\theta,j) \le u_a, \forall (i,j)\in A_v \tag{4.10}$$

where u_a is the capacity associated with transit vehicle used to serve link $a \in A_v$. However, doing so would heuristically bound the predicted flows without providing a realistic model of the risk of failing to board an arc and strategic behavior induced by that. Moreover, passengers in the transit network have continuance priority over other passengers. The above constraint would not be able to capture such behavior. To model such behavior, previous studies have proposed to use *failure-to-board* probabilities or *access* probabilities. They evaluate the probability with which a passenger waiting at a bus stop can access an outgoing link. Such access probabilities result in multiple paths that a traveler can take with positive probability. In the capacitated assignment, the hyperpaths/strategies are induced by both risks of denied boarding due to limited capacity and missing transfers due to unreliable service. A strategy helps passengers minimize their expected costs under various types of uncertainties. When passengers employ strategies to move between various origin-destination pairs and compete for the limited capacity, the strategic equilibrium occurs when no passenger can improve her expected cost by unilaterally switching to a different strategy. To compute the optimal strategy in the capacitated assignment, Chapter 3 formulated a Bellman equation 3.14 that can be solved using a label correcting algorithm. The overall steps of finding the optimal policy using a label correcting algorithm for the capacitated case are summarized in Algorithm 4 from Line 1-12. The worst-case complexity remains the same as $\mathcal{O}(|\hat{S}_C||A|)$. However, the state space is bigger in this case. The optimal policy μ_C^* is calculated using the optimal cost labels in Line 13. Further, the optimal cost of taking a certain action at any state Q_C^* is evaluated in line 14.

Algorithm 4 Label correcting algorithm for capacitated assignment

1: **procedure** CLC(d)
2: *(Initialize)* $\hat{J}_C(i,t) \leftarrow \infty, \forall (i,t) \in \hat{S} \backslash \{d\}$ and $\hat{J}(d) \leftarrow 0$
3: $SE \leftarrow BS(d)$
4: **while** $SE \neq \phi$ **do**
5: Remove an element i from SE
6: **for** $t \in \tilde{t}_{k(i)}(i)$ **do**
7: $tempJ \leftarrow 0$
8: **for** $\theta \in \Theta_i(t)$ **do**
9: **for** $x \in X_i^{\theta}(t)$ **do**
10: $tempJ \mathrel{+}= p^{\theta}\pi^{x} \min_{j \in u_C(i,t,\theta,x)} \left\{ c_{ij}^{\theta} + \hat{J}_C(j, t + c_{ij}^{\theta}) \right\}$
11: **if** $tempJ < \hat{J}_C(i,t)$ **then**
12: $\hat{J}_C(i,t) \leftarrow tempJ$; $SE \leftarrow SE \cup BS(i)$
13: $\mu_C^*(i,t,\theta,x) \leftarrow \underset{j \in u_C(i,t,\theta,x)}{\operatorname{argmin}} \{c_{ij}^{\theta} + \hat{J}_C^*(j, t + c_{ij}^{\theta})\}, \forall (i,t,\theta,x) \in S_C \backslash \{d\}$
14: $Q_C^*(s,j) \leftarrow c_{ij}^{\theta} + \hat{J}_C^*(j, t + c_{ij}^{\theta}), \forall j \in u_C(s), \forall s = (i,t,\theta,x) \in S_C$
15: $P_{s,j} \leftarrow 1.0/|\mu^*(s)|, \forall j \in \mu^*(s), \forall s = (i,t,\theta,x) \in S_C$
16: **for** every $g \in G : d_g = d$ **do**
17: **if** arrival penalties are included **then**
18: $t_g^* \leftarrow \underset{t_g^{ED} \leq t \leq t_g^{ED} + \delta_3}{\operatorname{argmin}} \{\hat{J}^*(o_g,t) + \eta_1 * \max(0, t_g^{EA} - (t + \hat{J}^*(o_g,t))) + \eta_2 \max(0, (t + \hat{J}^*(o_g,t)) - t_g^{LA})\}$
19: **else**
20: $t_g^* \leftarrow \underset{t_g^{ED} \leq t \leq t_g^{ED} + \delta_3}{\operatorname{argmin}} \{\hat{J}^*(o_g,t)\}$
21: $R_{g,t} \leftarrow 1.0/|t_g^*|, \forall t \in t_g^*$
22: **return** $\hat{J}_C^*, Q_C^*, \mu^*, P, R$

The route choice of passengers is characterized by the probability of taking an action at a particular state. Therefore, we introduce $P = \{P_{s,a}^d\}$ as the probability of taking an action $a \in u_C(s), \forall s \in S$, when going to destination $d \in D$ and $R = \{R_{g,t}\}$ as the probability of group $g \in G$ departing at time $t \in \left[t_g^{ED}, t_g^{ED} + \delta_3\right]$. These route choice probabilities are calculated in Algorithm 4 from lines 15-21. We can observe that when there are multiple actions $j \in \mu^*(i,t,\theta,x)$ at state (i,t,θ,x) that achieve optimal expected cost, we assign equal probability to each optimal action. Similarly, if

multiple departure times provide the same optimal expected cost for group $g \in G$, we assign equal probabilities to these departure times. However, if there is a single action that achieves minimum, then the probabilities are degenerate. The use of route choice probabilities allows us to use a flexible choice of selecting actions at various states. For example, one can use the logit-based route choice probabilities.

4.2.1 Network loading

The computation of optimal policies/strategies for individual destinations reveals the number of passengers using a specific strategy (since we know the number of passengers in a group, their departure time probabilities, and their route choice probabilities). In this section, we describe a NetworkLoading procedure that converts these strategic flows into link flows. The loading of passengers follows some behavioral rules that are described below:

1. At a transfer node, if a passenger according to her strategy, decides to continue on the same route r rather than taking a transfer or ending her trip, then that passenger should get the priority over other passengers who either want to transfer to r or begin their journey with r. Such priority is known as *continuance priority* [48]. At any node, the passengers with continuance priority are loaded first onto the outgoing links.

2. We assume that passengers without continuance priority have equal access to the outgoing links, and they are processed in random and a uniformly distributed single queue. Such loading of passengers is also known as *random loading*. One could try other loading approaches such as *First-Come-First-Serve*, *Regret*, etc. [54] discusses such exogenous priority rules for transit assignment, which we leave for future research to explore.

The NETWORKLOADING procedure is summarized in Algorithm 5. It takes passenger route choice probabilities for various destinations $\{\hat{P}^d\}_{d\in D}$ and departure time probabilities for individual groups $\{\hat{R}_g\}_{g\in G}$ as inputs and outputs link flows $\mathbf{v}$ and availability probabilities π. The procedure starts by initializing the state-action passenger flows $\mathbf{v}^d$ for various destinations, node-time priority passenger flows $\mathbf{V}_p$ and non-priority passenger flows $\mathbf{V}_n$. After this, we assign the group flows at various departure times according to the departure time choice probabilities $\hat{\mathbf{R}}$ (Lines 4-6). Then, we process various nodes in topological order to load the passenger demand on outgoing links. For each node $i \in N\backslash\{d\}$, we start by assuming an availability vector, where all the outgoing links $u(i,t,\theta)$ are available, i.e., we assign $\pi^x = 1, \forall x = (i,t,\theta,\{1\}^{|u(i,t,\theta)|}) \in S_C, 0$,otherwise. Then, we start the loading of the demand reached node i onto outgoing links. Overall, this loading process can be divided into two phases. In the first phase, we assign the priority flows (Lines 11-23). Depending on the strategy at various states (i,t,θ,x) for individual destination d, a fraction of flow $tempFlow$ is assigned to outgoing link $(i,j) : j \in u_C(i,t,\theta,x)$ according to route choice probabilities $\hat{P}^d_{(i,t,\theta),j}$. Then, a fraction of $tempFlow$ is further assigned to node j and transitioning time $t^{'}$ either as priority or non-priority flow depending on the strategy and route choice probabilities. Of course, for the origin nodes, there will be no priority flow to be assigned. The second phase of the loading procedure at node i is the loading of non-priority flows. This loading is performed using a single-queue processing procedure described by [50] and [51] for static auto networks. We first calculate the residual capacity $\tilde{\mathbf{u}}$ of outgoing links after the loading of priority flows. Then, based on the route choice probabilities, we evaluate $\tilde{\mathbf{v}}$, which describes the number of passengers trying to access outgoing links. The flow trying to access a particular link $\tilde{v}_{ij}$ may exceed the residual capacity $\tilde{u}_{ij}$. Assuming that all the non-priority passengers waiting at that node have an equal probability of accessing an outgoing link, we evaluate the access probability $(\frac{\tilde{u}_{ij}}{\tilde{v}_{ij}})$ of that link. Then,

a minimum access probability $\tilde{\beta}$ of any outgoing link is calculated using the expression given in Line 34. $\tilde{\beta}$ describes the proportion of passengers that can be loaded before one or more outgoing links get saturated. If the accessing flow of any outgoing link does not exceed its residual capacity, then $\tilde{\beta} = 1$, which means all the waiting passengers can access their optimal choice. For assigning the appropriate number of passengers onto outgoing links, we repeat a similar procedure as priority flows, where some of the passengers reach the outgoing node as a priority and some as non-priority flow. We update the residual capacity and the number of passengers to be loaded at various times $U(i,t)$. If $\tilde{\beta} < 1$, we evaluate the saturated outgoing links, prepare the availability vector, and update its probability π^x using $\tilde{\beta}$. The availability probabilities are updated based on the principle that only the $\tilde{\beta}$ proportion of passengers will observe the current state, and the rest of the passengers $(1 - \tilde{\beta})$ will observe a different state. We continue updating the state availability probabilities in this manner until all the accessing flow is assigned. Note that due to assumption 13 and the presence of walking links from transfer nodes, we will never observe the availability vector, where all the outgoing links get saturated and are not available. This procedure will evaluate π's, which will be further used in the label correcting algorithm for updating the strategies in the assignment algorithm.

Algorithm 5 Network loading

procedure NETWORKLOADING($\hat{\mathbf{P}}, \hat{\mathbf{R}}$)

$v^d(s,j) \leftarrow 0, \forall j \in u_C(s), \forall s = (i,t,\theta,x) \in S_C, \forall d \in D$

$V_p(i,t) \leftarrow 0, V_n(i,t) \leftarrow 0, \forall t \in \tilde{t}_{k(i)}(i), \forall i \in N$

for $g \in G$ **do**

for $t \in \left[t_g^{ED}, t_g^{ED} + \delta_3\right]$ **do**

$V_n^d(o_g,t)$ += $\hat{R}_{g,t} * d_g^{o_g d}$

Find the topological order of nodes in N

for $i \in$ topological order **do**

stop$\leftarrow$ FALSE; $U_n(i,t) \leftarrow V_n(i,t); \forall t \in \tilde{t}_{k(i)}(i)$

$x_{(i,t,\theta)} \leftarrow \{1\}^{|u(i,t,\theta)|}; \pi^x \leftarrow 1,$ if $x = x_{(i,t,\theta)}, 0,$ otherwise, $\forall (i,t,\theta) \in S_C$

for $d \in D$ **do**

for $t \in \tilde{t}_{k(i)}(i)$ **do**

for $\theta \in \Theta_i(t)$ **do**

for $j \in u_C(i,t,\theta,x_{(i,t,\theta)}) : k(i) == k(j)$ **do**

$tempFlow \leftarrow p^{\theta} * \pi^{x_{(i,t,\theta)}} * \hat{P}^d_{(i,t,\theta,x_{(i,t,\theta)}),j} * V_p^d(i,t);\ t^{'} = t + c_{ij}^{\theta}$

$v^d(i,t,\theta,x_{(i,t,\theta)},j)$ += $tempFlow$

for $\theta^{'} \in \Theta_j(t^{'})$ **do**

$x_{(j,t^{'},\theta^{'})} \leftarrow \{1\}^{|u(j,t^{'},\theta^{'})|}$

for $l \in u_C(j,t^{'},\theta^{'},x_{(j,t^{'},\theta^{'})})$ **do**

if $k(j) == k(l)$ **then**

$V_p^d(j,t^{'})$ += $p^{\theta^{'}} * \pi^{x_{(j,t^{'},\theta^{'})}} * \hat{P}^d_{(j,t^{'},\theta^{'},x_{(j,t^{'},\theta^{'})}),l} * tempFlow$

else

$V_n^d(j,t^{'})$ += $p^{\theta^{'}} * \pi^{x_{(j,t^{'},\theta^{'})}} * \hat{P}^d_{(j,t^{'},\theta^{'},x_{(j,t^{'},\theta^{'})}),l} * tempFlow$

for $j \in FS(i)$ **do**

$\tilde{u}_{ij} \leftarrow \mathfrak{C}(k(j)) - \sum_{d\in D} \sum_{t\in\tilde{t}_{k(i)}(i)} \sum_{x\in X_i^{\theta}(t)} \sum_{\theta\in\Theta_i(t)} v^d(i,t,\theta,x,j)$

while not stop **do**

for $d \in D$ **do**

$\tilde{v}_{ij}^d \leftarrow 0, \forall j \in FS(i), \forall d \in D$

for $t \in \tilde{t}_{k(i)}(i)$ **do**

for $\theta \in \Theta_i(t)$ **do**

for $j \in u_C(i,t,\theta,x_{(i,t,\theta)})$ **do**

$\tilde{v}_{ij}^d$ += $p^{\theta} * \pi^{x_{(i,t,\theta)}} * \hat{P}^d_{(i,t,\theta,x_{(i,t,\theta)}),j} * U_n(i,t)$

$\tilde{v}_{ij} = \sum_{d \in D} \tilde{v}_{ij}^d$ ▷ Flow that'll be competing to access (i, j)

$\tilde{\beta} \leftarrow \min\left\{1, \min_{j \in FS(i)} \left(\frac{\tilde{u}_{ij}}{\tilde{v}_{ij}}\right)\right\}$

for $j \in FS(i)$ **do**

$\tilde{u}_{ij} = \tilde{u}_{ij} - \tilde{\beta}\tilde{v}_{ij}$

for $d \in D$ **do**

for $t \in \tilde{t}_{k(i)}(i)$ **do**

for $\theta \in \Theta_i(t)$ **do**

for $j \in u_C(i, t, \theta, x_{i,t,\theta})$ **do**

$tempFlow \leftarrow \tilde{\beta} * p^{\theta} * \pi^{x_{(i,t,\theta)}} * \hat{P}^d_{(i,t,\theta,x_{(i,t,\theta)}),j} * U_n^d(i, t);\ t^{'} = t + c_{ij}^{\theta}$

$v^d(i, t, \theta, x_{(i,t,\theta)}, j)$ += $tempFlow$

for $\theta^{'} \in \Theta_j(t^{'})$ **do**

for $l \in u_C(j, t^{'}, \theta^{'}, x_{(j,t^{'},\theta^{'})})$ **do**

if $k(j) == k(l)$ **then**

$V_p^d(j, t^{'})$ += $p^{\theta^{'}} * \pi^{x_{(j,t^{'},\theta^{'})}} * \hat{P}_{(j,t^{'},\theta^{'},x_{(j,t^{'},\theta^{'})}),l} * tempFlow$

else

$V_n^d(j, t^{'})$ += $p^{\theta^{'}} * \pi^{x_{(j,t^{'},\theta^{'})}} * \hat{P}_{(j,t^{'},\theta^{'},x_{(j,t^{'},\theta^{'})}),l} * tempFlow$

$U_n^d(i, t) \leftarrow (1 - \tilde{\beta}) V_n^d(i, t)$

if $\tilde{\beta} < 1$ **then**

for $j^{'} \in \text{argmin}_{j \in FS(i)} \left(\frac{\tilde{u}_{ij}}{\tilde{v}_{ij}}\right)$ **do**

for $t \in \tilde{t}_{k(i)}(i)$ **do**

for $\theta \in \Theta_i(t)$ **do**

$p \leftarrow \pi^{x_{(i,t,\theta)}}$ ▷ Probability of current state

$\pi^{x_{(i,t,\theta)}} \leftarrow \tilde{\beta} p$ ▷ Update probability of current state

$x_{(i,t,\theta)}[j^{'}] \leftarrow 0$ ▷ New state

$\pi^{x_{(i,t,\theta)}} \leftarrow (1 - \tilde{\beta}) p$ ▷ Update prob of new state

else

stop ← TRUE

return $\pi, \mathbf{v}$

To explain the concept of network loading in capacitated assignment problem, let us consider the example given in Section 4.1. Using the policy computed in the previous sub-section, we evaluate the route choice probabilities as below:

$$P_{(o,0,\{0,0\},\{1,1\}),A_1} = 1 \qquad P_{(o,0,\{0,0\},\{1,1\}),E_2} = 0$$

$$P_{(o,0,\{0,0\},\{0,1\}),E_2} = 1$$

$$P_{(A_1,0,\{2\},\{1\}),B_1} = 1, \qquad P_{(A_1,0,\{8\},\{1\}),B_1} = 1$$

$$P_{(E_2,0,\{3\},\{1\}),D_2} = 1, \qquad P_{(E_2,0,\{5\},\{1\}),D_2} = 1 \qquad P_{(E_2,0,\{10\},\{1\}),D_2} = 1$$

$$P_{(B_1,2,\{15,1\},\{1,1\}),C_1} = 0, \qquad P_{(B_1,2,\{15,1\},\{1,1\}),D_2} = 1$$

$$P_{(B_1,2,\{15,3\},\{1,1\}),C_1} = 1, \qquad P_{(B_1,2,\{15,3\},\{1,1\}),D_2} = 0$$

$$P_{(B_1,2,\{15,8\},\{1,1\}),C_1} = 1 \qquad P_{(B_1,2,\{15,8\},\{1,1\}),D_2} = 0$$

$$P_{(B_1,8,\{15,\infty\},\{1,0\}),C_1} = 1,$$

$$P_{(B_1,8,\{15,2\},\{1,1\}),C_1} = 0.5, \qquad P_{(B_1,8,\{15,2\},\{1,1\}),D_2} = 0.5$$

$$P_{(D_2,3,\{13\},\{1\}),C_2} = 1, \qquad P_{(D_2,5,\{13\},\{1\}),C_2} = 1 \qquad P_{(D_2,10,\{13\},\{1\}),C_2} = 1$$

$$P_{(C_1,17,\{1\},\{1\}),d} = 1, \qquad P_{(C_1,23,\{1\},\{1\}),d} = 1$$

$$P_{(C_2,16,\{1\},\{1\}),d} = 1, \qquad P_{(C_2,18,\{1\},\{1\}),d} = 1 \qquad P_{(C_2,23,\{1\},\{1\}),d} = 1$$

Further, assume that the capacity of both trips is 60. For passenger loading, let us process the nodes in topological order. For node 1, we have, $V_n(o,0) = 100$. First, initialize $\pi^{(o,0,\{0,0\},\{1,1\})} = 1$. The accessing flow and residual capacities of outgoing links are $\tilde{v}_{o,A_1} = 100, \tilde{v}_{o,E_2} = 0$ and $\tilde{u}_{o,A_1} = 60, \tilde{u}_{o,E_2} = 60$. As the flow trying to access link (o, A_1) is more than its residual capacity, we have $\beta = 0.6$. Since, we have

$\beta = 0.6$, we have to run more than 1 iteration of the "while loop" to finish the loading at node A_1. As we know that $P_{(o,0,\{0,0\},\{1,1\}),A_1} = 1$, $P_{(o,0,\{0,0\},\{1,1\}),E_2} = 0$, we have $v_{(o,0,\{0,0\},\{1,1\}),A_1} = 0.6 * 100 = 60, v_{(o,0,\{0,0\},\{1,1\}),E_2} = 0$. Since all the flow that reaches A_1 want to continue on the same route, we assign $V_p(A_1, 0) = 60$. This gives us $\pi^{(o,0,\{0,0\},\{1,1\})} = 0.6$ and $\pi^{(o,0,\{0,0\},\{0,1\})} = 0.4$. After updating the state availability probabilities, we now run the second iteration of the while loop. The accessing flow and residual capacities of available outgoing links are $\tilde{v}_{o,E_2} = 40$ and $\tilde{u}_{o,E_2} = 60$. Therefore, $\beta = 1$. This means that all flow can access their first available choice, i.e., $v_{(o,0,\{0,0\},\{0,1\}),E_2} = 1.0 * 40 = 40$. Since all the flow that reaches E_2 want to continue on the same route, we assign $V_p(E_2, 0) = 40$.

The next node in the topological order is A_1. Since all the flow that needs to be assigned at this node is priority flow, we have $v_{(A_1,0,\{2\},\{1\}),B_1} = 0.6 * 60 = 36$ and $v_{(A_1,0,\{8\},\{1\}),B_1} = 0.4 * 60 = 24$. This makes $V_p(B_1, 2) = 0.8 * 36 = 28.8$ and $V_n(B_1, 2) = 0.2 * 36 = 7.2$. Similarly, $V_p(B_1, 8) = 24 * 0.5 * 1 + 24 * 0.5 * 0.5 = 18$ and $V_n(B_1, 8) = 6$. Processing the node E_2, we have $v_{(E_2,0,\{3\},\{1\}),D_2} = 0.2 * 40 = 8, v_{(E_2,0,\{5\},\{1\}),D_2} = 0.3 * 40 = 12$, and $v_{(E_2,0,\{10\},\{1\}),D_2} = 0.5 * 40 = 20$. Therefore, $V_p(D_2, 3) = 8, V_p(D_2, 5) = 12$, and $V_p(D_2, 10) = 20$.

The next node in the topological order is B_1. This is an important node as it has both priority as well as non-priority flow to assign. Let's start with the assignment of priority flow. We have, $v_{(B_1,2,\{15,1\},\{1,1\}),C_1} = 28.8$ and $v_{(B_1,8,\{15,1\},\{1,1\}),C_1} = 18$. Clearly, $V_p(C_1, 17) = 28.8$ and $V_p(C_1, 23) = 18$. Next, we process the non-priority flow. We have accessing flow $\tilde{v}_{(B_1,D_2)} = 7.2 + 6 = 13.2$, residual capacity $\tilde{u}_{(B_1,D_2)} = 60 - 40 = 20$, and $\beta = 1$. This gives $v_{(B_1,2,\{15,1\},\{1,1\}),D_2} = 7.2$ and $v_{(B_1,8,\{15,2\},\{1,1\}),D_2} = 6$. Following the same procedure, we have, $v_{(D_2,3,\{13\},\{1\}),C_2} = 15.2$, $v_{(D_2,3,\{5\},\{1\}),C_2} =$

12, $v_{(D_2,3,\{10\},\{1\}),C_2} = 26$.

Calculating the average flow, we have, $v(o, A_1) = 60, v(o, E_2) = 40$, $v(A_1, B_1) = 60$, $v(E_2, D_2) = 40$, $v(B_1, D_2) = 13.2$, $v(B_1, C_1) = 46.8$, $v(D_2, C_2) = 53.2$, $v(C_1, d) = 46.8$, $v(C_2, d) = 53.2$.

4.2.2 Assignment of passengers

The optimal strategy computed using Algorithm 4 helps evaluate the route choice $\mathbf{P}$ and departure time choice $\mathbf{R}$ probabilities using the probability of availability vectors π. Then, the NetworkLoading procedure in Algorithm 5 will update the values of π based on $\mathbf{P}$ and $\mathbf{R}$. An equilibrium is reached when, in each state, no passenger can improve her expected cost by changing the probability of taking a particular action in that state. This means that in equilibrium, all non-null choice probabilities $P^d_{s,a}$ and $R^d_{g,t}$ associated to a state and group resp. will have the same expected costs $Q^d_{s,a}$ and $\hat{J}^d_{o_g,t}$. To characterize equilibrium, let us define the feasible set of route choice and departure time choice probabilities $\mathfrak{P}$ as below:

$$\mathfrak{P} = \Bigg\{ \mathbf{P} \times \mathbf{R} \in \mathbb{R}^{|D| \times \sum_{s \in S_C} |u_C(s)|} \times \mathbb{R}^{|G| \times |T|} : \sum_{j \in u_C(s)} P^d_{s,j} = 1, \forall s \in S_C, \forall d \in D, \text{ and } \sum_{t \in \left[t^{ED}_g, t^{ED}_g + \delta_3\right]} R_{g,t} = 1, \forall g \in G \Bigg\} \tag{4.11}$$

Further, the expected cost of choice probability vector $(\mathbf{P}, \mathbf{R})$ denoted by $C(\mathbf{P}, \mathbf{R})$ can be evaluated using the following equation:

$$C(\mathbf{P}, \mathbf{R}) = \sum_{d \in D} \sum_{s \in S_C} \sum_{j \in u_C(s)} Q^d(s, j) \times P^d_{s,j} + \sum_{g \in G} \sum_{t \in \left[t^{ED}_g, t^{ED}_g + \delta_3\right]} \hat{J}^d(o_g, t) \times P_{g,t} \tag{4.12}$$

We call $(\mathbf{P}^*, \mathbf{R}^*)$ as the equilibrium probabilities if they satisfy the variational inequality given as:

$$\left\langle C(\mathbf{P}^*, \mathbf{R}^*), \begin{Bmatrix} \mathbf{P}^* - \mathbf{P} \\ \mathbf{R}^* - \mathbf{R} \end{Bmatrix} \right\rangle \leq 0, \forall (\mathbf{P}, \mathbf{R}) \in \mathfrak{P} \tag{4.13}$$

Since the expected cost of mapping $C(\mathbf{P}, \mathbf{R})$ cannot be evaluated in closed form as it depends on the availability probabilities π through the loading procedure, we cannot formulate the above VI problem into an equivalent optimization problem. However, there exists at least one solution to this VI problem because the set $\mathfrak{P}$ is compact, and mapping $C(\mathbf{P}, \mathbf{R})$ is continuous since it depends on the availability probabilities π, which is a function of continuous $(\mathbf{P}, \mathbf{R})$ [51]. Moreover, we cannot show that there exists a unique solution to the given variational inequality. To solve the assignment problem, we use an MSA-based heuristic approach. We start by initializing the entries of the initial $(\hat{\mathbf{P}}, \hat{\mathbf{R}})$ as zero. Before running the Algorithm 4, we assume that $\pi^{(i,t,\theta,x)} = 1,$ if $x = \{1\}^{|u(i,t,\theta)|}, 0$ otherwise $, \forall x \in X_i^{\theta}(t), \forall (i,t,\theta) \in S$. Then, we evaluate the best response choice probabilities $(\mathbf{P}, \mathbf{R})$ using the Algorithm 4, which are used for updating the current $(\hat{\mathbf{P}}, \hat{\mathbf{R}})$ based on the convex combination of the current values $(\hat{\mathbf{P}}, \hat{\mathbf{R}})$ and the best response $(\mathbf{P}, \mathbf{R})$ using $\alpha = \frac{1}{k+1}$, where k is the iteration number. Then, the updated $(\hat{\mathbf{P}}, \hat{\mathbf{R}})$ is used for the NetworkLoading procedure that further updates the availability probabilities π. We continue this procedure until the $gap(\hat{\mathbf{P}}, \hat{\mathbf{R}}, \mathbf{P}, \mathbf{R})$ calculated using (4.14) reaches below the tolerance level ϵ. The gap function is similar to the one used for the traffic assignment based on the user equilibrium principle. However, the latter uses link flow vectors, and the former uses the link choice probabilities. The overall MSA algorithm is summarized in Algorithm 6. The converged average link flow values can be calculated using (4.15).

$$gap(\hat{\mathbf{P}}, \hat{\mathbf{R}}, \mathbf{P}, \mathbf{R}) = \frac{\sum_{d \in D} \sum_{s \in S_C} \sum_{j \in u_C(s)} Q^d(s,j) \times (\hat{P}^d_{s,j} - P^d_{s,j}) + \sum_{g \in G} \sum_{t \in \left[t_g^{ED}, t_g^{ED} + \delta_3\right]} \hat{J}^d(o_g, t) \times (\hat{R}_{g,t} - R_{g,t})}{\sum_{d \in D} \sum_{s \in S_C} \sum_{j \in u_C(s)} Q^d(s,j) \times P^d_{s,j} + \sum_{g \in G} \sum_{t \in \left[t_g^{ED}, t_g^{ED} + \delta_3\right]} J^d(o_g, t) \times R_{g,t}} \tag{4.14}$$

$$v_{ij} = \sum_{d \in D} \sum_{t \in \tilde{t}_{k(i)}(i)} \sum_{\theta \in \Theta_i(t)} \sum_{x \in X_i^{\theta}(t)} v^d(i,t,\theta,x,j), \forall (i,j) \in A \tag{4.15}$$

Algorithm 6 Method of successive averages for capacitated assignment

procedure MSA(ϵ)
  (Initialization) $\hat{P}^d_{s,j} \leftarrow 0, \forall j \in u_C(s), \forall s \in S_C, \forall d \in D$
  $\hat{R}_{g,t} \leftarrow 0, \forall t \in \left[t^{ED}, t^{ED} + \delta_3\right], \forall g \in G$
  $k \leftarrow 0; gap \leftarrow \infty$
  while $gap > \epsilon$ **do**
    $\alpha = \frac{1}{k+1}$
    $\hat{J}^d, Q^d, \mu^d, P^d, R^d \leftarrow CLC(d), \forall d \in D$
    $\hat{\mathbf{P}} \leftarrow \alpha\hat{\mathbf{P}} + (1-\alpha)\hat{\mathbf{P}}$; $\hat{\mathbf{R}} \leftarrow \alpha\hat{\mathbf{R}} + (1-\alpha)\hat{\mathbf{R}}$
    $\pi, \mathbf{v} \leftarrow$NetworkLoading$(\hat{\mathbf{P}}, \hat{\mathbf{R}})$
    Calculate *gap* using the equation (4.14)
    $k \leftarrow k + 1$

Unlike previous studies on schedule-based transit assignment that maintains the flow vector based on a specific strategy for an origin-destination pair, the current research maintains the link flow vector based on local choice probabilities for each destination. This difference is akin to a path-based versus link-based algorithm for solving the traffic assignment problem.

4.3 Numerical experiments

For the application of proposed models, we consider the network given in Section 3.1.4 of Chapter 3. There are 6 origin-destination pairs in the network. A synthetic demand table is created for the assignment, which is shown in Table 4.1. It has 24 groups with different origins, destinations, earliest departure, and earliest and latest arrival time, with a total demand of 128 passengers.

Table 4.1: Demand table for transit assignment numerical experiment

Group	Origin	Dest.	t_g^{ED}	t_g^{EA}	t_g^{LA}	Dem.
1	1	6	09:55	10:07	10:12	5
2	1	6	09:55	10:10	10:15	6
3	1	6	09:50	10:04	10:09	4
4	1	6	10:05	10:17	10:22	8
5	1	6	10:05	10:20	10:25	2
6	1	6	10:00	10:14	10:19	2
7	1	6	09:58	10:27	10:32	8
8	1	6	10:00	10:24	10:30	9
9	1	6	10:05	10:37	10:42	11
10	1	5	09:52	10:05	10:10	1
11	1	5	10:05	10:15	10:20	5
12	1	5	09:55	10:25	10:30	4
13	1	5	09:57	10:36	10:40	5
14	7	6	09:58	10:10	10:15	7
15	7	6	10:08	10:20	10:25	2
16	7	6	10:15	10:35	10:40	4
17	7	5	10:13	10:33	10:38	4
18	7	5	10:03	10:18	10:23	7
19	14	6	10:05	10:14	10:20	9
20	14	6	10:13	10:22	10:28	5
21	14	6	10:20	10:36	10:40	8
22	14	5	10:05	10:12	10:18	4
23	14	5	10:10	10:20	10:25	2
24	14	5	10:15	10:30	10:36	6

4.3.1 Uncapacitated assignment

Based on the expected costs computed in Section 3.1.4, and whether or not late and early arrival penalty is being applied, we present the optimal departure time results for various passenger groups in Table 4.2. When the penalties are not applied, we look for a departure time that is above the earliest departure time and provides the least expected cost to the respective destination. It provides similar departure times for groups that have the same origin and similar earliest departure times. When the early and late arrival penalties are applied, we look for departure time that provides the least expected

cost based on (4.2). The penalties sometimes cause a passenger group to depart early or late to arrive at the destination in a given time interval.

Table 4.2: Optimal departure time of passenger groups

Group	Penalties not included	Penalties included	Group	Penalties not included	Penalties included
1	10:05	10:05	13	10:00	10:09
2	10:05	10:05	14	10:17	10:03
3	10:05	09:55	15	10:17	10:10
4	10:05	10:05	16	10:17	10:17
5	10:05	10:05	17	10:17	10:17
6	10:05	10:05	18	10:17	10:03
7	10:05	10:09	19	10:08	10:08
8	10:05	10:09	20	10:18	10:18
9	10:05	10:09	21	10:24	10:24
10	10:00	10:00	22	10:18	10:08
11	10:09	10:09	23	10:18	10:18
12	10:00	10:09	24	10:18	10:18

For assigning passengers, we use the departure times calculated based on the penalties. The passenger flow obtained after running Algorithm 3 is visualized in Figure 4.1. The flow of passengers on various links is varied according to the line width of the links in the figure. The transfer links are represented using dashed lines. If a link is not shown between two nodes, then either such link does not exist in the network or the flow of passengers on that link is zero. We can observe that most of the passengers prefer taking the first and second trips of various routes in the network. This is because most of the passenger groups have departure times from the origin closer to the departure times of the first and second trips of various transit routes departing from their origins. We observe the highest flow on the second trip of red and blue routes. This is because together these two routes connect both destinations (5 and 6) from origin 1. The passenger groups going from origin 14 to destination 6 prefer taking the orange route. The

passenger groups going from 1 to 5 or 6 prefer taking the transfer 3-4. However, we observe some flow on the first trip of the green route from origin 1 to destination 6 that takes a transfer to the orange route. The passengers going from origin 7 to destination 5 prefer taking the transfer 8-2 from violet to the red route. To go to destination 5, we observe some passengers taking transfers 12-4 and 3-4 from the second trip of orange and red routes to the third trip of the blue route. We do not observe a significant flow of passengers for the fourth trip of various transit routes. This is because most of the groups do not have a late departure time window as compared to the departure times of fourth trips of various routes departing from various origins.

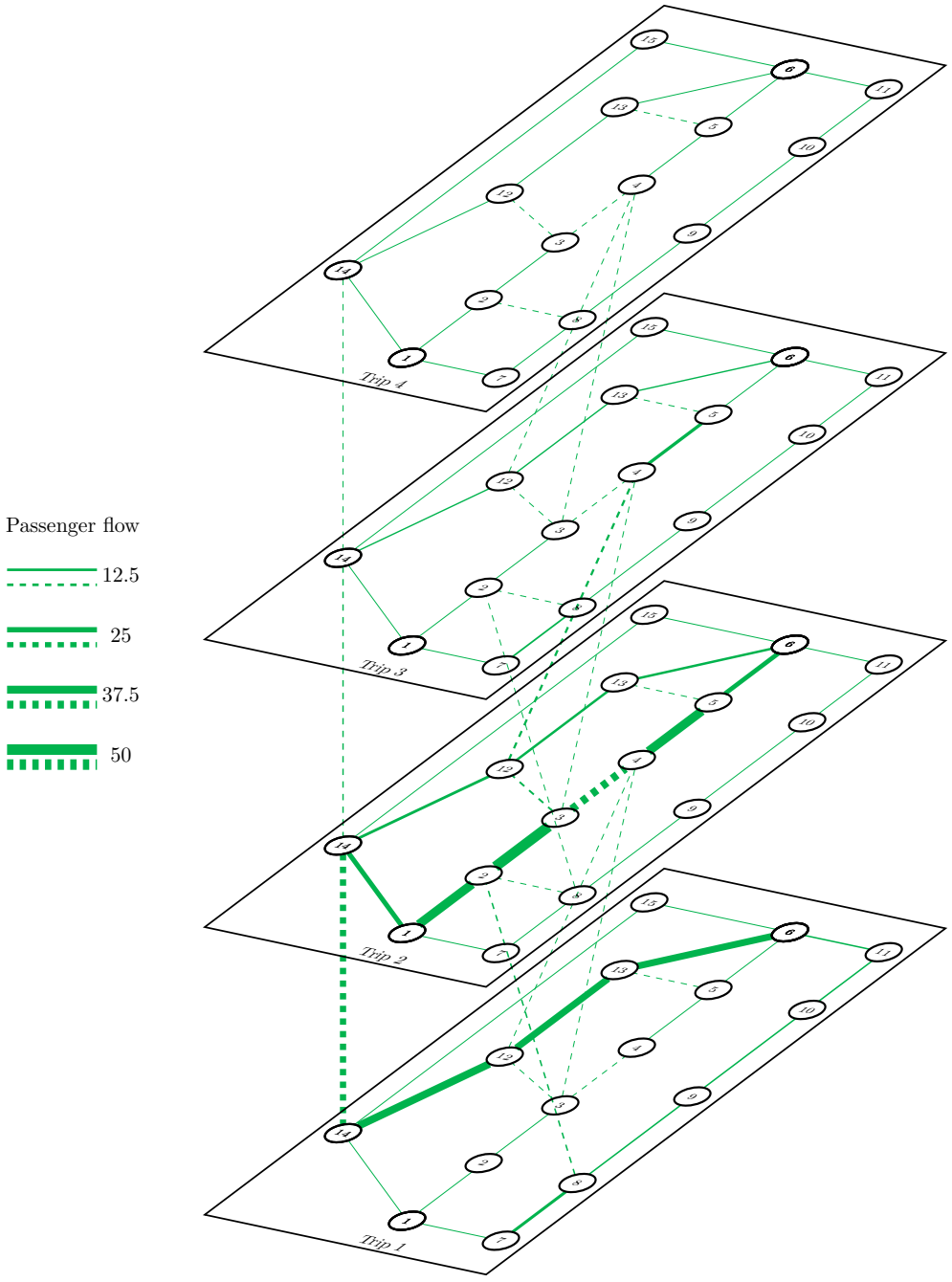


Figure 4.1: Passenger flow on various trips for uncapacitated transit assignment

4.3.2 Capacitated assignment

In this section, we present the results of the capacitated transit assignment. We ran the assignment Algorithm 6 with the gap tolerance value $\epsilon = 0.05\%$. It took 140 iterations and 8.5 minutes to converge to the solution up to the required tolerance gap. We plot the convergence behavior of the algorithm in Figure 4.2, where we can observe a steep decline of the gap value with every iteration. The overall convergence is achieved fairly quickly.

The converged departure time probabilities for various groups are visualized in Figure 4.3. Out of 24 groups, 10 groups have only one departure time, i.e., the probability of departing at a single departure time by these groups is 1. We further observe that 11 groups have two values in their departure time support and 3 groups have 3 or more values in their departure time support.

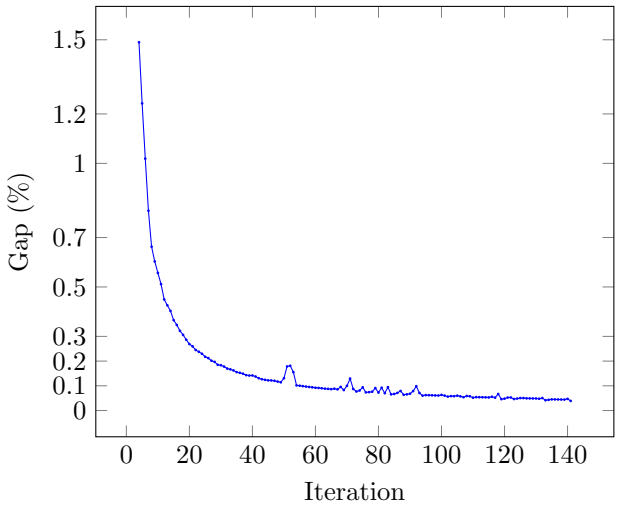


Figure 4.2: Convergence behavior of MSA algorithm

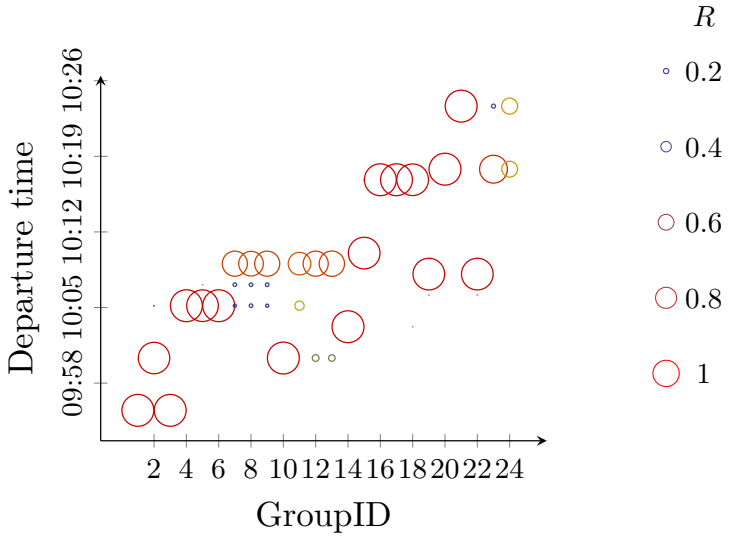


Figure 4.3: Converged values of departure time probabilities R

The uncapacitated assignment does not give us capacity-feasible flows. This is evident from the flow values visualized in Figure 4.1, where the first trip of the orange route and the second trip of red and violet routes carry more flow than their capacity (20 passengers). The capacitated assignment results in more realistic passenger flow on various trips and routes, which is visualized in Figure 4.4. Due to the limited capacity of transit routes, passengers have to shift from their most preferred choice to other choices. By looking at the figure, we can see that various segments of many attractive trip options are running at near or full capacity. This includes the first trip of the orange route, the second trip of red, green, orange, and blue routes, and the third trip of violet and blue routes. A lot of passengers taking the first trip of the orange route in case of the uncapacitated assignment are distributed to the second and third trip of the same route. To travel between 1-6, the orange route and combination of red and blue routes are the most popular choices. Both choices include one transfer provides improved expected costs as compared to direct routes (green and violet). To travel

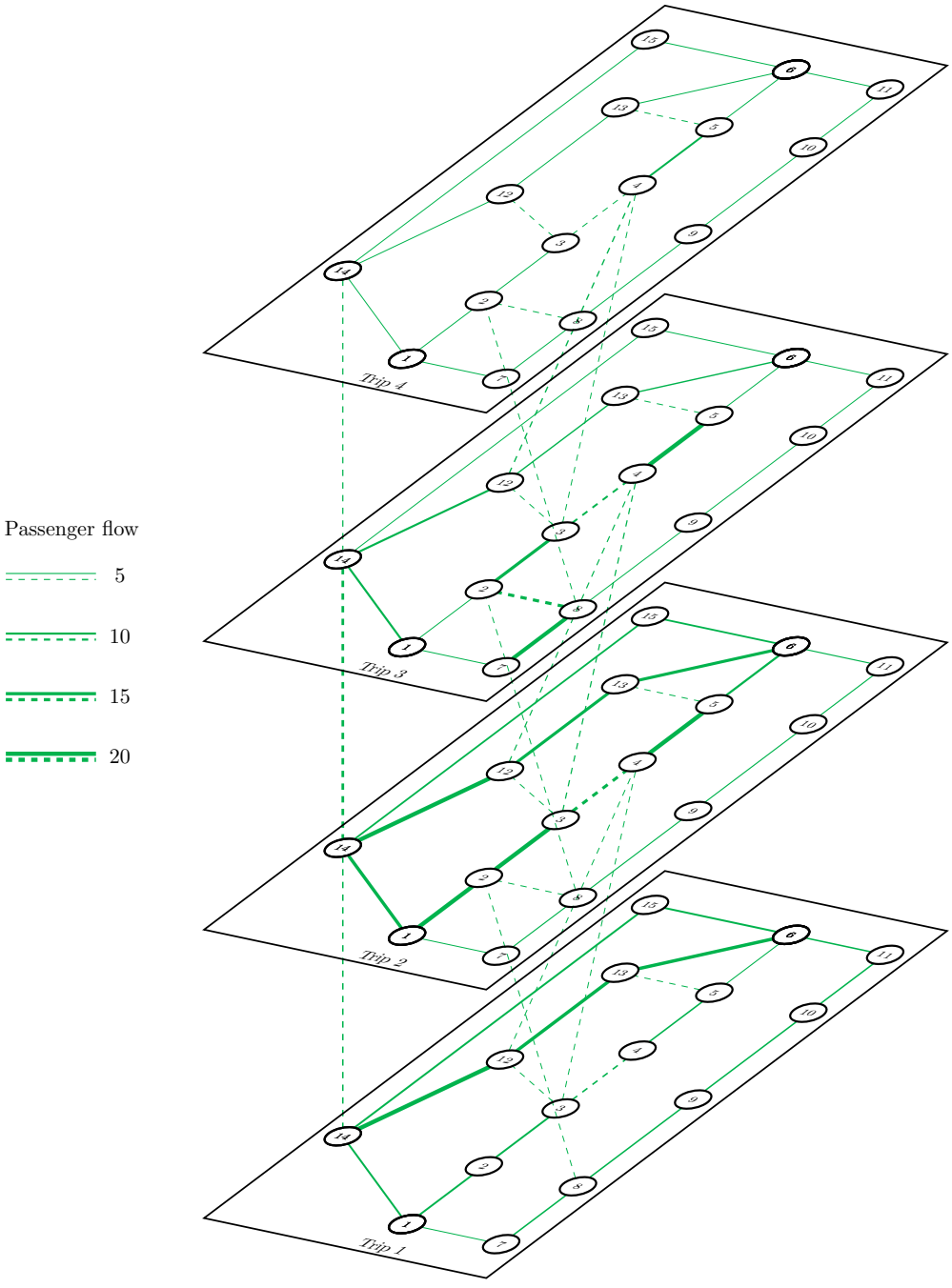


Figure 4.4: Passenger flow on various trips for capacitated transit assignment

between 7-5, passengers prefer taking two transfers 8-2 and 3-4 within second or second to third trips of the respective routes. Some passengers going from 14 to 5 have to face denied boarding on the blue route due to the only option to get to destination 5. This results in non-zero flow on transfer link 12-4 between various trips of orange and blue routes. Finally, we do not observe any passengers that have to walk to their destination due to failed transfer.

Chapter 5

Transit network design with strategic passenger assignment

The planning of a transit system is a complex problem if approached as a single monolithic unit. Therefore, it is common to divide this process into several stages to make it tractable to solve. Ceder and Wilson (1986) divided this process into five stages: Route design, Frequency setting, Timetable preparation, Fleet assignment, and Crew assignment [78]. The problem of designing routes and their frequencies is commonly referred to as the *Transit Network Design Problem* (TNDP), which is addressed in the current chapter. The main purpose of TNDP is to design a public transit network that can transport a given origin-destination demand cost-effectively. The route structure and associated frequency determine the level of service for passengers in terms of travel time, wait time, and the number of transfers, the total operating cost of the service, and determines whether it has sufficient capacity to serve the demand. Therefore, for this problem, one needs to consider several competing objectives from different perspectives. It includes minimizing the cost of operating the transit service, maximizing its geographical coverage, and maximizing the level of service for passengers under a

limited budget.

A critical aspect of the network design problem is to capture the interplay between transit service and passenger behavior through a *transit assignment* model. It would be naïve for a transit agency to design a network without incorporating strategic passenger behavior as it predicts the number of passengers using different routes in the network and estimates the total travel cost and the number of transfers experienced by passengers when traveling between various origin-destination pairs. The transfers are particularly important because they impose extra waiting time and inconvenience to passengers and could be reduced by increasing the number of direct routes at the expense of higher operating costs. Another important aspect is to predict the passenger choices in a network with a limited capacity of transit vehicles, which creates congestion, and passengers may have to face denied boarding to a desired transit route [36]. To incorporate strategic passenger behavior with transfers as well as capacity constraints, Cancela et al. (2015) suggest formulating TNDP as a bi-level optimization problem [3]. This is because the single-level TNDP formulation will force the passenger flow assignment to respect the transfer and capacity constraints, resulting in a prediction of unrealistic behavior.

In this chapter, we formulate TNDP as a *Bi-level Programming Problem* (BLPP). The upper level evaluates the route and frequency design by minimizing the composite operational and passenger cost while considering the limited budget, the number of transfers, and limited road and bus capacities. The lower level optimization problem predicts the effect of network design on passenger behavior through the *optimal strategy transit assignment model* [7]. The overall optimization model is a complex problem to solve due to the presence of non-linearities and non-convexities in the model along with combinatorial explosion arising from the discrete design decision variables. The chapter

develops relaxations, reformulations, and decomposition techniques to solve this problem.

The current chapter is structured as follows. Section 5.1 introduces the notations and concepts to be used in this chapter. Then, section 5.2 formulates the bi-level transit network design problem as a mathematical programming problem, which is followed by its reformulation in Section 5.3. Then, a branch-and-benders cut algorithm to solve the current problem is described in Section 5.4. Finally, the results of numerical experiments are presented Section 5.5.

5.1 Preliminaries

In this section, we introduce notations and concepts to be used in this chapter. Let us begin by considering a road network characterized by a directed graph $G_R(N_R, A_R)$, where N_R and A_R represent the set of nodes and links in the road network respectively. It is assumed that we have already identified a set of terminal[1] locations $\mathfrak{T} \subseteq N_R$. The passenger demand is assumed to be concentrated on a set of centroids of traffic analysis zones denoted by set Z. Let $O \subseteq Z$ and $D \subseteq Z$ be the set of origins and destinations respectively. The demand between any origin-destination pair $(o, d) \in O \times D$ is represented by d^{od}. The set of links coming out and going into a node i are denoted by sets $FS(i)$ and $BS(i)$ respectively. Let $\mathfrak{d} : A_R \mapsto \mathbb{R}$ and $c : A_R \mapsto \mathbb{R}$ be the length (in miles) and travel time (in minutes) of road links.

[1] A terminal is a transit station where buses/trains start or end their trip

5.1.1 Creation of candidate transit lines

The first step in the planning of a transit network is the creation of candidate transit routes L. We use words "route" and "line"[2] interchangeably in this chapter. A candidate transit route $(l \subset A_R) \in L$ is a set of connected arcs whose route length is given by $\Delta_l = \sum_{a \in l} \mathfrak{d}(a)$. We use the criteria described by [72] for creating a set of candidate transit routes, which is discussed below:

1. The route length should not be too long or too short, i.e., $\Delta_{\min} \leq \Delta_l \leq \Delta_{\max}$. This is because longer routes are susceptible to unreliable service, whereas shorter routes may require frequent turnovers at terminals resulting in higher operational costs.

2. The travel time between two nodes of a route should not deviate more than $\tilde{\delta}\%$ of the shortest road travel time between those nodes. This is important to not make transit less attractive to auto mode.

3. A route should not be completely included in another route, i.e., for any two routes $l, l' \in L,\ l \backslash l' \neq \phi$ and $l' \backslash l \neq \phi$.

4. A route should not be circular, i.e., starting node and ending node of the route are not the same.

5. Every origin-destination pair should be connected by not more than $\tilde{\alpha}$ transfers.

The algorithmic procedure to create candidate transit routes is summarized in Algorithm 7. The algorithm uses the depth-first approach to create routes between terminal nodes to all other nodes in the network and filter out routes that do not satisfy the

[2] A transit line/route is defined by a set of stops with a starting and an endpoint between which buses/trains run back and forth.

above criteria 1-3. Criterion 4 is automatically satisfied as all the routes generated by the algorithm are acyclic. The last criterion will be incorporated into the design model.

Algorithm 7 Algorithm for generating candidate transit routes

Input $G_R(N_R, A_R)$
Output L ▷ Set of candidate transit routes
$L \leftarrow []$
procedure CREATECANDIDATEROUTES(G_R)
    for $t \in \mathfrak{T}$ **do**
        for $i \in N_R$ **do**
            $visited \leftarrow \{\}$; $temp\_routes \leftarrow []$; $path \leftarrow []$
            $visited[n] \leftarrow$False, $\forall n \in N$
            ENUMERATEPATHS($t, i, visited, path$)
            Filter out paths in *temp_routes* that do not satisfy criterion 1-3 given above
            Append all the routes in *temp_routes* to L
function ENUMERATEPATHS($i, n, visited, path$)
    $visited[i] \leftarrow$ True
    Append i to *path*
    if $i == n$ **then**
        Append *path* to *temp_routes*
    else
        for $k \in FS(i)$ **do**
            ENUMERATEPATHS($k, n, visited, path$)
    Remove last node from *path*
    $visited[i] \leftarrow$False

5.1.2 Frequency-based transit network

The set of candidate routes L induces a transit network characterized by a digraph $G(N, A)$, where N and A denote the set of nodes and links in the transit network respectively. For this study, every transit route is assumed to be bi-directional. However, this can be easily relaxed for streets where traffic runs in a single direction. The set of nodes can be partitioned into three types of nodes, namely, access/egress nodes Z, in-vehicle nodes N_v, and transfer nodes $N_t \subseteq N_v$. The in-vehicle nodes are specific for

each route, and transfer nodes are the in-vehicle nodes that can be used to transfer to other in-vehicle nodes. The set of links can be partitioned into four types of links, namely, access links A_e, egress links A_{ee}, in-vehicle links A_v, and transfer links A_t. A transfer link is created between two in-vehicle nodes only if the distance between them is less than an acceptable walking distance (say 0.75mi), the route associated with both nodes are different, and its head node and tail node are not the first stop and last stop of their associated routes respectively. Let $S = Z \cup N_t$ be the set of waiting nodes. The waiting nodes are the nodes in the network whose some or all of the outgoing links incur waiting. Let $f : A_e \cup A_t \mapsto \mathbb{R}$ be frequency associated with the access and transfer links. Furthermore, we denote $c : A \mapsto \mathbb{R}$ as the cost of traversing links.

5.1.3 Optimal strategy frequency-based transit assignment

Spiess and Florian (1987) proposed an optimization program to model the behavior of passengers traveling in a frequency-based transit system. The model is particularly attractive not only because it predicts the strategic behavior of passengers, but also it has a linear programming formulation. They assumed that the link travel time in the network is constant, there is no denied boarding due to limited route capacity, and headways of various transit routes follow an exponential distribution. To present their model, let us define a variable $\{g_{ik}\}_{i\in N,k\in D}$ as below:

$$g_{ik} = \begin{cases} d^{ik}, & \text{if } i \neq k, (i,k) \in O \times D \\ -\sum_{o\in O} d^{ok}, & \text{if } i = k \\ 0, & \text{otherwise} \end{cases} \tag{5.1}$$

Furthermore, let us denote v_{ak} and W_{ik} as the flow of passengers on link $a \in A$ and total passenger waiting time at node $i \in S$ destined to $k \in D$ respectively. Then, the

assignment can be formulated as (5.2).

$$z^{FP} = \underset{\mathbf{v},\mathbf{W}}{\text{minimize}} \quad \sum_{k\in D}\left(\sum_{a\in A} c_a v_{ak} + \sum_{i\in S} W_{ik}\right) \tag{5.2a}$$

$$\text{subject to} \quad \sum_{a\in FS(i)} v_{ak} = \sum_{a\in BS(i)} v_{ak} + g_{ik}, \forall i \in N, \forall k \in D \tag{5.2b}$$

$$v_{ak} \leq f_a W_{ik}, \forall a \in FS(i), \forall i \in S, \forall k \in D \tag{5.2c}$$

$$v_{ak} \geq 0, \forall a \in A, \forall k \in D \tag{5.2d}$$

The assignment program (5.2) minimizes the total expected link costs and wait time at various transit stops experienced by passengers in a transit network subject to the flow conservation constraint at each node (5.2b), flow proportion constraints (5.2c), and the non-negativity constraints (5.2d).

5.2 Bi-level transit network design problem

In this section, we introduce and formulate the *bi-level transit network design problem* (BTNDP). The problem can be viewed as a static version of a non-cooperative two-person sequential game introduced by von Stackelberg in the context of unbalanced economic markets [139]. In the present context, the two players are the transit agency, which is responsible for the design of the network, is termed as the "leader" and the passengers, who are flowing in the given network, are termed as the "follower". The control or decision variables are partitioned amongst the players who seek to minimize their cost functions. This is a static game because both players have one move to make, i.e., the transit agency will make a decision about the network once-and-for-all, and the passengers will move themselves according to the optimal strategies. Now assuming common knowledge of rationality, what should a leader do? It would be naïve for the leader to minimize its cost based on some belief about the decisions of the follower

because it knows exactly how rational follower would respond to its design decisions. Therefore, the "leader" goes first and attempts to minimize the net operational cost and travel cost of the passengers subject to some feasibility constraints and anticipation of the response of its opponent. The follower then observes the leader's decision and reacts using optimal strategies in the network. By doing so, the leader may gain *first mover advantage*. Because the set of feasible choices available to either player is interdependent, the leader's decision affects both the players' cost and allowable actions and vice-versa. Furthermore, the leader can affect the decisions of the passengers but cannot fully control their actions. In what follows, we present the network design model. Before presenting it, we must make the following assumptions:

Assumption 1. *(Modeling assumptions)*

1. *The link costs are constant.*
2. *Leader and follower are expected cost minimizers.*
3. *The passenger demand between various origin-destination pairs is fixed.*
4. *Buses stop at each stop of their route. The boarding and alighting time of passengers are ignored.*
5. *The available number of buses is enough to serve all the passengers in the demand matrix.*
6. *The leader is assumed to be optimistic, i.e., when there are multiple optimal decisions of the follower for a given upper-level decision, it chooses the one that benefits it the most among all optimal solutions.*

5.2.1 Decisions

In BTNDP, the leader or the transit agency has control over two sets of decision variables, namely, whether a candidate route should be built/located or not ($\mathbf{x} \in \mathfrak{B}^{|L|}$) and which frequency $f \in \Theta$ should be adopted for a located transit route ($\mathbf{y} \in \mathfrak{B}^{|L \times \Theta|}$), where $\Theta = \{1, 2, 3, 6, 12\}$ is the set of possible frequencies in buses/hr. By defining frequency set in this way, one can define the lower and upper bound on the frequency values. The follower or passengers have control over the flow on links $\mathbf{v} \in \mathfrak{R}_{+}^{|A \times D|}$ and the total wait time at nodes ($\mathbf{W} \in \mathfrak{R}^{|S \times D|}$) which is the outcome of the optimal strategies of passengers moving in the network. Here, as a result of sequential decision making, the functions $\mathbf{v}(\mathbf{x}, \mathbf{y}), \mathbf{W}(\mathbf{x}, \mathbf{y})$ emphasize the fact that leader's decision is implicit in $\mathbf{v}$ and $\mathbf{W}$. For convenience, we will not show this dependency in the notations.

5.2.2 Upper level problem

The upper-level problem is the optimization problem solved by the transit agency for the design of the transit network. To do so, it must consider different perspectives, namely, operator and passenger perspectives. The operator's perspective is to provide transit service at a minimal operational cost, and the passenger's perspective is to get cheap and convenient service [72]. Therefore, we consider a composite objective function for the upper-level problem based on both perspectives. It is the sum of the operating cost of running buses and the total expected travel time and wait time experienced by the passengers.

$$F(\mathbf{x}, \mathbf{y}, \mathbf{v}, \mathbf{W}) = (1-\beta) \left[\sum_{l \in L} \sum_{f \in \Theta} (2 f \Delta_l \phi) y_{lf} \right] + \beta \left[\sum_{k \in D} \left(\sum_{a \in A} c_a v_{ak} + \sum_{i \in S} W_{ik} \right) \varphi \right] \quad (5.3)$$

where ϕ is the bus operating cost per mile, φ is the value of travel time in \$/hr, and $0 \leq \beta \leq 1$ is the weight/importance given passengers' perspective. The operating cost

is obtained by multiplying the total running distance (round trip travel distance times frequency) of a route with operating cost per mile, and passenger cost is obtained by adding the total expected travel time and wait time spent in the network. The upper-level problem is subject to the following constraints:

Fleet constraint

A transit agency has a limited budget to operate a service. We use the number of available buses B as a proxy for the budget.

$$\sum_{l \in L} \left(\sum_{f \in \Theta} \mathcal{B}(l, f) y_{lf} \right) \leq B \tag{5.4}$$

where, the mapping $\mathcal{B} : L \times \Theta \mapsto \mathfrak{N}$ used in (5.4) is defined as $\mathcal{B}(l, f) = \left(f \times \sum_{a \in l} 2c_a \right)$, which describes that the number of buses required to provide frequency $f \in \Theta$ for line $l \in L$ is the product of the frequency and round trip travel time. One can also include budget constraints associated with the construction of new lines if new train lines or BRT are constructed. By assuming a constant construction cost per mile, the constraint will take the form of a Knapsack constraint in terms of $\mathbf{x}$ similar to (5.4). For the sake of simplicity, we do not include it here.

Street capacity

There is a maximum number of buses that can pass through road link $a \in A_R$ per unit time known as street capacity κ_a. The following constraint restricts the total frequency of the buses on link a to be less than or equal to κ_a.

$$\sum_{l \in L : a \in l} \sum_{f \in \Theta} f y_{lf} \leq \kappa_a, \forall a \in A_R \tag{5.5}$$

Logical constraints

We need to include a logical constraint that a frequency for a route cannot be adopted if it is not located. Moreover, if a route is located, then we can only adopt one frequency value for that.

$$\sum_{f \in \Theta} y_{lf} = x_l, \forall l \in L \tag{5.6}$$

Transfer constraints

Transfers are inconvenient and induce extra waiting time for passengers that may make transit service unattractive to them. They can be modeled as a soft approach that restricts or minimizes the number of transfers taken by all passengers. For example, [79] specifies the minimum percentage of total demand that should be satisfied without transfers, [1] minimizes the total number of passengers on transfer links, and [3] specifies the percentage of demand that should be satisfied with no more than a specific number of transfers. We also restrict the average number of transfers per passenger to be less than or equal to $\tilde{\epsilon}$.

$$\frac{\sum_{k \in D} \sum_{a \in A^t} v_{ak}}{\sum_{od \in O \times D} d^{od}} \leq \tilde{\epsilon} \tag{5.7}$$

Minimizing or restricting the number of transferring passengers may not avoid the situation for some passengers in the network taking an unbounded number of transfers. Therefore, we call this approach the "soft approach." One can also formulate a constraint specifying the availability of a path with at most $\tilde{\alpha}$ transfers between every origin-destination pair. For this purpose, let us denote $R(od)$ and $TR(od)$ as the number of direct paths and transfer paths with at most $\tilde{\alpha}$ transfers between origin-destination pair $od \in O \times D$ in the network. Further, let $\mathcal{L}(p)$ be the set of lines used to serve path p in the network. To avoid the enormous number of paths, we take the value of $\tilde{\alpha} = 1$.

Then, the transfer constraints can be formulated as below:

$$\sum_{p \in R(od)} \sum_{l \in \mathcal{L}(p)} x_l + \sum_{tr \in TR(od)} \left(\prod_{l \in \mathcal{L}(tr)} x_l \right) \geq 1, \forall od \in O \times D \tag{5.8}$$

The above constraints make sure that for every origin-destination pair, at least one direct or transfer path (with at most one transfer) is available. We call this approach the "hard approach" of restricting transfers. However, the expression (5.8) has bilinear terms that remain non-convex even if the integrality constraints are relaxed. Fortunately, in this case, bilinear terms consist of the product of two binary variables that can be exactly relaxed using the McCormick relaxation. Let $z_{tr} = \prod_{l \in \mathcal{L}(tr)} x_l, \forall tr \in TR(od), \forall od \in O \times D$. Then, z_{tr} can be expressed as the following set of linear inequalities:

$$z_{tr} \leq x_l, \forall l \in \mathcal{L}(tr), \forall tr \in TR(od) \forall od \in O \times D \tag{5.9a}$$

$$z_{tr} \geq \sum_{l \in \mathcal{L}(tr)} x_l - \left(|\mathcal{L}(tr)| - 1 \right), \forall tr \in TR(od), \forall od \in O \times D \tag{5.9b}$$

The proof of the above reformulation is trivial and can be shown by considering two cases, namely, $x_l = 0$ and 1. For $\tilde{\alpha} > 1$, the expression (5.8) will have multilinear terms that can still be exactly relaxed as the set of linear constraints using the above technique at an expense of a large number of constraints. There are two disadvantages of using the hard constraints. First, if we only include the hard constraints, then it may result in a network where a high proportion of passengers have to take transfer. Second, since the constraints are included for every transfer path and every origin-destination pair, it can result in slow computational performance. We show the effect of both soft and hard approaches on the number of transferring passengers in the numerical experiments (Section 5.5).

We believe that the transfer constraints (5.7) should not be part of the lower level problem because by including them in the lower level, the assignment of passengers will be forced to satisfy transfer constraints, which may show unrealistic passenger behavior in some cases. For example, some passengers prefer taking a transfer path because it provides a lesser expected cost in comparison to a direct path.

Capacity constraints

The modeling of passenger behavior while incorporating the capacity constraints (congestion) is a difficult problem. The congestion is important to consider since it causes denied boarding, which leads to increased waiting time, travel time, and discomfort. Several authors have tried to include congestion into frequency-based transit assignment models through various approaches, namely, discomfort function [7], effective frequency [36,38,140,141], and failure-to-board probabilities [39]. Despite the effort, there is no tractable closed-form of congested frequency-based transit assignment model. On the other hand, it would not be ideal to include transit vehicle capacity constraints into the lower-level assignment program (e.g., in [1]) because doing so may lead to unrealistic passenger behavior, which previous studies on congested frequency-based transit assignments were trying to avoid. Therefore, indirectly, we avoid requiring a congested frequency-based transit assignment model at the lower level by designing a network that does not exceed capacity limits using capacity constraints in the upper level. Assuming that transit vehicles with equal capacities are deployed, we can formulate the capacity constraints as (5.10).

$$\sum_{k \in D} v_{ak} \leq \sum_{f \in \Theta} f y_{l(a)f} \mathfrak{C}, \forall a \in A_v \tag{5.10}$$

where, $\mathfrak{C}$ represent the capacity of one transit vehicle.

5.2.3 Lower level problem

The lower level problem is the optimization problem solved by the passengers. This is the same as the optimal strategy transit assignment model described in Section 5.1.3.

$$\mathcal{Z}^{FP} = \underset{\mathbf{v},\mathbf{W}}{\text{minimize}} \quad f(\mathbf{v},\mathbf{W}) = \sum_{k\in D}\left(\sum_{a\in A} c_a v_{ak} + \sum_{i\in S} W_{ik}\right) \tag{5.11a}$$

$$\text{subject to} \quad \sum_{a\in FS(i)} v_{ak} = \sum_{a\in BS(i)} v_{ak} + g_{ik}, \forall i \in N, \forall k \in D \tag{5.11b}$$

$$v_{ak} \leq \left(\sum_{f\in\Theta} f y_{l(a)f}\right) W_{ik}, \forall a \in FS(i), \forall i \in S, \forall k \in D \tag{5.11c}$$

$$v_{ak} \leq \tilde{M} x_{l(a)}, \forall a \in A_v, \forall k \in D \tag{5.11d}$$

$$v_{ak} \geq 0, \forall a \in A, \forall k \in D \tag{5.11e}$$

The constraints (5.11c) incorporate the condition that a route frequency can be used if only if it is adopted for that route and (5.11d) forces the flow on inactive in-vehicle links to be equal to zero. The value of big-M in (5.11d) can be taken as $\tilde{M} = \sum_{o\in O} d^{od}$. Although there is no need to include (5.11d) as they are redundant due to the presence of (5.10) in the upper level, they provide better linear relaxation bounds. Therefore, they are also known as *strong linking constraints* [142]. The above formulation introduces bilinear terms in (5.11c). By again employing McCormick relaxations, this can be exactly relaxed into a set of linear constraints. Let $t_{faik} = y_{l(a)f} W_{ik}, \forall f \in \Theta, \forall a \in FS(i), \forall i \in S, \forall k \in D$. Further, let us assume that there exists a finite upper and lower bound to the variable W_{ik} , i.e., $\underline{W}_{ik} \leq W_{ik} \leq \overline{W}_{ik}$. Then, t_{faik} can be expressed as

the set of linear constraints (5.12a)-(5.12d).

$$\overline{W}_{ik} - W_{ik} + t_{faik} - \overline{W}_{ik} y_{l(a)f} \geq 0 \tag{5.12a}$$

$$\overline{W}_{ik} y_{l(a)f} - t_{faik} \geq 0 \tag{5.12b}$$

$$t_{faik} - \underline{W}_{ik} y_{l(a)f} \geq 0 \tag{5.12c}$$

$$W_{ik} - \underline{W}_{ik} - t_{faik} + \underline{W}_{ik} y_{l(a)f} \geq 0 \tag{5.12d}$$

We can assume $\underline{W}_{ik} = 0$ since the wait time cannot be negative and $\overline{W}_{ik} = \left(\sum_{o \in O} d^{od}\right) \times \frac{60}{1}$ assuming all the passengers destined to k flow through node i with total expected frequency of only 1 bus/hr. Moreover, we can ignore constraints (5.12a) because it provides an upper bound on W_{ik} and since we are minimizing W_{ik}, it will either be redundant (when $y_{l(a)f} = 1$) or not tight (when $y_{l(a)f} = 0$). After this reformulation, (5.11) can be stated as (5.13).

.

$$\mathcal{Z}^{FP} = \underset{\mathbf{v},\mathbf{W},\mathbf{t}}{\text{minimize}} \quad f(\mathbf{v},\mathbf{W},\mathbf{t}) = \sum_{k \in D} \left(\sum_{a \in A} c_a v_{ak} + \sum_{i \in S} W_{ik} \right) \tag{5.13a}$$

$$\text{subject to} \quad \sum_{a \in FS(i)} v_{ak} = \sum_{a \in BS(i)} v_{ak} + g_{ik}, \forall i \in N, \forall k \in D \tag{5.13b}$$

$$v_{ak} \leq \sum_{f \in \Theta} f t_{faik}, \forall a \in FS(i), \forall i \in S, \forall k \in D \tag{5.13c}$$

$$t_{faik} \leq \overline{W}_{ik} y_{l(a)f}, \forall f \in \Theta, \forall a \in FS(i), \forall i \in S, \forall k \in D \tag{5.13d}$$

$$W_{ik} - t_{faik} \geq 0, \forall f \in \Theta, \forall a \in FS(i), \forall i \in S, \forall k \in D \tag{5.13e}$$

$$v_{ak} \leq \left(\sum_{o \in O} d^{od} \right) x_{l(a)}, \forall a \in A_v, \forall k \in D \tag{5.13f}$$

$$v_{ak} \geq 0, \forall a \in A, \forall k \in D \tag{5.13g}$$

$$t_{faik} \geq 0, \forall f \in \Theta, \forall a \in FS(i), \forall i \in S, \forall k \in D \tag{5.13h}$$

By putting down all the pieces together, we can state the overall BTNDP as follows:

$$\mathcal{Z}^* = \underset{\mathbf{x},\mathbf{y},\mathbf{z},\mathbf{v}}{\text{minimize}} \quad (1-\beta)\left[\sum_{l\in L}\sum_{f\in\Theta}(2f\Delta_l\phi)y_{lf}\right] + \beta\left[\sum_{k\in D}\left(\sum_{a\in A}c_a v_{ak} + \sum_{i\in S}W_{ik}\right)\varphi\right] \tag{5.14a}$$

$$\text{subject to} \quad \sum_{l\in L}\left(\sum_{f\in\Theta}\mathcal{B}(l,f)y_{lf}\right) \leq B \tag{5.14b}$$

$$\sum_{l\in L:a\in l}\sum_{f\in\Theta} f y_{lf} \leq \kappa_a, \forall a \in A_R \tag{5.14c}$$

$$\sum_{f\in\Theta} y_{lf} = x_l, \forall l \in L \tag{5.14d}$$

$$\sum_{p\in R(od)}\sum_{l\in\mathcal{L}(p)} x_l + \sum_{tr\in TR(od)} z_{tr} \geq 1, \forall od \in O\times D \tag{5.14e}$$

$$z_{tr} \leq x_l, \forall l \in \mathcal{L}(tr), \forall tr \in TR(od) \forall od \in O\times D \tag{5.14f}$$

$$z_{tr} \geq \sum_{l\in\mathcal{L}(tr)} x_l - \left(|\mathcal{L}(tr)| - 1\right), \forall tr \in TR(od), \forall od \in O\times D \tag{5.14g}$$

$$-\sum_{k\in D}\sum_{a\in A^t} v_{ak} \geq -\tilde{\epsilon}\left(\sum_{od\in O\times D} d^{od}\right) \tag{5.14h}$$

$$-\sum_{k\in D} v_{ak} \geq -\sum_{f\in\Theta} f y_{l(a)f}\mathfrak{C}, \forall a \in A_v \tag{5.14i}$$

$$\mathbf{x} \in \mathfrak{B}^{|L|}, \mathbf{y} \in \mathfrak{B}^{|\Theta\times L|} \tag{5.14j}$$

$$(\mathbf{v},\mathbf{W},\mathbf{t}) \in \text{argmin} \quad \sum_{k\in D}\left(\sum_{a\in A} c_a v_{ak} + \sum_{i\in S} W_{ik}\right) \tag{5.14k}$$

$$\text{subject to} \sum_{a\in FS(i)} v_{ak} = \sum_{a\in BS(i)} v_{ak} + g_{ik}, \forall i \in N, \forall k \in D \tag{5.14l}$$

$$-v_{ak} \geq -\sum_{f\in\Theta} f t_{faik}, \forall a \in FS(i), \forall i \in S, \forall k \in D \tag{5.14m}$$

$$-t_{faik} \geq -\overline{W}_{ik} y_{l(a)f}, \forall f \in \Theta, \forall a \in FS(i), \forall i \in S, \forall k \in D \tag{5.14n}$$

$$W_{ik} - t_{faik} \geq 0, \forall f \in \Theta, \forall a \in FS(i), \forall i \in S, \forall k \in D \tag{5.14o}$$

$$-v_{ak} \geq -\left(\sum_{o\in O} d^{od}\right) x_{l(a)}, \forall a \in A_v, \forall k \in D \tag{5.14p}$$

$$t_{faik} \geq 0, \forall f \in \Theta, \forall a \in FS(i), \forall i \in S, \forall k \in D \tag{5.14q}$$

$$\mathbf{v} \in \mathfrak{R}_+^{|A\times D|} \tag{5.14r}$$

5.3 BTNDP reformulation

In this section, we describe a reformulation of BTNDP that can be solved using an off-the-shelf solver. Before moving forward, it is convenient to define a few sets to help develop the solution algorithm. They are defined as follows:

1. Constraint region of $(\mathbf{x}, \mathbf{y})$ that does not depend on the lower level decision variables.

$$\mathcal{U} = \underset{\mathbf{x},\mathbf{y}}{\text{proj}}\{(\mathbf{x}, \mathbf{y}, \mathbf{z}) : (5.14\text{b}) - (5.14\text{g}), (5.14\text{j})\} \tag{5.15}$$

   where, $\underset{\mathbf{x},\mathbf{y}}{\text{proj}}(U)$ represents the projection of the set U onto x and y variables.

2. Constraint region of the bi-level problem (5.14)

$$\mathcal{S} = \{(\mathbf{x}, \mathbf{y}, \mathbf{z}, \mathbf{v}, \mathbf{W}, \mathbf{t}) : (5.14\text{b}) - (5.14\text{j}), (5.14\text{l}) - (5.14\text{r})\} \tag{5.16}$$

3. Feasible set of follower decisions for each fixed $(\mathbf{x}, \mathbf{y}) \in \mathcal{U}$

$$\mathcal{S}(\mathbf{x}, \mathbf{y}) = \{(\mathbf{v}, \mathbf{W}, \mathbf{t}) : (5.14\text{l}) - (5.14\text{r})\} \tag{5.17}$$

4. Follower's reaction set for each fixed $(\mathbf{x}, \mathbf{y}) \in \mathcal{U}$

$$\mathcal{P}(\mathbf{x}, \mathbf{y}) = \left\{ (\mathbf{v}, \mathbf{W}, \mathbf{t}) \in \text{argmin} \left[\sum_{k \in D} \left(\sum_{a \in A} c_a v_{ak} + \sum_{i \in S} W_{ik} \right) \right] : (\mathbf{v}, \mathbf{W}, \mathbf{t}) \in \mathcal{S}(\mathbf{x}, \mathbf{y}) \right\} \tag{5.18}$$

5. Inducible region

$$IR = \{(\mathbf{x}, \mathbf{y}, \mathbf{z}, \mathbf{v}, \mathbf{W}, \mathbf{t}) \in \mathcal{S} : (\mathbf{v}, \mathbf{W}, \mathbf{t}) \in \mathcal{P}(\mathbf{x}, \mathbf{y})\} \tag{5.19}$$

IR is the region over which the leader can optimize. Therefore, BTNDP can be compactly written as below:

$$\underset{\mathbf{x},\mathbf{y},\mathbf{z},\mathbf{v},\mathbf{W},\mathbf{t}}{\text{minimize}} \quad F(\mathbf{x},\mathbf{y},\mathbf{z},\mathbf{v},\mathbf{W},\mathbf{t}) \tag{5.20a}$$

$$\text{subject to} \quad (\mathbf{x},\mathbf{y},\mathbf{z},\mathbf{v},\mathbf{W},\mathbf{t}) \in IR \tag{5.20b}$$

An important problem related to any bi-level problem, known as the *High Point Problem* (HPP), is the problem without the lower-level objective function constraint. In our context, HPP is minimizing objective function (5.14a) over $\mathcal{S}$. For the existence of the solution to BTNDP, we make the following assumption.

Assumption 2. *HPP is feasible and bounded.*

Assumption (2) is a necessary requirement for the existence of an optimal solution to BTNDP (5.14). Moreover, since HPP is a relaxation of (5.14), it guarantees a finite lower bound to the problem (5.14).

Assumption 3. *The dual of the follower's problem for a fixed* $(\mathbf{x},\mathbf{y}) \in \mathcal{U}$ *is feasible.*

As the feasible region of the follower's dual problem does not depend on $(\mathbf{x},\mathbf{y}) \in \mathcal{U}$ (see (5.21)), its feasibility ensures that the follower's primal problem (5.13) is bounded from below. Furthermore, this assumption promises a strong duality of the follower's problem for a fixed $(\mathbf{x},\mathbf{y}) \in \mathcal{U}$.

A feasible upper level design solution $(\hat{\mathbf{x}},\hat{\mathbf{y}},\hat{\mathbf{z}})$ to HPP may or may not be feasible to BTNDP (5.14). When it is feasible to (5.14), then we call that solution as *bi-level feasible design* solution. To be precise, any $(\hat{\mathbf{x}},\hat{\mathbf{y}},\hat{\mathbf{z}})$ is a bi-level feasible design solution if $\mathcal{S}(\hat{\mathbf{x}},\hat{\mathbf{y}}) \neq \phi$ and $\exists(\hat{\mathbf{v}},\hat{\mathbf{W}},\hat{\mathbf{t}}) \in \mathcal{P}(\hat{\mathbf{x}},\hat{\mathbf{y}})$ such that (5.14h) and (5.14i) are satisfied. Any bi-level feasible design solution $(\hat{\mathbf{x}},\hat{\mathbf{y}},\hat{\mathbf{z}})$ along with the follower reaction

$(\hat{\mathbf{v}}, \hat{\mathbf{W}}, \hat{\mathbf{t}}) \in \mathcal{P}(\hat{\mathbf{x}}, \hat{\mathbf{y}})$ will provide a finite upper bound to the BTNDP (5.14). Next, we prove that the follower problem is always feasible for a given $(\mathbf{x}, \mathbf{y}) \in \mathcal{U}$.

Proposition 5. $\mathcal{S}(\mathbf{x}, \mathbf{y}) \neq \phi, \forall (\mathbf{x}, \mathbf{y}) \in \mathcal{U}$

Proof. The set $\mathcal{S}(\mathbf{x}, \mathbf{y})$ can be empty in two cases: either there is no flow balance, i.e., $\sum_{k \in D} \sum_{i \in N} g_{ik} \neq 0$ or there does not exists a directed path between some O-D pair $(o, k) \in O \times D$. However, it is not possible to have any of these cases because from the definition of (5.1), we have $\sum_{k \in D} \sum_{i \in N} g_{ik} = 0$ and there always exists at least one directed path between every O-D pair due to transfer constraints (5.14e)-(5.14g). □

Associating dual variables μ_{ik}, λ^0_{aik}, $\lambda^1_{faik}, \lambda^2_{faik}$, and τ_{ak} to constraints (5.14l), (5.14m), (5.14n), (5.14o), and (5.14p) respectively, the dual of the follower problem can be written as below:

$$\mathcal{Z}^{DFP}(\mathbf{x}, \mathbf{y}) = \underset{\mu, \boldsymbol{\lambda^0}, \boldsymbol{\lambda^1}, \boldsymbol{\lambda^2}, \tau}{\text{maximize}} \quad \sum_{k \in D} \left[\sum_{i \in N} \mu_{ik} g_{ik} - \sum_{i \in S} \sum_{a \in FS(i)} \sum_{f \in \Theta} \left\{ \overline{W}_{ik} y_{l(a)f} \lambda^1_{faik} \right\} - \sum_{a \in A_v} \tau_{ak} \left(\sum_{o \in O} d^{ok} \right) x_{l(a)} \right] \tag{5.21a}$$

$$\text{subject to} \quad -\mu_{ik} + \mu_{jk} + \lambda^0_{aik} \geq -c_a, \forall a = (i,j) \in A_a \cup A_t, \forall i \in N, \forall k \in D \tag{5.21b}$$

$$-\mu_{ik} + \mu_{jk} + \tau_{ak} \geq -c_a, \forall a = (i,j) \in A_v, \forall i \in N, \forall k \in D \tag{5.21c}$$

$$-\mu_{ik} + \mu_{jk} \geq -c_a, \forall a = (i,j) \in A_e, \forall i \in N, \forall k \in D \tag{5.21d}$$

$$\sum_{a \in FS(i)} \sum_{f \in \Theta} \lambda^2_{faik} = 1, \forall i \in S, \forall k \in D \tag{5.21e}$$

$$-f\lambda^0_{aik} + \lambda^1_{faik} + \lambda^2_{faik} \geq 0, \forall f \in \Theta, \forall a \in FS(i), \forall i \in S, \forall k \in D \tag{5.21f}$$

$$\lambda^0_{aik} \geq 0, \forall a \in FS(i), \forall i \in S, \forall k \in D \tag{5.21g}$$

$$\lambda^1_{faik} \geq 0, \lambda^2_{faik} \geq 0, \forall f \in \Theta, \forall a \in FS(i), \forall i \in S, \forall k \in D \tag{5.21h}$$

$$\tau_{ak} \geq 0, \forall a \in A_v, \forall k \in D \tag{5.21i}$$

5.3.1 Single-level reformulation of BTNDP

In this section, we reformulate (5.14) as a single-level problem using KKT conditions of the follower problem, which is presented as below:

$$\mathcal{Z}^* = \underset{\substack{\mathbf{x},\mathbf{y},\mathbf{z},\mathbf{v},\mathbf{W},\mathbf{t},\\ \mu,\lambda^{\mathbf{0}},\lambda^{\mathbf{1}},\lambda^{\mathbf{2}},\tau}}{\text{minimize}} \quad (1-\beta)\left[\sum_{l\in L}\sum_{f\in\Theta}(2f\Delta_l\phi)y_{lf}\right] + \beta\left[\sum_{k\in D}\left(\sum_{a\in A}c_a v_{ak} + \sum_{i\in S}W_{ik}\right)\varphi\right] \tag{5.22a}$$

$$\text{subject to} \quad (5.14\text{b}) - (5.14\text{j}) \tag{5.22b}$$

$$(5.14\text{l}) - (5.14\text{r}) \tag{5.22c}$$

$$\sum_{k\in D}\left(\sum_{a\in A}c_a v_{ak} + \sum_{i\in S}W_{ik}\right) \le \sum_{k\in D}\left[\sum_{i\in N}\mu_{ik}g_{ik} - \sum_{i\in S}\sum_{a\in FS(i)}\sum_{f\in\Theta}\left\{\overline{W}_{ik}y_{l(a)f}\lambda^1_{faik}\right\} - \sum_{a\in A_v}\tau_{ak}\left(\sum_{o\in O}d^{ok}\right)x_{l(a)}\right] \tag{5.22d}$$

$$(5.21\text{b}) - (5.21\text{i}) \tag{5.22e}$$

In this reformulation, (5.22c) and (5.22e) represent the primal and dual feasibility conditions of the follower problem respectively. (5.22d) ensures the strong duality of the follower problem while optimizing leader's objective (5.22a) [143]. The feasible region of this problem (5.22b)-(5.22e) is also an explicit representation of IR. An important observation about (5.22) is that it is feasible only when (5.22d) is satisfied at equality. The following proposition shows that result.

Proposition 6. *For a fixed* $(\hat{\mathbf{x}},\hat{\mathbf{y}}) \in \mathcal{U}$*, the following inequality holds for* (5.22)*.*

$$\sum_{k\in D}\left(\sum_{a\in A}c_a v_{ak} + \sum_{i\in S}W_{ik}\right) \ge \sum_{k\in D}\left[\sum_{i\in N}\mu_{ik}g_{ik} - \sum_{i\in S}\sum_{a\in FS(i)}\sum_{f\in\Theta}\left\{\overline{W}_{ik}y_{l(a)f}\lambda^1_{faik}\right\} - \sum_{a\in A_v}\tau_{ak}\left(\sum_{o\in O}d^{ok}\right)x_{l(a)}\right] \tag{5.23}$$

Proof. Let $S_L(\hat{\mathbf{x}},\hat{\mathbf{y}}) = \{(\mathbf{v},\mathbf{W},\mathbf{t}) : (5.14\text{h}) - (5.14\text{i}), (5.14\text{l}) - (5.14\text{r})\}$, $S_F(\hat{\mathbf{x}},\hat{\mathbf{y}}) = \{(\mathbf{v},\mathbf{W},\mathbf{t}) : (5.14\text{l}) - (5.14\text{r})\}$, and $\mathcal{Z}^1(\hat{\mathbf{x}},\hat{\mathbf{y}}) = \max\left\{\sum_{k\in D}\left(\sum_{a\in A}c_a v_{ak} + \sum_{i\in S}W_{ik}\right) :\right.$

$(\mathbf{v}, \mathbf{W}, \mathbf{t}) \in S_L(\hat{\mathbf{x}}, \hat{\mathbf{y}})\}$. As the follower problem is feasible for any fixed $(\hat{\mathbf{x}}, \hat{\mathbf{y}}) \in \mathcal{U}$ and strong duality holds, we have, $\mathcal{Z}^{DFP}(\hat{\mathbf{x}}, \hat{\mathbf{y}}) = \mathcal{Z}^{FP}(\hat{\mathbf{x}}, \hat{\mathbf{y}}) =$
$\max\left\{\sum_{k\in D}\left(\sum_{a\in A} c_a v_{ak} + \sum_{i\in S} W_{ik}\right) : (\mathbf{v}, \mathbf{W}, \mathbf{t}) \in S_F(\hat{\mathbf{x}}, \hat{\mathbf{y}})\right\}$. Since $S_L(\hat{\mathbf{x}}, \hat{\mathbf{y}}) \subseteq IR$ and $S_F(\hat{\mathbf{x}}, \hat{\mathbf{y}}) \subseteq IR$, we have $\mathcal{Z}^1 \leq \mathcal{Z}^*$ and $\mathcal{Z}^{DFP} \leq \mathcal{Z}^*$. Further, $S_F \subseteq S_L$, we must have $\mathcal{Z}^{DFP} \leq \mathcal{Z}^1 \leq \mathcal{Z}^*$. This shows the required result. □

The exact single-level reformulation of BTNDP (5.22) is a mixed-integer non-linear program. The non-linearity arises due to the bi-linear terms in (5.22d), which we further relax as the set of linear constraints by again employing McCormick relaxations. Let $\xi^1_{faik} = y_{l(a)f}\lambda^1_{faik}$ and $\xi^2_{ak} = \tau_{ak}x_{l(a)}$ and assuming an upper bound on λ^1_{faik} and τ_{ak} as $\overline{\lambda}^1_{faik}$ and $\overline{\tau}_{ak}$ resp., we can write (5.22) after this reformulation as (5.24).

$$\mathcal{Z}^* = \underset{\substack{\mathbf{x},\mathbf{y},\mathbf{z},\mathbf{v},\mathbf{W},\mathbf{t},\\ \mu,\lambda^{\mathbf{0}},\lambda^{\mathbf{1}},\lambda^{\mathbf{2}},\tau,\xi^{\mathbf{1}},\xi^{\mathbf{2}}}}{\text{minimize}} \quad (1-\beta)\left[\sum_{l\in L}\sum_{f\in\Theta}(2f\Delta_l\phi)y_{lf}\right] + \beta\left[\sum_{k\in D}\left(\sum_{a\in A}c_a v_{ak} + \sum_{i\in S}W_{ik}\right)\varphi\right] \tag{5.24a}$$

$$\text{subject to} \quad (5.14b) - (5.14j) \tag{5.24b}$$

$$(5.14l) - (5.14r) \tag{5.24c}$$

$$\sum_{k\in D}\left[\sum_{i\in N}\mu_{ik}g_{ik} - \sum_{i\in S}\sum_{a\in FS(i)}\sum_{f\in\Theta}\left\{\overline{W}_{ik}\xi^1_{faik}\right\} - \sum_{a\in A_v}\left(\sum_{o\in O}d^{ok}\right)\xi^2_{ak}\right]$$
$$-\sum_{k\in D}\left(\sum_{a\in A}c_a v_{ak} + \sum_{i\in S}W_{ik}\right) \geq 0 \tag{5.24d}$$

$$(5.21b) - (5.21i) \tag{5.24e}$$

$$-\lambda^1_{faik} + \xi^1_{faik} \geq -\overline{\lambda}^1_{faik}(1-y_{l(a)f}), \forall f\in\Theta, \forall a\in FS(i), \forall i\in S, \forall k\in D \tag{5.24f}$$

$$\lambda^1_{faik} - \xi^1_{faik} \geq 0, \forall f\in\Theta, \forall a\in FS(i), \forall i\in S, \forall k\in D \tag{5.24g}$$

$$-\xi^1_{faik} \geq -\overline{\lambda}^1_{faik}y_{l(a)f}, \forall f\in\Theta, \forall a\in FS(i), \forall i\in S, \forall k\in D \tag{5.24h}$$

$$-\xi^2_{ak} + \tau_{ak} \geq 0, \forall a\in A_v, \forall k\in D \tag{5.24i}$$

$$\xi^2_{ak} - \tau_{ak} \geq -\overline{\tau}_{ak}(1-x_{l(a)}), \forall a\in A_v, \forall k\in D \tag{5.24j}$$

$$-\xi^2_{ak} \geq -\overline{\tau}_{ak}x_{l(a)}, \forall a\in A_v, \forall k\in D \tag{5.24k}$$

$$\xi^1_{faik} \geq 0, \forall f\in\Theta, \forall a\in FS(i), \forall i\in S, \forall k\in D \tag{5.24l}$$

$$\xi^2_{ak} \geq 0, \forall a\in A_v, \forall k\in D \tag{5.24m}$$

5.4 Branch-and-benders cut algorithm for BTNDP

Benders decomposition is an elegant way of solving a large scale MILP by iteratively solving two simpler subproblems: the relaxed master problem (RMP), which is a relaxation of the original problem, and a subproblem (SP), which provides optimality and feasibility cuts to strengthen the RMP [144]. In this section, we describe how Benders cuts can be incorporated in a branch-and-cut framework to achieve the bi-level optimal

solution.

Usually, in Benders decomposition, MILP is partitioned into RMP that consists of discrete variables and SP that consists of continuous variables. However, for the problem in hand, this leads to a multitude of feasibility cuts due to strict requirements of (5.14h) and (5.14i) which results in higher computational time. Instead, we restrict $(\mathbf{x}, \mathbf{y}, \mathbf{z}, \mathbf{v}, \mathbf{W}, \mathbf{t})$ in RMP and $(\mu, \lambda^{\mathbf{0}}, \lambda^{\mathbf{1}}, \lambda^{\mathbf{2}}, \tau, \xi^{\mathbf{1}}, \xi^{\mathbf{2}})$ in SP. Therefore, the initial RMP will be the same as HPP. This decomposition strategy offers the following advantages:

1. There are no feasibility cuts due (5.14h) and (5.14i) and the feasibility cuts are generated only for bi-level infeasible design solutions.

2. This decomposition strategy leads to multiple optimality cuts as SP can further be decomposed for every destination.

For a fixed value of $(\hat{\mathbf{x}}, \hat{\mathbf{y}}, \hat{\mathbf{z}}, \hat{\mathbf{v}}, \hat{\mathbf{W}}, \hat{\mathbf{t}})$, we can write the Benders subproblem as (5.25). Since the follower problem can be decomposed for every destination, the constraint (5.25b) is written for every $k \in D$.

$$\mathcal{Z}^{SP}(\hat{\mathbf{x}},\hat{\mathbf{y}},\hat{\mathbf{z}},\hat{\mathbf{v}},\hat{\mathbf{W}},\hat{\mathbf{t}}) = \underset{\mu,\boldsymbol{\lambda^0},\boldsymbol{\lambda^1},\boldsymbol{\lambda^2},\tau,\boldsymbol{\xi^1},\boldsymbol{\xi^2}}{\text{minimize}} 0 \tag{5.25a}$$

$$\text{subject to} \quad \sum_{i\in N}\mu_{ik}g_{ik} - \sum_{i\in S}\sum_{a\in FS(i)}\sum_{f\in\Theta}\left\{\overline{W}_{ik}\xi^1_{faik}\right\} - \sum_{a\in A_v}\left(\sum_{o\in O}d^{ok}\right)\xi^2_{ak}$$

$$\geq \left(\sum_{a\in A}c_a\hat{v}_{ak} + \sum_{i\in S}\hat{W}_{ik}\right), \forall k\in D \tag{5.25b}$$

$$-\mu_{ik}+\mu_{jk}+\lambda^0_{aik} \geq -c_a, \forall a=(i,j)\in A_a\cup A_t, \forall i\in N, \forall k\in D \tag{5.25c}$$

$$-\mu_{ik}+\mu_{jk}+\tau_{ak} \geq -c_a, \forall a=(i,j)\in A_v, \forall i\in N, \forall k\in D \tag{5.25d}$$

$$-\mu_{ik}+\mu_{jk} \geq -c_a, \forall a=(i,j)\in A_e, \forall i\in N, \forall k\in D \tag{5.25e}$$

$$\sum_{a\in FS(i)}\sum_{f\in\Theta}\lambda^2_{faik} = 1, \forall i\in S, \forall k\in D \tag{5.25f}$$

$$-f\lambda^0_{aik}+\lambda^1_{faik}+\lambda^2_{faik} \geq 0, \forall f\in\Theta, \forall a\in FS(i), \forall i\in S, \forall k\in D \tag{5.25g}$$

$$-\lambda^1_{faik}+\xi^1_{faik} \geq -\overline{\lambda}^1_{faik}(1-\hat{y}_{l(a)f}), \forall f\in\Theta, \forall a\in FS(i), \forall i\in S, \forall k\in D \tag{5.25h}$$

$$\lambda^1_{faik}-\xi^1_{faik} \geq 0, \forall f\in\Theta, \forall a\in FS(i), \forall i\in S, \forall k\in D \tag{5.25i}$$

$$-\xi^1_{faik} \geq -\overline{\lambda}^1_{faik}\hat{y}_{l(a)f}, \forall f\in\Theta, \forall a\in FS(i), \forall i\in S, \forall k\in D \tag{5.25j}$$

$$-\xi^2_{ak}+\tau_{ak} \geq 0, \forall a\in A_v, \forall k\in D \tag{5.25k}$$

$$\xi^2_{ak}-\tau_{ak} \geq -\overline{\tau}_{ak}(1-\hat{x}_{l(a)}), \forall a\in A_v, \forall k\in D \tag{5.25l}$$

$$-\xi^2_{ak} \geq -\overline{\tau}_{ak}\hat{x}_{l(a)}, \forall a\in A_v, \forall k\in D \tag{5.25m}$$

$$\lambda^0_{aik} \geq 0, \forall a\in FS(i), \forall i\in S, \forall k\in D \tag{5.25n}$$

$$\lambda^1_{faik}\geq 0, \lambda^2_{faik}\geq 0, \xi^1_{faik}\geq 0, \forall f\in\Theta, \forall a\in FS(i), \forall i\in S, \forall k\in D \tag{5.25o}$$

$$\xi^2_{ak}\geq 0, \tau_{ak}\geq 0 \forall a\in A_v, \forall k\in D \tag{5.25p}$$

Associating the dual variables $\{\gamma_k\}, \{\nu_{ak}\}, \{\mathcal{W}_{ik}\}, \{\mathfrak{t}_{faik}\}, \{\sigma^1_{faik}\}, \{\sigma^2_{faik}\}, \{\sigma^3_{faik}\}$,

$\{\sigma^4_{ak}\}, \{\sigma^5_{ak}\}, \{\sigma^6_{ak}\}$ to constraints (5.25b)-(5.25m), we can write the Benders dual subproblem as (5.26).

$$\mathcal{Z}^{DSP}(\hat{\mathbf{x}}, \hat{\mathbf{y}}, \hat{\mathbf{z}}, \hat{\mathbf{v}}, \hat{\mathbf{W}}, \hat{\mathbf{t}}) = \underset{\substack{\gamma,\nu,\mathcal{W},\mathbf{t},\boldsymbol{\sigma^1},\boldsymbol{\sigma^2},\\ \boldsymbol{\sigma^3},\boldsymbol{\sigma^4},\boldsymbol{\sigma^5},\boldsymbol{\sigma^6}}}{\text{maximize}} \sum_{k\in D}\Bigg[\left(\sum_{a\in A} c_a\hat{v}_{ak} + \sum_{i\in S}\hat{W}_{ik}\right)\gamma_k - \sum_{a\in A} c_a\nu_{ak} + \sum_{i\in S}\mathcal{W}_{ik}$$
$$- \sum_{i\in S}\sum_{a\in FS(i)}\sum_{f\in\Theta}\left(\overline{\lambda}^1_{faik}(1-\hat{y}_{l(a)f})\sigma^1_{faik} + \overline{\lambda}^1_{faik}\hat{y}_{l(a)f}\sigma^3_{faik}\right)$$
$$- \sum_{a\in A_v}\left(\overline{\tau}_{ak}(1-\hat{x}_{l(a)})\sigma^5_{ak} + \overline{\tau}_{ak}\hat{x}_{l(a)}\sigma^6_{ak}\right)\Bigg] \tag{5.26a}$$

$$\text{subject to} \quad \sum_{a\in FS(i)}\nu_{ak} - \sum_{a\in BS(i)}\nu_{ak} = \gamma_k g_{ik}, \forall i\in N, \forall k\in D \tag{5.26b}$$

$$\nu_{ak} - \sum_{f\in\Theta} f\mathfrak{t}_{faik} \le 0, \forall a\in FS(i), \forall i\in S, \forall k\in D \tag{5.26c}$$

$$\mathfrak{t}_{faik} - \sigma^1_{faik} + \sigma^2_{faik} \le 0, \forall f\in\Theta, \forall a\in FS(i), \forall i\in S, \forall k\in D \tag{5.26d}$$

$$\mathcal{W}_{ik} + \mathfrak{t}_{faik} \le 0, \forall f\in\Theta, \forall a\in FS(i), \forall i\in S, \forall k\in D \tag{5.26e}$$

$$-\overline{W}_{ik}\gamma_k + \sigma^1_{faik} - \sigma^2_{faik} - \sigma^3_{faik} \le 0, \forall f\in\Theta, \forall a\in FS(i), \forall i\in S, \forall k\in D \tag{5.26f}$$

$$\nu_{ak} + \sigma^4_{ak} - \sigma^5_{ak} \le 0, \forall a\in A_v, \forall k\in D \tag{5.26g}$$

$$-\left(\sum_{o\in O} d^{ok}\right)\gamma_k - \sigma^4_{ak} + \sigma^5_{ak} - \sigma^6_{ak} \le 0, \forall a\in A_v, \forall k\in D \tag{5.26h}$$

$$\nu_{ak} \ge 0, \forall a\in A, \forall k\in D \tag{5.26i}$$

$$\mathcal{W}_{ik} \ge 0, \forall i\in S, \forall k\in D \tag{5.26j}$$

$$\sigma^1_{faik}, \sigma^2_{faik}, \sigma^3_{faik} \ge 0, \forall f\in\Theta, \forall a\in FS(i), \forall i\in S, \forall k\in D \tag{5.26k}$$

$$\sigma^4_{ak}, \sigma^5_{ak}, \sigma^6_{ak} \ge 0, \forall a\in A_v, \forall k\in D \tag{5.26l}$$

Let P and R denote the indices of extreme points and extreme rays of the polyhedron and recession cone associated to the feasible region of (5.26). Then, the problem (5.24) can be reformulated as:

$$\mathcal{Z}^* = \underset{\mathbf{x},\mathbf{y},\mathbf{z},\mathbf{v},\mathbf{W},\mathbf{t},\zeta}{\text{minimize}} \quad (1-\beta)\left[\sum_{l\in L}\sum_{f\in\Theta}(2f\Delta_l\phi)y_{lf}\right] + \beta\left[\sum_{k\in D}\left(\sum_{a\in A}c_a v_{ak} + \sum_{i\in S}W_{ik}\right)\varphi\right] + \sum_{k\in D}\zeta_k \tag{5.27a}$$

$$\text{subject to} \quad (5.14\text{b}) - (5.14\text{j}) \tag{5.27b}$$

$$(5.14\text{l}) - (5.14\text{r}) \tag{5.27c}$$

$$\begin{aligned}\zeta_k \geq \Bigg[&\left(\sum_{a\in A}c_a v_{ak} + \sum_{i\in S}W_{ik}\right)\gamma_k - \sum_{a\in A}c_a(\nu_{ak})^e + \sum_{i\in S}(\mathcal{W}_{ik})^e \\ &- \sum_{i\in S}\sum_{a\in FS(i)}\sum_{f\in\Theta}\left(\overline{\lambda}^1_{faik}(1-y_{l(a)f})(\sigma^1_{faik})^e + \overline{\lambda}^1_{faik}y_{l(a)f}(\sigma^3_{faik})^e\right) \\ &- \sum_{a\in A_v}\left(\overline{\tau}_{ak}(1-x_{l(a)})(\sigma^5_{ak})^e + \overline{\tau}_{ak}x_{l(a)}(\sigma^6_{ak})^e\right)\Bigg], \forall k\in D, \forall e\in P\end{aligned} \tag{5.27d}$$

$$\begin{aligned}0 \geq \sum_{k\in D}\Bigg[&\left(\sum_{a\in A}c_a v_{ak} + \sum_{i\in S}W_{ik}\right)\gamma_k - \sum_{a\in A}c_a(\nu_{ak})^h + \sum_{i\in S}(\mathcal{W}_{ik})^h \\ &- \sum_{i\in S}\sum_{a\in FS(i)}\sum_{f\in\Theta}\left(\overline{\lambda}^1_{faik}(1-y_{l(a)f})(\sigma^1_{faik})^h + \overline{\lambda}^1_{faik}y_{l(a)f}(\sigma^3_{faik})^h\right) \\ &- \sum_{a\in A_v}\left(\overline{\tau}_{ak}(1-x_{l(a)})(\sigma^5_{ak})^h + \overline{\tau}_{ak}x_{l(a)}(\sigma^6_{ak})^h\right)\Bigg], \forall h\in R\end{aligned} \tag{5.27e}$$

(5.27) is HPP with the set of optimality cuts (5.27d) and feasibility cuts (5.27e) and decision variables $\{\zeta_k\}$, also known as the Benders master problem. The issue with the Benders reformulation is that there could be a large number of extreme points and rays of the polyhedron and recession cone associated with the feasible region of (5.26). Therefore, one starts solving the HPP and generates rows (5.27d) and (5.27e) as we find feasible solutions to this relaxed problem.

Most state-of-the-art MIP solvers offer a `Callback` option, which can be used to alter the behavior of the branch-and-cut algorithm used to solve a given MILP. The Benders cuts can be applied as `LazyCuts`, which eliminate feasible solutions of the master problem that are otherwise bi-level infeasible. The overall process is summarized

in the flow chart given in Figure 5.1 [145]. It starts by preparing and solving the (5.27) without (5.27d) and (5.27e). As we find a new feasible solution $(\hat{\mathbf{x}}, \hat{\mathbf{y}}, \hat{\mathbf{z}}, \hat{\mathbf{v}}, \hat{\mathbf{W}}, \hat{\mathbf{t}}, \hat{\zeta})$ to this problem, it is sent for the further investigation using a `Callback` procedure. The procedure is described as follows. For the new feasible solution $(\hat{\mathbf{x}}, \hat{\mathbf{y}}, \hat{\mathbf{z}}, \hat{\mathbf{v}}, \hat{\mathbf{W}}, \hat{\mathbf{t}}, \hat{\zeta})$ to the master problem, DSP (5.26) is solved. If (5.26) is found to be unbounded, then a lazy infeasibility cut (5.27e) using an unbounded ray to the current DSP is added to the (5.27). If (5.26) is found to be optimal but $\mathcal{Z}^{DSP} > \hat{\zeta} = \sum_{k \in D} \hat{\zeta}_k$, then using its solution multiple lazy optimal cuts (5.27d) are added for every destination. Otherwise, $(\hat{\mathbf{x}}, \hat{\mathbf{y}}, \hat{\mathbf{z}}, \hat{\mathbf{v}}, \hat{\mathbf{W}}, \hat{\mathbf{t}}, \hat{\zeta})$ is accepted as a new incumbent and used to update the upper bound of the overall problem. By continuing to add possible new cuts, the final solution obtained after the termination criterion is satisfied is optimal to (5.24).

5.4.1 Other improvements

Various cuts associated with solving the multi-commodity flow network design problem, such as `FlowCover` and `FlowPath` cuts, could help boost the lower bound of RMP. These cuts are available with state-of-the-art solvers such as Gurobi and CPLEX, which can be forced to generate such cuts aggressively. In this section, we propose cut-set-based inequalities and normalized Benders infeasibility cuts to accelerate the convergence. They are described below:

Cut-set based inequalities

A network cut is a partition of nodes N into two non-empty subsets N_1 and $N_2 = N \backslash N_1$. A cut-set and their associated commodity subset are denoted by $(N_1, N_2) = \{(i,j) \in A : i \in N_1, j \in N_2\}$ and $OD(N_1, N_2) = \{(o,d) \in O \times D : o \in N_1, d \in N_2\}$ respectively. A cut-set inequality represents that the total capacity provided on the edges in the cut-set should be greater or equal to the total demand of the associated commodity

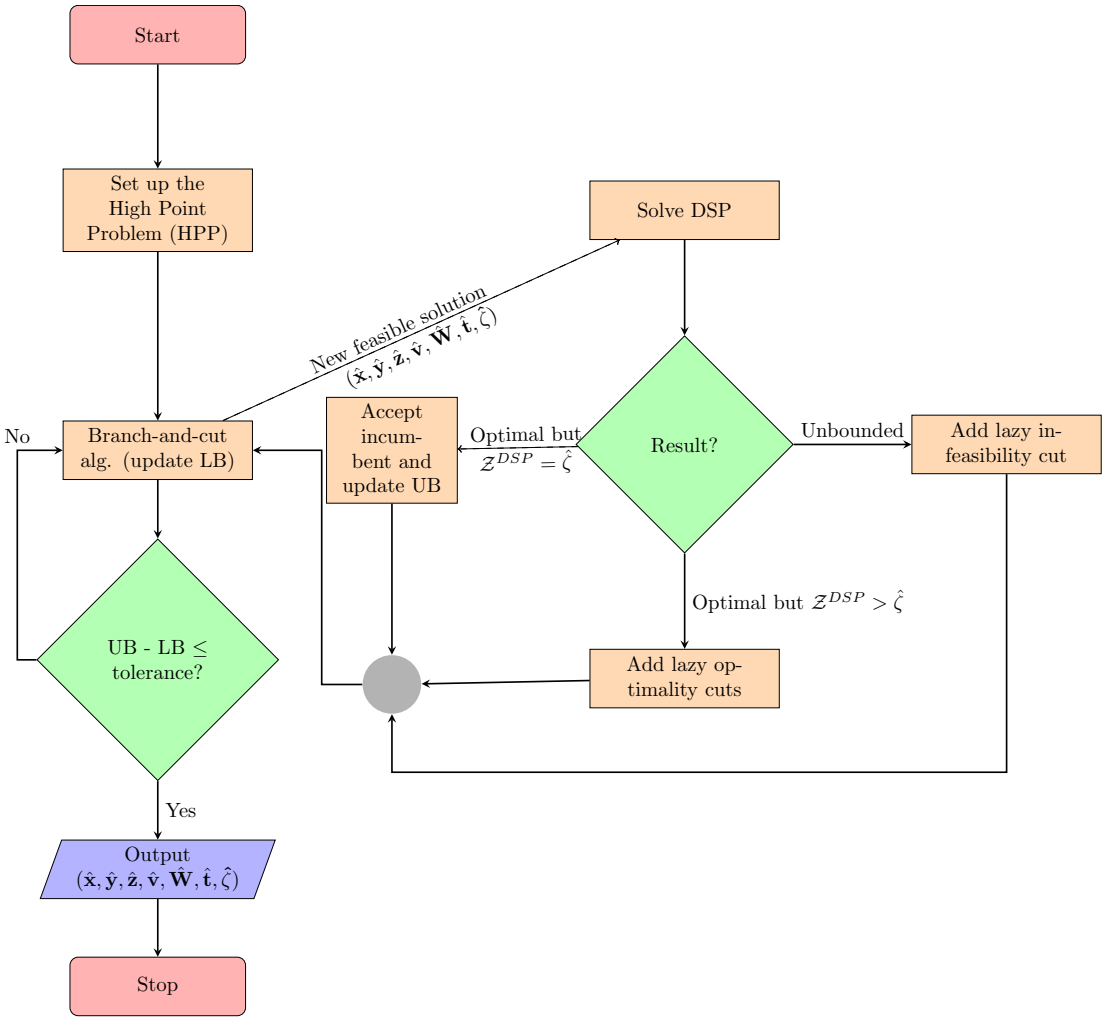


Figure 5.1: Branch-and-benders cut algorithm

subsets [142]. Since the number of cut-set inequalities can be exponentially large and are not sufficient to characterize the optimal solution to the BTNDP, we add a few cut-set-based inequalities that can be generated easily. They are specified as below:

$$\sum_{a \in FS(o)} \sum_{f \in \Theta} y_{l(a)f} \mathfrak{C} \geq \sum_{d \in D} d^{od}, \forall o \in O \tag{5.28}$$

$$\sum_{a \in BS(d)} \sum_{f \in \Theta} y_{l(a)f} \mathfrak{C} \geq \sum_{o \in O} d^{od}, \forall d \in D \tag{5.29}$$

(5.28) and (5.29) describe that the total capacity of the links coming out of every origin and going respectively into every destination should be greater than or equal to the demand to be sent and received respectively at these nodes.

Normalizing Benders infeasibility cuts

An efficient Benders infeasibility cut can be generated by maximizing the cut violation with a normalization condition to deal with the unboundedness of the dual problem [146,147]. In this case, the cut generation problem can be stated as a *separation problem* as below:

$$0 \leq \hat{\zeta} \tag{5.30}$$

$$(5.25\text{b}) - (5.25\text{p}) \tag{5.31}$$

The above system of inequations is infeasible, so their dual should be unbounded. Associating dual variables π_0 and rest same as (5.26), we can state the dual problem as (5.32) with a normalization condition (5.32c), where each variable is assigned a weight and the total normalized sum is restricted to 1. [147] proposed to use weights $w_i = 1$

for all the variables which appears in the objective function (5.32a), and 0, otherwise.

$$\mathcal{Z}^{DSP}(\hat{\mathbf{x}}, \hat{\mathbf{y}}, \hat{\mathbf{z}}, \hat{\mathbf{v}}, \hat{\mathbf{W}}, \hat{\mathbf{t}}) =$$

$$\underset{\gamma,\nu,\mathcal{W},\mathbf{t},\boldsymbol{\sigma^1},\boldsymbol{\sigma^2},\boldsymbol{\sigma^3},\boldsymbol{\sigma^4},\boldsymbol{\sigma^5},\boldsymbol{\sigma^6},\boldsymbol{\pi_0}}{\text{maximize}} \quad -\pi_0\hat{\zeta} + \sum_{k\in D}\left[\left(\sum_{a\in A} c_a\hat{v}_{ak} + \sum_{i\in S}\hat{W}_{ik}\right)\gamma_k - \sum_{a\in A} c_a\nu_{ak} + \sum_{i\in S}\mathcal{W}_{ik}\right.$$
$$-\sum_{i\in S}\sum_{a\in FS(i)}\sum_{f\in\Theta}\left(\overline{\lambda}^1_{faik}(1-\hat{y}_{l(a)f})\sigma^1_{faik} + \overline{\lambda}^1_{faik}\hat{y}_{l(a)f}\sigma^3_{faik}\right)$$
$$\left.-\sum_{a\in A_v}\left(\overline{\tau}_{ak}(1-\hat{x}_{l(a)})\sigma^5_{ak} + \overline{\tau}_{ak}\hat{x}_{l(a)}\sigma^6_{ak}\right)\right] \tag{5.32a}$$

$$\text{subject to} \quad (5.26b) - (5.26l) \tag{5.32b}$$

$$w_0\pi_0 + \mathbf{w}^T\text{vec}(\gamma,\nu,\mathcal{W},\mathbf{t},\boldsymbol{\sigma^1},\boldsymbol{\sigma^2},\boldsymbol{\sigma^3},\boldsymbol{\sigma^4},\boldsymbol{\sigma^5},\boldsymbol{\sigma^6}) = 1 \tag{5.32c}$$

This helps in generating the stronger cut in the following form:

$$0 \geq -\pi_0\hat{\zeta} + \sum_{k\in D}\left[\left(\sum_{a\in A} c_a\hat{v}_{ak} + \sum_{i\in S}\hat{W}_{ik}\right)\gamma_k - \sum_{a\in A} c_a\nu_{ak} + \sum_{i\in S}\mathcal{W}_{ik}\right.$$
$$-\sum_{i\in S}\sum_{a\in FS(i)}\sum_{f\in\Theta}\left(\overline{\lambda}^1_{faik}(1-\hat{y}_{l(a)f})\sigma^1_{faik} + \overline{\lambda}^1_{faik}\hat{y}_{l(a)f}\sigma^3_{faik}\right)$$
$$\left.-\sum_{a\in A_v}\left(\overline{\tau}_{ak}(1-\hat{x}_{l(a)})\sigma^5_{ak} + \overline{\tau}_{ak}\hat{x}_{l(a)}\sigma^6_{ak}\right)\right] \tag{5.33}$$

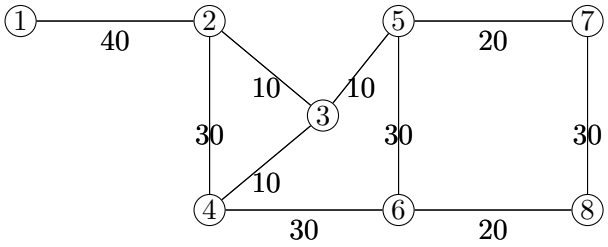
5.5 Numerical Results

In this section, we present the results of the numerical experiments performed for BT-NDP. The network design problem is solved using the instances provided by [3], [148], and [4]. The numerical results section is organized as follows. The first sub-section solves a small case study and presents the results of sensitivity analysis on various parameters used in the optimization model. In subsection 5.5.2, we present the results of computational performance of the current methodology on different network instances. All implementations were coded in Python 3.8, using Gurobi 9.1.2 as the optimization

solver. The tests were executed on Intel(R) Xenon(R) CPU running at 2.2 GHz with 128 GB RAM under a Windows operating system.

5.5.1 BTNDP results: a small case study

In this section, we consider a small network given in [3]. It has 8 nodes, 10 links, and 4 O-D pairs. The network and demand table is shown in Figure 5.2. The demand shown in the table is the peak hour demand between various origin-destination pairs. For



(a) Small network

From \To	1	7	8	Total
3	60	0	120	180
4	0	30	0	30
6	30	0	0	30
Total	90	30	120	240

(b) Demand table

Figure 5.2: Network and demand table [3]

creating the candidate transit routes, we set the shortest path deviation factor $\tilde{\delta} = 20\%$ and put no restrictions on the length of a transit route. The set of terminals is given by $\mathfrak{T} = \{3, 4, 6\}$. Using the Algorithm 7, 8 candidate transit routes are generated which are specified below:

a. 4-3-5-7

b. 6-4-3-2-1

c. 6-5-3-2-1

d. 6-5-7

e. 6-8-7

f. 3-4-6-8

g. 3-5-6-8

h. 3-5-7-8

The transit network created using the candidate transit routes has 40 nodes and 226 links. The parameters used in the BTNDP model for this experiment are provided in Table 5.1.

Table 5.1: Parameter values used in the experiment

Parameter	Value
Set of possible frequencies, Θ	$\{1, 2, 3, 6, 12\}$ buses/hr
Commercial speed of bus	0.25 mi/min
Operating cost, ϕ	\$15/mi
Fleet size, B	100
Road link capacity, κ	60
Bus passenger capacity, $\mathfrak{C}$	60
Average no. of transfers allowed, $\tilde{\epsilon}$	1
Value of travel time, φ	\$15/hr
Passengers' perspective weight, β	0.3

The selected transit routes in the optimal network design and their corresponding frequencies are a, b, and e and 2, 2, and 1 bus/hr respectively. We observe that the frequencies are assigned based on the ridership of located routes, given by 345, 630, and 370. The total operating cost and number of buses required for providing the proposed service are \$2,326 and 11 respectively. The total time spent in the system is equal to 297 passenger-hours, including 235 passenger-hours of travel time spent on various links and 62 passenger-hours of wait time spent at various nodes of the network. The average number of transfers per passenger is found to be equal to 1, which shows that the soft transfer constraint (5.7) is binding.

Most studies model TNDP as a single-level problem, which in this case, is the high

point problem (HPP). It is interesting to note the difference between the solution of BTNDP and HPP. The total operator and passenger cost in dollars by solving the HPP is \$6,900 as compared to BTNDP's total cost value, which is equal to \$6,789. We call the difference between both objective values, which is \$111, as the price of modeling the problem as a bi-level problem. Moreover, we found that the passenger flow values obtained by solving the optimal strategy transit assignment using the single-level design solution are not able to satisfy (5.7) and (5.8). This shows that it is important to model the problem as a bi-level problem for predicting realistic passenger behavior.

Sensitivity analysis

For sensitivity analysis, we selected various parameters used in this study. This includes B, β, κ, and $\mathfrak{C}$. We found that changing κ and $\mathfrak{C}$ did not significantly affect the final results, and therefore, we do not include the sensitivity analysis of these parameters.

Effect of B

For this experiment, we solve (5.24) for varying bus fleet size of 15, 25, 50, 75, 100, 120, 150, and 200. To emphasize the passengers' perspective, we set $\beta = 1$. The results of sensitivity analysis are plotted in Figure 5.3. In particular, it shows the effect of fleet size on various types of costs, number of buses, and number of routes located. Overall, the total passengers' cost reduces with the increase in the fleet size, as evident from Figure 5.3(a). The reduction of passenger cost is caused by the reduction in the wait time (Figure 5.3(c)) with increasing deployment of buses. On the other hand, the in-vehicle travel time does not change with varying fleet sizes (Figure 5.3(b)). This is because the increase in the frequency does not affect the in-vehicle travel time of parallel routes. The overall operating cost increases with the increase in the fleet size (Figure 5.3(d)). This

is intuitive as the higher fleet size provides an opportunity to assign a higher frequency to the routes, which results in more capacity on the attractive routes and lesser wait time. Figure 5.3(e) and (f) provide details of the effect of fleet size on located routes and the number of buses deployed. Sometimes, the number of located routes decreases with the increase in the fleet size. This is because sometimes less number of routes with higher frequencies leads to better passenger costs as compared to more number of routes with lesser frequencies. For example, the located routes and their frequencies at the fleet size of 75 are a, b, f, g, h and 1, 3, 12, 12, 12 respectively. On the other hand, the located routes and their frequencies at the fleet size of 80 are a, b, f, g, and 12, 12, 12, 12 respectively. The increase in the frequency with increasing fleet size can be observed in Figure 5.3(f), where we see that all buses in the fleet are always utilized.

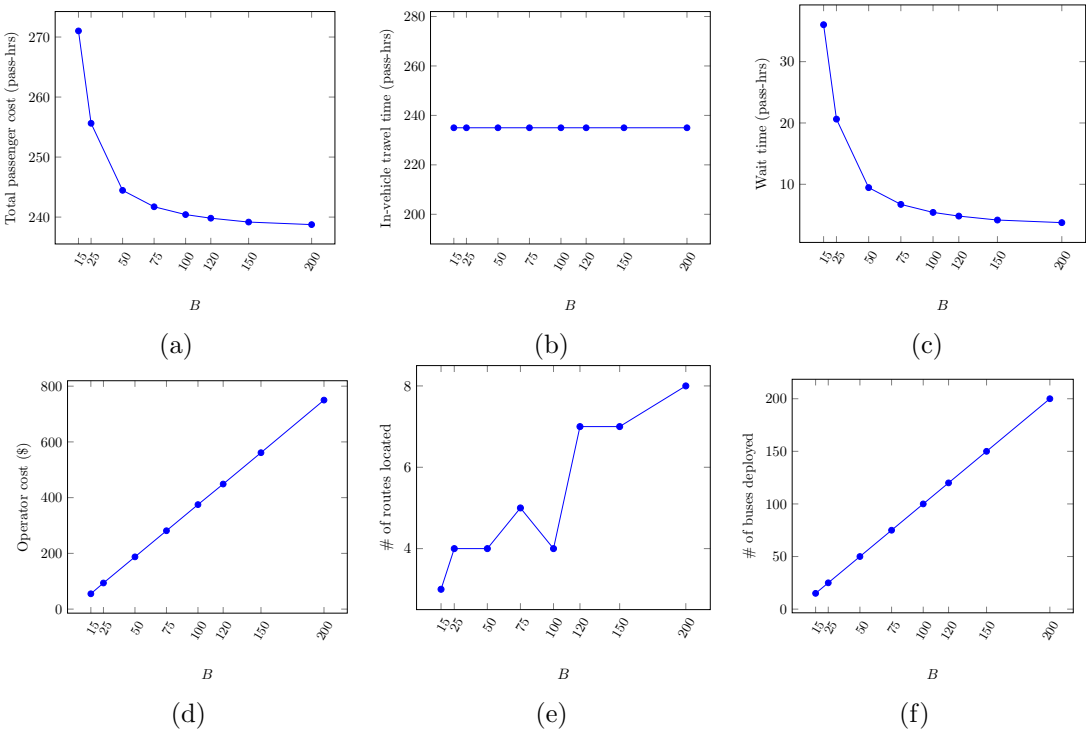


Figure 5.3: Sensitivity of B on passenger cost, operator cost, number of located routes, and number of buses deployed

Effect of β

To analyze the effect of competing perspectives of operator and passengers, we conduct the sensitivity analysis on the weights assigned to the passengers' and operator's perspectives in the BTNDP model. For this purpose, the value of β is varied from 0 to 1 and the results are plotted in Figure 5.4. Similar to the fleet size, the increase in the weight to the passengers' perspective leads to decrease in overall passengers cost (Figure 5.4(a)) and increase in the operator cost (Figure 5.4(d)). The decrease in the passenger cost is due to the reduction of wait time of passengers with the adoption of a higher frequency of routes. The in-vehicle time is not affected by the value of β. Similarly, the number of located routes does not change significantly as they vary between 3 and 4. By looking at the plots, an important observation about the range of β, for which we observe prominent effects, can be made. The value of β between 0.9 and 1 has a drastic effect on the evaluated values. For example, the number of buses deployed increases from 12 to 100 between this range. On the other hand, the range of β between 0 to 0.9 does not have a comparable effect on the design. Overall, the plots show that including the operator's perspective is important to obtain an economical design with an acceptable level of service to passengers.

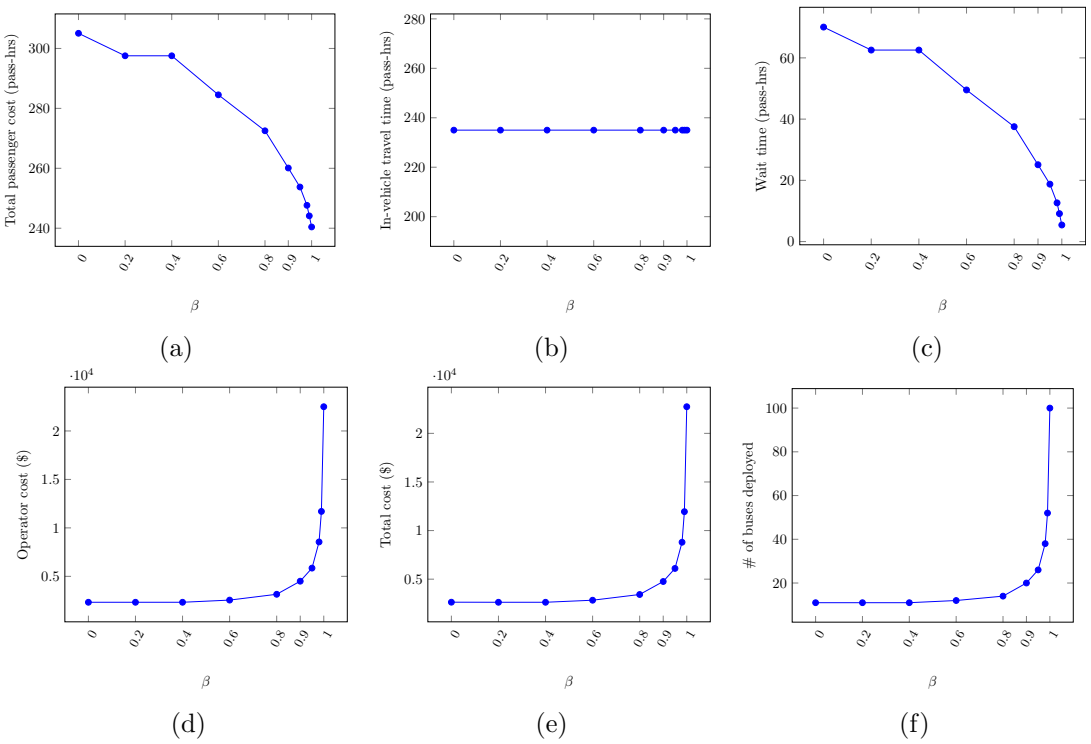


Figure 5.4: Sensitivity of β on passenger cost, operator cost, and number of buses deployed

Effect of transfer constraints

For an aggregated model, it is difficult to track passengers and evaluate the exact number of transfers for each passenger. Therefore, we analyze the total number of passengers on transfer links and observe the effect of different transfer constraints proposed in Section 5.2.2. To emphasize the passengers' perspective, we set $\beta = 1$. Table 5.2 shows the total passenger flow on transfer links by employing different transfer constraints. Without any transfer constraints, the total transferring flow is equal to 750. This gets reduced to 495 when we apply the hard transfer constraints only. On the other hand, with only soft transfer constraints, the total passenger flow on transfer links is equal to 240, which remains the same when we include both hard and transfer constraints. This shows that the soft transfer constraint is binding in both cases. This analysis shows that even

though hard transfer constraints ensure the existence of a direct or transfer path with at most one transfer, some passengers still prefer to take more transfers to save time. On the other hand, with soft transfer constraints, the passengers comparatively take a lesser number of transfers.

Table 5.2: Effect of transfer constraints on passenger flow

Transfer constraints	Total flow on transfer links
No transfer constraints	750
Hard transfer constraints only	495
Soft transfer constraints only	240
Both soft and hard transfer constraints	240

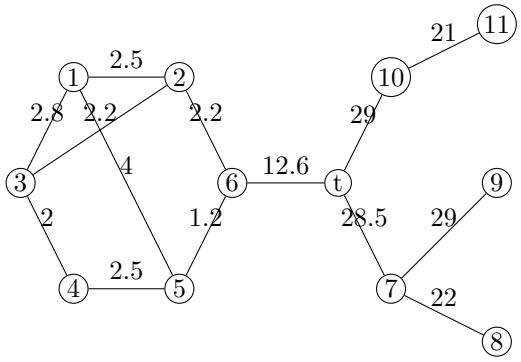
5.5.2 Computational performance

For analyzing the computational performance of the method proposed in this chapter, we solve different network instances. The parameter values are kept the same for every instance as given in Table 5.1. The details of these instances are provided below:

1. *Network 1*: This is the same as the network instance used in the previous section. It has 8 nodes, 10 links, and 4 OD pairs. The network is visualized in Figure 5.2. The number of candidate routes considered is 8. The transit network created using the candidate routes has 40 nodes and 226 links. The number of decision variables and constraints in the initial relaxed master problem are equal to 1,258 (including 1,182 continuous, 76 binary) and 1,547 respectively.

2. *Network 2* [4]: This network has 12 nodes, 14 links, and 30 O-D pairs. The network is visualized in Figure 5.5. The number of candidate routes considered is 9. The transit network created using the candidate routes has 87 nodes and 766 links. The number of decision variables and constraints in the initial relaxed

master problem are equal to 5,858 (including 5,766 continuous, 92 binary) and 7,468 respectively.



(a) Network 2

From \To	7	8	9	10	11	Total
1	192	148	102	94	149	685
2	100	74	78	56	102	410
3	87	77	71	46	113	394
4	96	63	49	34	85	327
5	33	24	19	15	34	125
6	19	14	14	9	23	79
Total	527	400	333	254	506	2020

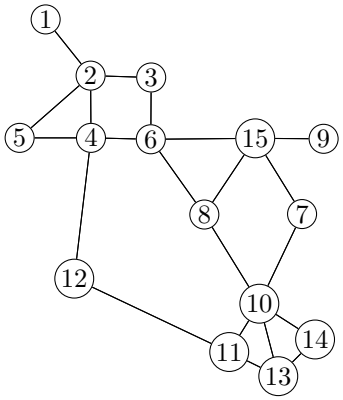
(b) Demand table

Figure 5.5: Network and demand table [4]

3. *Network 3* [5]: This network has 15 nodes, 21 links, and 172 O-D pairs. The network is visualized in Figure 5.6. As per [79], the time horizon of the demand is 24 hours. To obtain the peak hourly demand, we divide the demand values by a factor of 8. For creating candidate transit routes, we consider the minimum and maximum length of the routes as 40 and 100 units. The number of candidate routes generated is 21. The transit network created using the candidate routes has 181 nodes and 2,485 links. The number of decision variables and constraints in the initial relaxed master problem are equal to 46,588 (including 46,243 continuous, 345 binary) and 86,411 respectively.

Before delving into the final design of the networks described above, we compare the performance of four different strategies. These strategies are described below:

1. *Strategy 1*: Solving BTNDP directly using Gurobi solver. For this case, we keep the bi-linear terms, i.e., do not use any McCormick relaxations anywhere, and reformulate the bi-level model into a single-level model using the KKT conditions



(a) Network 3

0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	Total
1	0	50	25	7	10	18	9	9	3	20	3	3	4	0	161
2	50	0	6	15	2	22	11	11	1	16	2	1	1	0	138
3	25	6	0	5	7	22	11	11	1	5	2	1	1	0	97
4	7	15	5	0	6	12	6	6	1	30	5	3	1	0	97
5	10	2	7	6	0	6	3	3	1	15	2	1	0	0	56
6	18	22	22	12	6	0	12	12	3	110	7	1	1	1	227
7	9	11	11	6	3	12	0	6	1	55	4	1	1	0	120
8	9	11	11	6	3	12	6	0	1	55	4	1	1	0	120
9	3	1	1	1	1	3	1	1	0	17	2	0	0	0	31
10	20	16	5	30	15	110	55	55	17	0	75	31	62	25	516
11	3	2	2	5	2	7	4	4	2	75	0	9	11	1	127
12	3	1	1	3	1	1	1	1	0	31	9	0	8	0	60
13	4	1	1	1	0	1	1	1	0	62	11	8	0	5	96
14	0	0	0	0	0	1	0	0	0	25	1	0	5	0	32
Total	161	138	97	97	56	227	120	120	31	516	127	60	96	32	1878

(b) Demand table

Figure 5.6: Network and demand table [5]

of the follower problem. The resulting model is a mixed-integer non-linear program. Gurobi 9.1 has the functionality to solve mixed-integer bilinear programs. For that purpose, we set the Gurobi parameter `NonConvex = 2`.

2. *Strategy 2*: Solving BTNDP model (5.24) using Gurobi solver. This is a mixed-integer linear program that can be solved using Gurobi.

3. *Strategy 3*: Solving BTNDP model (5.24) using Branch-and-Benders cut algorithm described in Section 5.4 without any improvements mentioned in Section 5.4.1. For this purpose, we define the HPP model and set the Gurobi parameter `lazyConstraints = 2`.

4. *Strategy 4*: Solving BTNDP model (5.24) using Branch-and-Benders cut algorithm described in Section 5.4 with improvements such as cut-set-based inequalities and normalizing infeasibility cuts mentioned in Section 5.4.1. For this purpose, we define the HPP model with extra cut-set-based inequalities (5.28)-(5.29) and set the Gurobi parameter `lazyConstraints =2`.

For all the above tests, the maximum time limit was set to 12 hours. The results

of the computational performance are shown in Table 5.3. The computational time is recorded in seconds, and the gap value used by the Gurobi solver, which is defined as $\frac{(UB-LB)*100}{UB}$, is reported in percentage.

Table 5.3: Computational performance of different strategies

Network		*Network 1*	*Network 2*	*Network 3*
Strategy 1	Best objective value ($)	2,967	22,986	-**
	Gap (%)	0	0	-**
	Computational time (s)	110	925	Timed out*
Strategy 2	Best objective value ($)	2,967	22,986	15,249
	Gap (%)	0	0	15.4
	Computational time (s)	2	96	Timed out*
Strategy 3	Best objective value ($)	2,967	22,986	14,956
	Gap (%)	0	0	2.4
	Computational time (s)	2	105	Timed out*
	User cuts added	4	149	178
Strategy 4	Best objective value($)	2,967	22,986	14,956
	Gap (%)	0	0	0
	Computational time (s)	2	65	35,887
	User cuts added	10	35	121

*Maximum time limit = 12 hours.
**No valid upper or lower bound was found in 12 hours.

The mixed-integer bilinear program for BTNDP is hard to solve directly using Gurobi, and therefore, Strategy 1 shows the worst performance among all the strategies. It was able to solve the first two instances in a reasonable amount of time but could not even compute the lower as well as upper bound in the time limit for Network 3. The Network 1 instance is small, and therefore, all the strategies apart from Strategy 1 can solve this in 2 seconds to optimality. The Network 2 instance is a bit more complex than the Network 1 instance due to the higher number of origin-destination pairs. For this instance, Strategy 4 shows the best performance among all the strategies and solves (5.24) in just 65 seconds. Strategy 2 can solve this instance faster than Strategy 3. The

Network 3 instance has a high number of origin-destination pairs, which results in a large-scale optimization problem. For this instance, Strategy 4 is the only strategy that can solve the model to optimality in time limit. Strategy 3 also shows good performance and can find the optimal solution within the time limit. Although, it could solve the model only up to 2.4 % gap. Finally, Gurobi can find a feasible solution close to the optimal solution but could only reach the optimality gap of 15.4 % in the time limit. We observed that both strategies 3 and 4 can find the optimal solution for Network 3 instance in 5 and 2 hours respectively. However, they spend most of the time closing the gap. This shows that the lower bound of the linear relaxation of the problem is poor, and strong valid inequalities are needed to further improve the convergence. We observed that adding extra cut-set-inequalities in Strategy 4 helped boost the initial lower bound. Further, normalizing benders infeasibility cuts produces deeper cuts than the classic Benders cuts. This is evident from the lower number of user cuts added by Strategy 4 to achieve optimality than Strategy 3.

When performing the computations, we found that the number of trips in the network directly affects the computational time. To show this effect, we conduct a sensitivity analysis on the demand intensity. For this purpose, we multiply the demand between various origin-destination pairs using a demand factor equal to 0.5, 1, and 2. We solve the model (5.24) using Strategy 4 for various network instances and visualize the computational times in Figure 5.7. The results clearly show that the computational time is proportional to the demand intensity, i.e., increasing the demand increases the computational time to solve the model.

The optimal design results of different network instances are shown in Table 5.4. The total number of routes located in various networks is given as 3, 6, and 7 respectively. Due to heavy demand in the Network 2 instance, we observe that high frequency

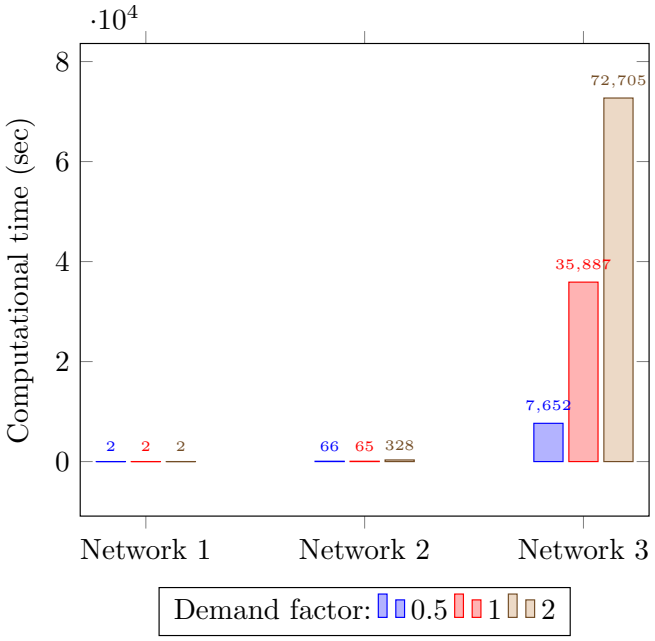


Figure 5.7: Computational time versus demand intensity (solved using Strategy 4)

is assigned to the located routes. This leads to more buses deployed in the network to serve the demand. The operating cost is also higher in Network 2 due to longer routes in the network. For Network 3 instance, we observe that most of the routes are assigned a frequency equal to 1 bus/hr except one route, which is assigned a frequency of 3 buses/hr. To provide this service, only 11 buses are needed.

One of the reasons to select the instance given in [1] (Network 2) for experiments is because they also present a bi-level transit network design model. Moreover, they use the optimal strategy transit assignment for modeling the passenger behavior in the lower level. To calculate the benchmark result, they use a brute force method (that enumerates all possible solutions). This benchmark result is presented in Table 5.5. The results of the benchmark model are better than the result computed by our model. This is because of the following reasons. First, [1]'s model frequency as a continuous

Table 5.4: Network design results of various instances

Network	Route	Frequency (buses/hr)	Operator cost ($)	Passenger cost (pass-hrs)	Total buses deployed
Network 1	4-3-5-7	2	2,326	297	11
	1-2-3-4-6	2			
	6-8-7	1			
Network 2	2-1-3-4-5-6-t-7-8	6	18,546	2,223	86
	1-2-3-4-5-6-t-7-8	3			
	2-1-3-4-5-6-t-10-11	12			
	2-1-3-4-5-6-t-7-9	6			
	2-3-4-5-6-t-10	1			
	2-3-4-5-6-t-7	6			
Network 3	12-11-10-7	1	16,662	731	11
	5-4-6-8-15-9	1			
	14-13-11-10-8-15-9	1			
	14-13-11-10-7-15-9	1			
	11-13-14-10-8-6-3-2-4-5	1			
	14-13-11-10-8-6-3-2-1	3			
	12-4-6-8-15-7	1			

variable not an integer value. Second, their model does not consider the explicit transfer, street capacity, and route capacity constraints. This makes the feasible region of the model larger to obtain a better objective value. Furthermore, we believe communicating fractional frequencies to the passengers would be difficult in real-life.

Table 5.5: Benchmark network design results from [1]

Network	Route	Frequency (buses/hr)	Operator cost ($)	Passenger cost (pass-hrs)	Total buses deployed
Network 2 benchmark results	1-3-2-6-t-7-8	4.51	13,696	2,084	62
	1-3-2-6-t-7-9	4.84			
	1-3-2-6-t-10-11	4.54			
	2-3-4-5-6-t-10	4.88			
	2-3-4-5-6-t-7	10.17			

Chapter 6

Design of integrated transit systems with strategic passenger assignment

The introduction of Mobility-on-Demand (MoD) services such as Uber, Lyft, and others as transportation alternatives has created many opportunities as well as challenges. On one hand, they provide a seamless mobility service with just a few taps on a cellphone application. On the other hand, it has increased congestion in densely populated areas due to an increase in the relocation and pickup trips made by the participating drivers in the network [149]. Furthermore, the transportation agencies envision the introduction of Autonomous Vehicles as a shared mobility service in the near future [150], which would lead to severe congestion in densely populated areas as predicted by various simulation studies [113, 151, 152].

Public transportation, which can carry multiple passengers, is widely considered as

a practical solution to the congestion problem by reducing vehicle-miles traveled (VMT) on roads [153]. However, due to its fixed routes and schedules, limited network coverage, and waiting time, sometimes, it is less attractive to travelers in comparison to the auto mode. The limited network coverage makes it difficult or sometimes impossible to access transit service in some areas. This inaccessibility problem is also known as the *first mile/last mile (FMLM) problem for transit*. The problem is commonly faced by travelers commuting from low-density areas where transit service is not available or less frequent because of the economic in-viability of providing such service.

A few studies have argued that the Mobility-on-demand service provided using autonomous vehicles would become a competitor of public transit mode [113, 154, 155], reducing its ridership, and other studies have even raised the question of whether urban mobility is possible without the classical public transit service [115,156]. However, [119] showed that the integration of the MoD system with transit could help in achieving better results, such as a significant reduction in travel time, emissions, and costs as compared to the standalone MoD system. Through the current research, we also envisage an integrated MoD and transit system that aims to achieve the following potential benefits:

1. Providing fast and reliable mobility in low-density areas (i.e., by providing a first mile/last mile service) by means of characteristics of MoD service such as demand responsiveness, fleet repositioning, and reachability.

2. Allocation of resources from less congested areas to providing high-frequency transit service in congested areas through such integration.

3. Using existing transit infrastructure to reduce the number of vehicles needed for serving trips.

4. Reducing congestion and carbon emissions in the network, improving the mobility of travelers, and reducing the overall cost of providing transit service.

To achieve the above-mentioned benefits, we focus on the strategic planning of the transportation network that allows for intermodal trips with the first or last leg of the trips being served by the MoD service. To be specific, we try to answer questions such as which transit routes to operate when MoD vehicles are deployed to serve the FMLM connection, what should be the size of the vehicle fleet to be deployed, and what should be the frequency of operating transit routes.

6.1 Preliminaries and Background

The notations introduced in Chapter 5 are also used in the current chapter. Let us begin by considering a multimodal transportation network characterized by a digraph $G(V, E)$, where V denotes the set of nodes that includes road intersections N_R, transit stops/stations N, and centroids of traffic analysis zones Z and A denotes the set of links. We associate every node $i \in V$ in the network with exactly one zone $Z(i)$. Depending on the mode, the links are also divided into three categories, namely transit, road, mode transfer, and access/egress links represented by $A, A_R, A_e,$ and M respectively. Let $O \subset Z$ and $D \subset Z$ be the subsets of centroids representing the origins and destinations respectively. The demand between various origin-destination pairs is represented by $\{d^{od}\}_{(o,d)\in O\times D}$. The overall network can be divided into three sub-networks which are described below:

1. *Transit network*: The transit network is characterized by the subgraph $G(N, A)$ which consists of a set of candidate transit lines/routes denoted by the set L. The terms "route" and "line" are used interchangeably throughout this chapter. Each line $l \in L$ is composed of a set of stops $N^l \subset N$ which are connected by edges

$A^l \subset A$. The network also consists of transfer links A_t between two nodes if the walking distance between those is less than the acceptable walking distance ζ (say 0.75mi).

2. *Road network*: The road network is characterized by the subgraph $G_R(N_R, A_R)$, where N_R denotes the set of nodes and A_R denotes the set of links in the road network.

3. *Walking links*: The access and egress walking links A_e connect the centroids of various zones with the road/transit nodes and vice-versa, whereas mode transfer links M are used to transfer between nodes of various modes.

6.1.1 Costs

There is a subset of nodes in the network where passengers have to wait for the service. The collection of head nodes of links in the sets A_e, A_t, and M constitutes the *waiting nodes* N^w. Let us assume that $c : E \mapsto \mathfrak{R}_+$ and $w : N^w \mapsto \mathfrak{R}_+$ denote the cost (e.g., walking time, in-vehicle time, and fare) associated with the links in E and waiting time associated with the nodes in N^w respectively. The cost of links is known beforehand (and is computed by adding the travel time and possible fare multiplied by the value of time). On the other hand, the wait time depends on the availability of MoD or transit service.

6.1.2 Waiting time computation

Unlike a personal vehicle, the MoD or transit service is not readily available, and passengers have to wait to access these services. So, it is important to quantify the expected wait time of these services, the computation of which is discussed below:

MoD service

We assume MoD operations in a network as a queuing system to compute the average waiting time experienced by the passengers to access such service. The average wait time may not be justified for the planning of day-to-days operations but can be used to approximate the actual wait time experienced by the passengers for long-term strategic planning of the network, which is the focus of the current study. Therefore, we consider a stationary state of an MoD system, where the number of waiting customers $\mathcal{C}$ and vacant vehicles $\mathcal{V}$ are time-invariant. Using the Cobb-Douglas production function, the matching time between the customers and the vacant vehicles can be expressed as a function of $\mathcal{C}$ and $\mathcal{V}$.

$$m^{c-v} = \mathcal{A}(\mathcal{V})^{\alpha_1}(\mathcal{C})^{\alpha_2} \tag{6.1}$$

where, α_1 and α_2 are defined as the elasticities of the matching function and $\mathcal{A}$ is a parameter specific to a zone, which is a function of the market area divided by the running speed in that zone [157]. According to Little's law, the long-term average number of customers/drivers in a stationary system is equal to the long-term average arrival rate Q multiplied by the average wait time (w^c/w^t) that a customer/driver spends in the system before being matched ([157]).

$$\mathcal{V} = Qw^t \tag{6.2}$$

$$\mathcal{C} = Qw^c \tag{6.3}$$

Using (6.3) and assuming $\alpha_1 = \alpha_2 \approx 1$ [158], we can represent the stationary state

$(m^{c-v} = Q)$ as below:

$$Q = \mathcal{AV}(Qw^c) \tag{6.4}$$

$$\implies w^c = \frac{1}{\mathcal{AV}} \tag{6.5}$$

Equation (6.5) shows that the average waiting time of customers waiting in a zone to access the MoD service is a function of the vacant number of vehicles. To achieve the desired level of service (i.e., average waiting time), a transportation agency needs to provide $\mathcal{V}$ vehicles at any point in time.

Transit service

Let us now discuss the wait time computation to access transit service at the head node of an access or transfer link in the transit network. Let $f : A \mapsto \mathfrak{R}$ be the frequency of the transit line associated with various links of the transit network. Let $\mathfrak{g}_i(w)$ be the probability distribution function of the waiting time for line i. According to [159], for the passengers arriving randomly at a node, the probability density function of the waiting time of line i is related to the headway or bus inter-arrival time distribution $\mathfrak{h}_i(h)$ as:

$$\mathfrak{g}_i(w) = \frac{\int_w^\infty \mathfrak{h}_i(h)dh}{\mathbb{E}[\mathfrak{h}_i]} \tag{6.6}$$

To evaluate the waiting time distribution, we make the following assumptions:

Assumption 4. *The inter-arrival time of a transit line $i \in L$ follows an exponential distribution with rate f_i.*

Assumption 5. *Passengers want to minimize the expected wait time to get to their destination. Therefore, at any node, passengers waiting to be served by the transit*

service have selected a list of attractive transit lines that can help them to get to their destination.

Both assumption 4 and 5 are common in the transit assignment literature (e.g., see [160]). By using the assumption 4 and equation (6.6), one can evaluate the distribution function of the wait time $\mathfrak{g}_i(w)$ as:

$$\mathfrak{g}_i(w) = f_i e^{-f_i w}, w \geq 0 \tag{6.7}$$

Proposition 7. *[7,55] Assuming that a passenger waiting at node $n \in N^w$ is served by the set of attractive transit lines $FS^*(n)$ and let $\mathfrak{F} = \sum_{j \in FS^*(n)} f_j$. With assumptions 4 and 5, the following holds:*

1. *The probability that a passenger would choose transit line $i \in FS^*(n)$ is given by*

$$P_i = \frac{f_i}{\mathfrak{F}} \tag{6.8}$$

2. *The expected wait time conditional to boarding line $i \in FS^*(n)$ is given by*

$$EW_i = \frac{f_i}{\mathfrak{F}^2} \tag{6.9}$$

3. *The probability of wait time at node n follows an exponential distribution with rate $\mathfrak{F}$. Therefore, the expected wait time at stop n is given by $EW_n = \frac{1}{\mathfrak{F}}$*

Proof. The probability of choosing line $i \in FS(n)$ is equal to the probability of waiting time for line $i \in FS(n)$ to be less than or equal to waiting time of other lines $j \neq i$, i.e.,

$$P_i = Prob(w_i \leq \min_{j \neq i} w_j) = \int_0^\infty \mathfrak{g}_i(w) \Pi_{j \neq i} Prob(w_j \geq w) dw = \int_0^\infty \gamma_i(w) dw \tag{6.10}$$

where, $\gamma_i(w) = g_i(w)\Pi_{j\neq i} Prob(w_j \geq w) = f_i e^{-f_i w}\Pi_{j\neq i} e^{-f_j w} = f_i e^{-(\sum_j f_j)w}$. The value of $\gamma_i(w)$ can be interpreted as the probability density function of the waiting time at the stop n conditional to boarding line i. Using (6.10), the probability of choosing line $i \in FS(n)$ can be evaluated as:

$$P_i = \int_0^\infty f_i e^{-(\sum_j f_j)w} dw = \frac{f_i}{\mathfrak{F}} \tag{6.11}$$

The expected wait time conditional to boarding line i is:

$$EW_i = \int_0^\infty w\gamma_i(w)dw = \int_0^\infty w f_i e^{-(\sum_j f_j)w} dw = \frac{f_i}{\mathfrak{F}^2} \tag{6.12}$$

Summing over all the lines $FS(n)$ gives us the expected wait time at the stop, i.e.,

$$EW_n = \sum_{i\in FS(n)} \int_0^\infty w\gamma_i(w)dw = \int_0^\infty w \sum_i \gamma_i(w)dw \tag{6.13}$$

where, $\sum_i \gamma_i(w)$ is the probability density function of the waiting time at stop n.

$$\sum_{i\in FS(n)} \gamma_i(w) = \sum_{i\in FS(n)} f_i e^{-(\sum_j f_j)w} = \mathfrak{F} e^{-\mathfrak{F}w}, w \geq 0 \tag{6.14}$$

Therefore, the expected wait time at stop n if given by $EW_n = \frac{1}{\mathfrak{F}}$. □

Combined MoD and transit service wait time

Before discussing the computation of the expected wait time involving both modes, we need to make an assumption about the wait time distribution of MoD service by utilizing the value of the average wait time of MoD service calculated in equation (6.5).

Assumption 6. *The wait time distribution of MoD service for passengers waiting at node n follows an exponential distribution with rate* $f_{MoD} = \mathcal{A}_{Z(n)}\mathcal{V}_{Z(n)}$*, where* $Z(n)$ *is the zone associated to node n and* $\mathcal{V}_{Z(n)}$ *is the number of vehicles deployed in zone* $Z(n)$*.*

A passenger waiting at the head node of an access link faces the choice between MoD or transit mode. This is because the wait time of both services can vary based on the frequency provided, and a passenger will include one or both modes in their strategy to reduce the overall expected cost. This assumption simplifies the operation of MoD service as a transit service available at any stop of the network. The following proposition evaluates the expected wait time of that passenger.

Proposition 8. *Given that the waiting time for transit and MoD mode follow an exponential distribution with rate* $\mathfrak{F}$ *and* f_{MoD} *respectively and* $\mathbb{F} = \mathfrak{F} + f_{MoD}$ *, the following holds:*

1. *The probabilities of taking transit and MoD are given by* $P_{MoD} = \frac{f_{MoD}}{\mathbb{F}}$ *and* $P_{transit} = \frac{\mathfrak{F}}{\mathbb{F}}$ *respectively.*

2. *The expected wait time of the passenger departing from an access node* n *served by both MoD and transit service is given by* $EW_n = \frac{1}{\mathbb{F}}$

Proof. The probability of taking transit is given by:

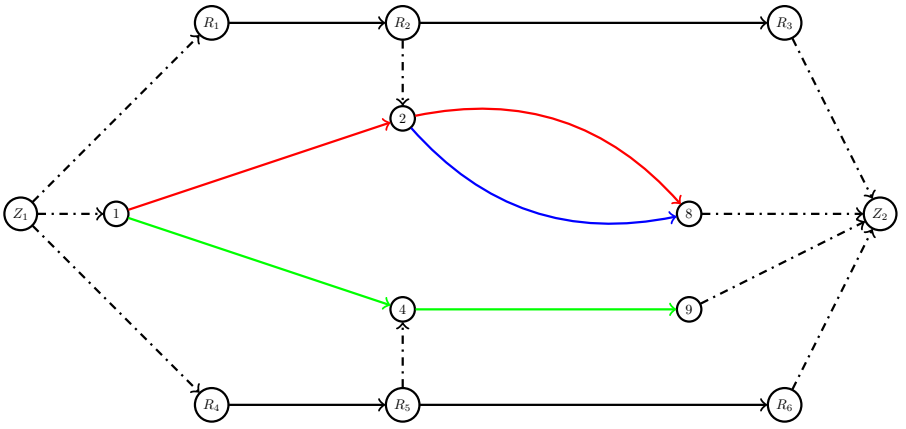
$$P_{transit} = \int_0^\infty (\sum_i \gamma_i(w)) Prob(w \leq w_{MoD}) dw \tag{6.15}$$

$$P_{transit} = \int_0^\infty \mathfrak{F} e^{-\mathfrak{F}w} \times e^{-(f_{MoD})w} dw = \frac{\mathfrak{F}}{\mathbb{F}} \tag{6.16}$$

Similarly, the probability of taking MoD is given by $P_{MoD} = \frac{\mathfrak{F}}{\mathbb{F}}$. The expected wait time of the passenger departing from an access node n is given by:

$$EW_n = \int_0^\infty w \left(\mathfrak{F} e^{-(f_{MoD}+\mathfrak{F})} + f_{MoD} e^{-(f_{MoD}+\mathfrak{F})} \right) dw = \frac{1}{\mathbb{F}} \tag{6.17}$$

□



Line	Frequency
Red	1/6
Green	1/2
Blue	1/3

Zone	Vehicles
Z_1	100
Z_2	50

Figure 6.1: An illustrative example of an integrated MoD and transit network

To get more insights into the wait time computation, let us consider an example. Figure 6.1(a) shows an illustration of a multimodal transportation network. It consists of 2 zones, 6 nodes and 8 links as part of the road network, and 5 nodes and 10 links as part of the transit network. The transit network has 3 transit lines (color-coded) whose frequencies are shown in Figure 6.1(b). There are 100 and 50 vehicles deployed in zone 1 and 2 respectively. By using Prop 7, we can evaluate the probability of passengers taking various transit lines in the network. For example, the probabilities of choosing red line and green line at stop 1 are $\frac{1/6}{1/6+1/2} = 0.25$ and $\frac{1/2}{1/6+1/2} = 0.75$ respectively. The expected wait time at stop 1 is equal to $12/8 = 1.5$ minutes. Similarly, using Proposition 8, the probabilities of choosing MoD and transit at Z_1 are $\frac{0.0017*100}{0.0017*100+8/12} = 0.2$ and 0.8 respectively (assuming $\mathcal{A}_1 = 0.0017$). The overall expected wait time at node Z_1 is 1.19

minutes which is less than 1.5 minutes by only considering transit service as part of the strategy.

We further use Proposition 7 and 8 to formulate the multimodal passenger assignment model. For this purpose, we extend the frequency-based transit assignment model proposed by [7] to a multimodal transportation system. Before moving forward, we must make the following assumptions:

Assumption 7. *(a) Ridepooling is not allowed, i.e., the MoD service serves one passenger at a time.*

(b) The transit lines are assumed to have unlimited capacity.

(c) Passengers want to reduce their expected generalized travel cost consisting of travel time, wait time, and fare to get to their destination.

The ridepooling problem requires matching of customers using a specific algorithm. This is an important aspect to accurately estimate the cost of day-to-day operations. Nevertheless, ignoring ridepooling will give us an upper bound on the number of vehicles required to serve various zones. The modeling of passenger behavior while incorporating the capacity constraints (congestion) is a difficult problem. The congestion is important to consider since it causes denied boarding, which leads to increased waiting time, travel time, and discomfort. Several authors have tried to include congestion into frequency-based transit assignment models through various approaches, namely, discomfort function [7], effective frequency [36, 38, 140, 141], and failure-to-board probabilities [39]. Despite the effort, there is no tractable closed-form of congested frequency-based transit assignment model. On the other hand, it would not be ideal to include transit vehicle capacity constraints into the assignment program (e.g., in [1]) because doing so may lead to unrealistic passenger behavior, which previous studies on congested frequency-based

transit assignments were trying to avoid. Therefore, we use an uncapaciated assignment for the design problem. Assumption 7(c) is a common in the assignment literature. To proceed further, let us define a variable $\{g_{ik}\}_{i\in N,k\in D}$ as below:

$$g_{ik} = \begin{cases} d^{ik}, & \text{if } i \neq k, (i,k) \in O \times D \\ -\sum_{o\in O} d^{ok}, & \text{if } i = k \\ 0, & \text{otherwise} \end{cases}$$

Furthermore, let us denote v_{ak} and W_{ik} as the flow of passengers on link $a \in A$ and waiting at node $i \in N^w$ resp. destined to $k \in D$. The assignment optimization program is presented below:

$$\underset{v,W}{\text{minimize}} \quad \sum_{k\in D}\left(\sum_{a\in E} c_a v_{ak} + \sum_{i\in N^w} W_{ik}\right) \tag{6.18a}$$

$$\text{subject to} \quad \sum_{a\in FS(i)} v_{ak} = \sum_{a\in BS(i)} v_{ak} + g_{ik}, \forall i \in V, \forall k \in D \tag{6.18b}$$

$$v_{ak} \leq f_a W_{ik}, \forall a \in FS(i) : a \in A, \forall i \in N^w, \forall k \in D \tag{6.18c}$$

$$v_{ak} \leq \mathcal{A}_{Z(i)}\mathcal{V}_{Z(i)} W_{ik}, \forall a \in FS(i) : a \in A_R, \forall i \in N^w, \forall k \in D \tag{6.18d}$$

$$v_{ak} \geq 0, \forall a \in E, \forall k \in D \tag{6.18e}$$

The assignment program (6.18) minimizes the total expected link costs and wait time at waiting nodes experienced by the passengers in a mulimodal network subject to the flow conservation constraint at each node (6.18b), flow proportion constraints (6.18c)-(6.18d), and the non-negativity and binary constraints (6.18e). The flow proportion constraints uses the probability of selecting an option $a \in FS(i)$ (if that option is a part of the strategy of the passengers traveling to destination $k \in D$) and multiplies it with the number of passengers waiting at that node. Note that the probability of selecting

an option (MoD or transit line) is calculated in Proposition 8.

6.2 Design of an integrated MoD and transit system

In this section, we present an optimization model incorporating the assignment program proposed in previous section for the design of an integrated MoD and transit system. The optimization program is formalized as a Mixed Integer Non-linear Program (MINLP). In this model, we determine which transit routes to keep operating among the current transit routes in the city network, decide the optimal frequency of those operating routes, and finally, determine the fleet size of vehicles required to provide MoD service in various zones. Note that one can also include new candidate transit routes as part of the design plan. The sets, parameters, and decision variables for the optimization model are summarized in Table 6.1.

The design of an integrated transit and MoD system should consider both passenger and operator perspectives. The operator's perspective is to provide the service at minimum cost, and the passengers' perspective is to minimize the overall cost of travel (including travel time, wait time, and fare). Based on these perspectives, the design optimization model is presented as (6.19). The objective function is the sum of the total expected travel cost and wait time experienced by the passengers in the network. The mapping $\mathcal{B} : L \times \Theta \mapsto \mathbb{N}$ used in (6.19b) is defined as $\mathcal{B}(l, f) = \left(f \times \sum_{a \in A^l} 2t_a\right)$, which describes that the number of buses required to provide frequency $f \in \Theta$ for a line $l \in L$ is equal to the product of the frequency and round trip travel time. (6.19b) constrain the total number of buses needed to be less than or equal to $\bar{B}$, which can be evaluated for a given budget. (6.19c) describes the flow conservation constraints at every node

Table 6.1: Sets, decision variables and parameters used in the integrated network design model

		Sets
$\mathfrak{B}$	$\triangleq$	Set of binary values
L	$\triangleq$	Set of candidate transit lines
$\Theta = \{2, 3, 4, 6, 12\}$	$\triangleq$	Set of possible frequencies of a line (buses/hr)
$\Omega = \{0.01, 50, 100, 200, 500\}$	$\triangleq$	Set of possible number of vehicles deployed in a zone
		Parameters
$\bar{B}$	$\triangleq$	Total number of buses available
$\bar{F}$	$\triangleq$	Total number of vehicles available
		Decision Variables
x_l	$=$	$\begin{cases} 1, & \text{if line } l \in L \text{ is kept operating} \\ 0, & \text{otherwise} \end{cases}$
y_{lf}	$=$	$\begin{cases} 1, & \text{if frequency } f \in \Theta \text{ is adopted for line } l \in L \\ 0, & \text{otherwise} \end{cases}$
$\mathcal{N}_{zn}$	$=$	$\begin{cases} 1, & \text{if a fleet of size } n \in \Omega \text{ is deployed in zone } z \in Z \\ 0, & \text{otherwise} \end{cases}$
v_{ak}	$=$	Flow of passengers on link $a \in E$ destined to $k \in D$
W_{ik}	$=$	Wait time of passengers waiting at node $i \in N^w$ destined to $k \in D$

for every destination. For a given MoD and bus fleet assignment, (6.19d)-(6.19e) describe the passenger flow on each link based on the frequency of the bus route and MoD service. A frequency value can be assigned to a route if that route is kept operating as constrained by (6.19f). (6.19g) describe that exactly one of the fleet sizes can be adopted for each zone. (6.19h) constrain the required number of vehicles to be less than or equal to $\bar{F}$. Finally, (6.19i), (6.19j) and (6.19k)-(6.19m) are the non-negativity constraints of the flow, wait time being free variables, and binary constraints of design variables respectively. One can also incorporate other constraints related to the budget of operating MoD and transit service but for the sake of simplicity, we do not include them here.

$$\underset{v,W,x,y,\mathcal{N}}{\text{minimize}} \quad \sum_{k\in D}\left(\sum_{a\in E} c_a v_{ak} + \sum_{i\in N^w} W_{ik}\right) \tag{6.19a}$$

$$\text{subject to} \quad \sum_{l\in L}\sum_{f\in\Theta} \mathcal{B}(l,f)\times y_{lf} \leq \bar{B} \tag{6.19b}$$

$$\sum_{a\in FS(i)} v_{ak} = \sum_{a\in BS(i)} v_{ak} + g_{ik}, \forall i \in V, \forall k \in D \tag{6.19c}$$

$$v_{ak} \leq \left(\sum_{f\in\Theta} f y_{l(a)f}\right) W_{ik}, \forall a \in FS(i) : a \in A, \forall i \in N^w, \forall k \in D \tag{6.19d}$$

$$v_{ak} \leq \mathcal{A}_{Z(i)}\left(\sum_{n\in\Omega} n\mathcal{N}_{Z(i)n}\right) W_{ik}, \forall a \in FS(i) : a \in A_R, \forall i \in N^w, \forall k \in D \tag{6.19e}$$

$$\sum_{f\in\Theta} y_{lf} = x_l, \forall l \in L \tag{6.19f}$$

$$\sum_{n\in\Omega} \mathcal{N}_{zn} = 1, \forall z \in Z \tag{6.19g}$$

$$\sum_{z\in Z}\left(\sum_{n\in\Omega} n\mathcal{N}_{zn}\right) \leq \bar{F} \tag{6.19h}$$

$$v_{ak} \geq 0, \forall a \in E, \forall k \in D \tag{6.19i}$$

$$W_{ik} \text{ free }, \forall i \in N^w, \forall k \in D \tag{6.19j}$$

$$x_l \in \mathfrak{B}, \forall l \in L \tag{6.19k}$$

$$y_{lf} \in \mathfrak{B}, \forall f \in \Theta, \forall l \in L \tag{6.19l}$$

$$\mathcal{N}_{zn} \in \mathfrak{B}, \forall n \in \Omega, z \in Z \tag{6.19m}$$

The optimization program (6.19) is a mixed-integer non-linear program (MINLP). The non-linearity arise from the constraints (6.19d)-(6.19e). It is computationally difficult to solve this program for large instances, which can be attributed to the integer constraints (6.19k)-(6.19m) and the bilinear constraints (6.19d)-(6.19e). The bilinear constraints are particularly difficult to handle due to the non-convex nature

even if the integrality constraints of the involved variables are relaxed. Fortunately, in this case, the non-convexity arises due to the product of continuous and binary variables, which can be exactly relaxed by employing McCormick relaxations. Let $t_{faik} = y_{l(a)f}W_{ik}, \forall f \in \Theta, \forall a \in FS(i) : a \in A, \forall i \in N^w, \forall k \in D$ and $\omega_{ink} = \mathcal{N}_{Z(i)n}W_{ik}, \forall n \in \Omega, \forall i \in N^w \cap N_R, \forall k \in D$. Further, let us assume that there exists a finite upper and lower bound on the variable W_{ik} , i.e., $\underline{W}_{ik} \leq W_{ik} \leq \overline{W}_{ik}$. Then, t_{faik} and ω_{ink} can be expressed as the set of linear constraints (6.20a)-(6.20d) and (6.21a)-(6.21d) respectively:

$$\overline{W}_{ik} - W_{ik} + t_{faik} - \overline{W}_{ik}y_{l(a)f} \geq 0 \tag{6.20a}$$

$$\overline{W}_{ik}y_{l(a)f} - t_{faik} \geq 0 \tag{6.20b}$$

$$t_{faik} - \underline{W}_{ik}y_{l(a)f} \geq 0 \tag{6.20c}$$

$$W_{ik} - \underline{W}_{ik} - t_{faik} + \underline{W}_{ik}y_{l(a)f} \geq 0 \tag{6.20d}$$

$$\overline{W}_{ik} - W_{ik} + \omega_{ink} - \overline{W}_{ik}\mathcal{N}_{Z(i)n} \geq 0 \tag{6.21a}$$

$$\overline{W}_{ik}\mathcal{N}_{Z(i)n} - \omega_{ink} \geq 0 \tag{6.21b}$$

$$\omega_{ink} - \underline{W}_{ik}\mathcal{N}_{Z(i)n} \geq 0 \tag{6.21c}$$

$$W_{ik} - \underline{W}_{ik} - \omega_{ink} + \underline{W}_{ik}\mathcal{N}_{Z(i)n} \geq 0 \tag{6.21d}$$

6.3 Solution methodology

After reformulating the bilinear constraints (6.19d)-(6.19e), the resulting model is a Mixed Integer Linear Program (MILP). The program is still difficult to solve efficiently for large instances. However, the structure of the problem allows us to use decomposition techniques such as Benders decomposition to efficiently solve it. In this section, we present the details of the Benders reformulation for this problem, along with the

proposed algorithmic enhancements.

6.3.1 Benders Reformulation

Benders decomposition [161] is an elegant way of solving a large scale MILP by iteratively solving two simpler subproblems: the relaxed master problem (RMP), which is a relaxation of the original problem and a subproblem (SP) which provides inequalities/cuts to strengthen the RMP. The subproblem should possess strong duality properties. Let us consider the network design problem described in the previous section. For a given feasible value of design decision variables $\hat{x}, \hat{y}, \hat{\mathcal{N}}$ and with $\underline{W}_{ik} = 0$ (wait time cannot be negative), we can rewrite the original problem as a *Benders subproblem* (6.22).

$$z^{SP}(\hat{x},\hat{y},\hat{\mathcal{N}}) = \min_{v,W,\omega,t} \quad \sum_{k\in D}\left(\sum_{a\in A} c_a v_{ak} + \sum_{i\in N^w} W_{ik}\right) \tag{6.22a}$$

subject to

$$\sum_{a\in FS(i)} v_{ak} = \sum_{a\in BS(i)} v_{ak} + g_{ik}, \forall i \in V, \forall k \in D \tag{6.22b}$$

$$v_{ak} \leq \left(\sum_{f\in\Theta} ft_{faik}\right), \forall a \in FS(i) : a \in A, \forall i \in N^w, \forall k \in D \tag{6.22c}$$

$$W_{ik} - t_{faik} \leq \overline{W}_{ik}(1-\hat{y}_{l(a)f}), \forall f \in \Theta, \forall a \in FS(i) : a \in A, \forall i \in N^w, \forall k \in D \tag{6.22d}$$

$$t_{faik} \leq \overline{W}_{ik}\hat{y}_{l(a)f}, \forall f \in \Theta, \forall a \in FS(i) : a \in A, \forall i \in N^w, \forall k \in D \tag{6.22e}$$

$$W_{ik} - t_{faik} \geq 0, \forall f \in \Theta, \forall a \in FS(i) : a \in A, \forall i \in N^w, \forall k \in D \tag{6.22f}$$

$$v_{ak} \leq \mathcal{A}_{Z(i)}\left(\sum_{n\in\Omega} n\omega_{ink}\right), \forall a \in FS(i) : a \in A_R, \forall i \in N^w, \forall k \in D \tag{6.22g}$$

$$W_{ik} - \omega_{ink} \leq \overline{W}_{ik}(1-\hat{\mathcal{N}}_{Z(i)n}), \forall n \in \Omega, \forall i \in N^w \cap N_R, \forall k \in D \tag{6.22h}$$

$$\omega_{ink} \leq \overline{W}_{ik}\hat{\mathcal{N}}_{Z(i)n}, \forall n \in \Omega, \forall i \in N^w \cap N_R, \forall k \in D \tag{6.22i}$$

$$W_{ik} - \omega_{ink} \geq 0, \forall n \in \Omega, \forall i \in N^w \cap N_R, \forall k \in D \tag{6.22j}$$

$$v_{ak} \geq 0, \forall a \in E, \forall k \in D \tag{6.22k}$$

$$t_{faik} \geq 0, \forall f \in \Theta, \forall a \in FS(i) : a \in A, \forall i \in N^w, \forall k \in D \tag{6.22l}$$

$$\omega_{ink} \geq 0, \forall n \in \Omega, \forall i \in N^w \cap N_R, \forall k \in D \tag{6.22m}$$

Let $\mathcal{X}^{SP} = \{(v, W, \omega, t) : (6.22\text{b}) - (6.22\text{m})\}$ be the feasible region of the Benders subproblem (6.22). Further, let us denote $\mathcal{X}^{MA} = \{(v, W) : (6.18\text{b}) - (6.18\text{e})\}$ as the feasible region of the multimodal assignment linear program. We can show the following result:

Proposition 9. *The projection of the feasible region of the subproblem* (6.22) *on to the space of* v *and* W *is same as the feasible region of the multimodal assignment problem* (6.18) *i.e.,*

$$proj_{v,W}\mathcal{X}^{SP} = \mathcal{X}^{MA}$$

Proof. To prove this, we need to show that (a) $proj_{v,W}\mathcal{X}^{SP} \subseteq \mathcal{X}^{MA}$ and (b) $\mathcal{X}^{MA} \subseteq$

$proj_{v,W}\mathcal{X}^{SP}$. Let us first start by proving (b). Let $(v, W) \in \mathcal{X}^{MA}$. For all $a \in FS(i) : a \in A, \forall i \in N^w, \forall k \in D$, we have $v_{ak} \leq f_a W_{ik}$. Let $y_{l(a)f} = 1$ if the frequency of the line associated to arc a is $f \in \Theta$ and 0, otherwise. Let $t_{faik} = \hat{y}_{l(a)f} W_{ik}$, then $f_a W_{ik} = \sum_{f\in\Theta} f\hat{y}_{l(a)f} W_{ik} = \sum_{f\in\Theta} f t_{faik}$, which is same as (6.22c). Also, $t_{faik} = \hat{y}_{l(a)f} W_{ik}$ can be expressed as (6.22d)- (6.22f) and (6.22l). Using a similar argument, we can show that for all $a \in FS(i) : a \in A_R, \forall i \in N^w, \forall k \in D$, the inequality $v_{ak} \leq f_a W_{ik}$ can be expressed as (6.22g)-(6.22j) and (6.22m). This shows that $\mathcal{X}^{MA} \subseteq proj_{v,W}\mathcal{X}^{SP}$. To prove part (a), let $(v, W, \omega, t) \in \mathcal{X}^{SP}$, then using Fourier-Motzkin elimination, we have $(v, W) \in \mathcal{X}^{MA}$ ([162, Chapter 3]). □

Proposition 9 shows that one can use the efficient [7]'s primal-dual algorithm designed for the transit assignment problem to solve the current Benders subproblem. To speed up the process of Benders decomposition by avoiding feasibility cuts, we need to put some restrictions so that the subproblem (6.22) is always feasible.

Proposition 10. *Given that* $0 \notin \Omega$ *and* $G_R(N_R, A_R)$ *is connected, then* $\mathcal{X}^{SP}$ *is non-empty for any given feasible value of* $\hat{x}, \hat{y}, \hat{\mathcal{N}}$.

Proof. The set $\mathcal{X}^{SP}$ can be empty in two cases i.e., when there is no flow balance ($\sum_{k\in D}\sum_{i\in N} g_{ik} \neq 0$) or there does not exist a directed path from any node $i \in N$ to any destination k. However, it is not possible to have any of these cases because from the definition of g_{ik}, we have $\sum_{k\in D}\sum_{i\in N} g_{ik} = 0$ and since there is always at least 0.01 vehicle assigned to all the zones and the road network is connected, there always exists a path from any node $i \in N$ to any destination $k \in D$. Therefore, $\mathcal{X}^{SP} \neq \phi$.

□

Proposition 10 makes the Benders subproblem feasible for any feasible value of $\hat{x}, \hat{y}, \hat{\mathcal{N}}$. This is an important result to make the Benders decomposition implementation faster.

Let $\{\mu_{ik}\}$, $\{\lambda^1_{aik}\}$, $\{\lambda^2_{faik}\}$, $\{\lambda^3_{faik}\}$, $\{\lambda^4_{faik}\}$, $\{\lambda^5_{aik}\}$, $\{\lambda^6_{nik}\}$, $\{\lambda^7_{nik}\}$, and $\{\lambda^8_{nik}\}$ be the dual variables associated with the constraints (6.22b) -(6.22j) respectively. Then, the dual of the subproblem DSP can be stated as below:

$$z^{DSP}(\hat{x},\hat{y},\hat{\mathcal{N}}) = \max_{\mu,\lambda} \quad \sum_{k\in D}\Bigg[\sum_{i\in N}\mu_{ik}g_{ik} + \sum_{i\in N^w}\sum_{a\in FS(i):a\in A}\sum_{f\in\Theta}\big(\overline{W}_{ik}(1-\hat{y}_{l(a)f})\lambda^2_{faik} + \overline{W}_{ik}\hat{y}_{l(a)f}\lambda^3_{faik}\big) + \sum_{i\in N^w\cap N_R}\sum_{n\in\Omega}\Big(\overline{W}_{ik}(1-\hat{\mathcal{N}}_{Z(i)n})\lambda^6_{nik} + \overline{W}_{ik}\hat{\mathcal{N}}_{Z(i)n}\lambda^7_{nik}\Big)\Bigg] \tag{6.23a}$$

subject to

$$\mu_{ik} - \mu_{jk} + \lambda^1_{aik} + \lambda^5_{aik} \le c_a, \forall a = (i,j) \in E, \forall k \in D \tag{6.23b}$$

$$-\sum_{f\in\theta}\sum_{\substack{a\in FS(i):\\ a\in A}}\big(\lambda^2_{faik} + \lambda^4_{faik}\big) + \sum_{n\in\Omega}\big(\lambda^6_{nik} + \lambda^8_{nik}\big) = 1, \forall, \forall i \in N^w, \forall k \in D \tag{6.23c}$$

$$-f\lambda^1_{aik} - \lambda^2_{faik} + \lambda^3_{faik} - \lambda^4_{faik} \le 0, \forall f \in \Theta, \forall a \in FS(i) : a \in A, \quad \forall i \in N^w, \forall k \in D \tag{6.23d}$$

$$-n\lambda^5_{aik} - \lambda^6_{nik} + \lambda^7_{nik} - \lambda^8_{nik} \le 0, \forall n \in \Omega, \forall i \in N^w \cap N_R, \forall k \in D \tag{6.23e}$$

$$\lambda^1_{aik}, \lambda^5_{aik} \le 0, \forall a \in FS(i), \forall i \in N^w, \forall k \in D \tag{6.23f}$$

$$\lambda^2_{faik}, \lambda^3_{faik} \le 0, \lambda^4_{faik} \ge 0, \forall f \in \Theta, \forall a \in FS(i) : a \in A, \forall i \in N^w, \forall k \in D \tag{6.23g}$$

$$\lambda^6_{nik}, \lambda^7_{nik} \le 0, \lambda^8_{nik} \ge 0, \forall n \in \Omega, \forall i \in N^w \cap N_R, \forall k \in D \tag{6.23h}$$

Let us denote the feasible region of DSP as $\Pi = \{(\mu, \lambda^1, \lambda^2, \lambda^3, \lambda^4, \lambda^5, \lambda^6, \lambda^7, \lambda^8) : (6.23b) - (6.23h)\}$. Note that Π does not depend on the value of $x, y, \mathcal{N}$. From Proposition 10, we know that SP is always feasible for any given feasible value of $(\hat{x}, \hat{y}, \hat{\mathcal{N}})$, then by linear programming duality, DSP should be bounded. The implication is that the polyhedron describing Π is bounded and can be described as the convex hull of a set of extreme points only (from Minkowski-Weyl's theorem on characterization of polyhedra ([162, Chapter 3])). Let $\{(\mu^\pi, (\lambda^1)^\pi, (\lambda^2)^\pi, (\lambda^3)^\pi, (\lambda^4)^\pi, (\lambda^5)^\pi, (\lambda^6)^\pi, (\lambda^7)^\pi, (\lambda^8)^\pi)\}_{\pi\in\mathcal{K}}$

be the set of extreme points of polytope Π, where $\mathcal{K}$ represents the set of indices of extreme points. By applying an outer linearization procedure to the inner (sub) problem of the original problem, we can restate it as (6.24), which is referred to as the *Benders Master problem* (MP).

Theorem 1. *[144] The problem* (6.19) *can be reformulated as below:*

$$\underset{x,y,\mathcal{N},\eta}{\textit{minimize}} \quad \eta \tag{6.24a}$$

$$\textit{subject to} \quad \sum_{l\in L}\sum_{f\in\Theta}\mathcal{B}(l,f)\times y_{lf} \leq \bar{B} \tag{6.24b}$$

$$\sum_{f\in\Theta} y_{lf} = x_l, \forall l \in L \tag{6.24c}$$

$$\sum_{n\in\Omega}\mathcal{N}_{zn} = 1, \forall z \in Z \tag{6.24d}$$

$$\sum_{z\in Z}\left(\sum_{n\in\Omega} n\mathcal{N}_{zn}\right) \leq \bar{F} \tag{6.24e}$$

$$\eta \geq \sum_{k\in D}\Bigg[\sum_{i\in N}(\mu_{ik})^{\pi}g_{ik} + \sum_{i\in N^w}\sum_{a\in FS(i):a\in E}\sum_{f\in\Theta}\big(\overline{W}_{ik}(1-\hat{y}_{l(a)f})(\lambda^2_{faik})^{\pi} + \overline{W}_{ik}\hat{y}_{l(a)f}(\lambda^3_{faik})^{\pi}\big) + \sum_{i\in N^w\cap N_R}\sum_{n\in\Omega}\Big(\overline{W}_{ik}(1-\hat{\mathcal{N}}_{Z(i)n})(\lambda^6_{nik})^{\pi} + \overline{W}_{ik}\hat{\mathcal{N}}_{Z(i)n}(\lambda^7_{nik})^{\pi}\Big)\Bigg], \forall \pi \in \mathcal{K} \tag{6.24f}$$

$$x_l \in \mathfrak{B}, \forall l \in L \tag{6.24g}$$

$$y_{lf} \in \mathfrak{B}, \forall f \in \Theta, \forall l \in L \tag{6.24h}$$

$$\mathcal{N}_{in} \in \mathfrak{B}, \forall n \in \Omega, i \in Z \tag{6.24i}$$

Proof. See [144]. □

6.3.2 Classic Benders decomposition implementation

The issue with the Benders reformulation is that there could be a large number of extreme points of the polyhedron associated with the feasible region of DSP, therefore, one applies an iterative process of solving two problems, namely, the relaxed master problem (RMP) and the subproblem (SP) repeatedly. The relaxed master problem is the master problem with constraints (6.24f) being defined only for a subset of extreme points, i.e., $\mathcal{K}^{'} \subset \mathcal{K}$. The overall implementation of the classic Benders Decomposition is summarized in Algorithm 8. We start by finding the feasible value of $(x^0, y^0, \mathcal{N}^0)$. This can be done by solving (6.24) without (6.24f) and including a constraint $\eta \geq 0$. Then, in each iteration t, the algorithm solves RMP with the given set of extreme points and then SP with the current value of $(x^t, y^t, \mathcal{N}^t)$. Since RMP is relaxation and SP is solved for a feasible value $(x^t, y^t, \mathcal{N}^t)$, they provide a lower bound and upper bound respectively to the original problem. The subproblem also provides inequalities (optimality cuts) to strengthen the formulation of RMP in each iteration. Thus, it is guaranteed to have non-decreasing lower bounds. In our case, there are no feasibility cuts since our subproblem is always feasible (Proposition 10). The algorithm terminates when both the upper bound and lower bound are close to each other.

Algorithm 8 Classic Benders decomposition implementation

1: (*Initialize*) Let $t = 0, UB = -\infty, LB = \infty, \mathcal{K}^{'} = \phi$. Assume an initial feasible value $(x^0, y^0, \mathcal{N}^0)$. Solve the SP (6.22), obtain the optimal dual solution and append that to set $\mathcal{K}^{'}$.
2: **while** $UB - LB > \epsilon$ **do** ▷ ϵ is the tolerance parameter
3: Set $t = t + 1$. Solve RMP (6.24) and obtain its optimal solution $(x^t, y^t, \mathcal{N}^t)$.
4: Set LB $= \eta$
5: Solve SP (6.22) for $(x^t, y^t, \mathcal{N}^t)$, obtain dual solutions and append that to $\mathcal{K}^{'}$.
6: Set $UB = \sum_{k \in D} \left(\sum_{a \in A} c_a v_{ak}^t + \sum_{i \in N^w} W_{ik}^t \right)$

6.3.3 Enhanced Benders decomposition implementation

The classic Benders decomposition may take prohibitive computational effort to converge, thus making it difficult to solve the problem for large instances. The slow convergence can be attributed to the low strength of the optimality cuts, degeneracy in the subproblem, no guarantee of non-decreasing upper bounds in each iteration, or not formulating the problem "properly" [163–165]. To accelerate the Benders decomposition algorithm, we make use of several enhancements that are described below:

Use of multiple cuts via disaggregated cuts

For this design problem, we can further utilize the decomposable structure of the Benders subproblem (6.19) as it is decomposable for each destination $k \in D$. That is, we can solve several (smaller) subproblems and generate multiple optimality cuts for the master problem. The disaggregated cuts have a higher probability of finding facet-defining inequalities characterizing Π. For this purpose, we modify RMP to allow for the disaggregated cuts as (6.25):

$$\underset{x,y,\mathcal{N},\eta}{\text{minimize}} \quad \sum_{k\in D} \eta_k \tag{6.25a}$$

$$\text{subject to} \quad (6.24\text{b}) - (6.24\text{e}) \tag{6.25b}$$

$$\eta_k \geq \Bigg[\sum_{i\in N} (\mu_{ik})^\pi g_{ik} + \sum_{i\in N^w} \sum_{a\in FS(i):a\in A} \sum_{f\in\Theta} \big(\overline{W}_{ik}(1-\hat{y}_{l(a)f})(\lambda^2_{faik})^\pi + \overline{W}_{ik}\hat{y}_{l(a)f}(\lambda^3_{faik})^\pi \big) + \sum_{i\in N^w\cap N_R} \sum_{n\in\Omega} \Big(\overline{W}_{ik}(1-\hat{\mathcal{N}}_{Z(i)n})(\lambda^6_{nik})^\pi + \overline{W}_{ik}\hat{\mathcal{N}}_{Z(i)n}(\lambda^7_{nik})^\pi \Big) \Bigg], \forall \pi \in \mathcal{K}_k, \forall k \in D \tag{6.25c}$$

$$(6.24\text{g}) - (6.24\text{i}) \tag{6.25d}$$

Note that by adding the disaggregated cuts for every destination, we can get back

the optimality cuts defined in the classic Benders relaxed master problem.

Use of multiple cuts via multiple solutions

To further improve the convergence of the algorithm, [166] used a strategy known as multiple cuts via multiple solutions. When solving the RMP, any commercial solver such as AIMMS or GUROBI can be asked to generate multiple solutions of an integer program (optimal as well as suboptimal) by using `pool solution option`. These multiple solutions can be used to generate multiple classic (6.24f) or disaggregated cuts (6.25c) to be added in next iteration of RMP. This strategy is expected to decrease the overall iterations and possibly the solution time of the algorithm.

Use of clique/cover cuts

Due to the limited availability of bus and vehicle fleet, one can use the clique/cover cuts to tighten the feasible region of the master problem.

Proposition 11. *For every $n \in \Omega$, if $\lfloor \frac{\bar{F}}{n} \rfloor < |Z|$ then the clique inequality $\sum_{z \in Z} \mathcal{N}_{zn} \leq \lfloor \frac{\bar{F}}{n} \rfloor$ is valid for* (6.19).

Proof. Due to limited fleet available, one can allocate $n \in \Omega$ vehicles in at most $\lfloor \frac{\bar{F}}{n} \rfloor$ zones. □

If for any $n \in \Omega$, we have $\lfloor \frac{\bar{F}}{n} \rfloor > |Z|$, then the inequality $\sum_{z \in Z} \mathcal{N}_{zn} \leq \lfloor \frac{\bar{F}}{n} \rfloor$ will be redundant and therefore, we do not add it to the model.

The inequality which constrain the number of buses (6.19b) is a Knapsack constraint. A set $C \subseteq L \times \Theta$ is a *cover* for inequality (6.19b) if $\sum_{(l,f) \in C} \mathcal{B}(l,f) > \bar{B}$ and it is *minimal cover* if $\sum_{(l,f) \in C \setminus \{(l',f')\}} \mathcal{B}(l,f) \leq \bar{B}$, for all $(l', f') \in C$

Proposition 12. *For any minimal cover* $C \subseteq L \times \Theta$*, the inequality* $\sum_{(l,f)\in C} y_{lf} \leq |C|-1$ *is valid for* (6.19)*.*

Proof. The proposition follows from the definition of minimal cover that we cannot provide the number of buses required in the minimal cover. □

To generate some of the minimal cover cuts, one can use the heuristic given in Algorithm 9. In this algorithm, for each frequency $f \in \Theta$, we keep the list of lines G for which the number of buses required to provide the frequency f does not exceed $\bar{B}$. Then, any line which is not in G, along with G forms a minimal cover.

Algorithm 9 Cover cut generation heuristic

```
procedure
    Compute the value of mapping B(l, f) for all (l, f) ∈ L × Θ.
    CC ← φ
    for f ∈ Θ do
        G ← []; temp ← 0
        for l ∈ L : B(l, f) in an ascending order do
            temp = temp + B(l, f)
            if temp ≤ B̄ then
                append (l, f) to G
            else
                break
        for l ∈ L\G do
            C ← G ∪ {(l, f)}
            append C to CC
return CC
```

Furthermore, one can use other efficient techniques to produce maximal clique or minimal cover cuts for the problem.

Other recommendations

One of the problems with the Benders subproblem (6.22) is that it assumes the value of $\overline{W}_{ik}$ as a given upper bound. The value of $\overline{W}_{ik}$ is a big-M introduced to relax the

non-linearity in the original model. If the value of the big-M is not chosen properly, then one can face serious issues with the convergence of the algorithm. For example, choosing $\overline{W}_{ik} < W_{ik}$ can make the subproblem infeasible, and choosing $\overline{W}_{ik}$ too high would generate weak optimality cuts, which would increase the computational time of the algorithm. One way to avoid this issue is to solve the assignment problem (6.18) for given design variables $(x, y, \mathcal{N})$ and compute the optimal value of W_{ik} and use that as an upper bound. Further improvements in the Benders decomposition method can involve the use of *pareto-optimal* cuts proposed by [165]. They help in avoiding the generation of multiple optimality cuts for a degenerate subproblem. We tried this strategy, however, we did not find any significant improvement in the solution time using these cuts, therefore, we do not discuss it here. Finally, when RMP is loaded with a large number of cuts we recommend removing the non-active cuts from the model by checking the dual value. There is no guarantee that they will not be generated again, but it will be faster to solve the RMP. The overall steps of the Benders implementation with possible acceleration techniques are summarized in Algorithm 10.

Algorithm 10 Enhanced Benders decomposition implementation

(*Initialize*) Let $t = 0, UB = -\infty, LB = \infty, \mathcal{K}_k^{'} = \phi, \forall k \in D$.
Prepare the master problem with clique and cover inequalities.
Assume an initial feasible value $(x^0, y^0, \mathcal{N}^0)$. Solve the SP (6.22), obtain the optimal dual solutions and append that to the set $\mathcal{K}_k^{'}, \forall k \in D$.
while $UB - LB > \epsilon$ **do** ▷ ϵ is the tolerance parameter
    Set $t = t + 1$. Solve RMP (6.24), obtain its optimal solution $(x_0^t, y_0^t, \mathcal{N}_0^t)$ and other optimal/suboptimal solutions $\{(x_s^t, y_s^t, \mathcal{N}_s^t)\}_{1 \leq s \leq l}$, where l is specified by the user.
    Set LB $= \sum_{k \in D} \eta_k$
    for $s = 0, 1, ..., l$ **do**
        Solve SP (6.22) for $(x_s^t, y_s^t, \mathcal{N}_s^t)$, obtain dual solution and append that to $\mathcal{K}_k^{'}, \forall k \in D$.
    Set $UB = \sum_{k \in D} \left(\sum_{a \in A} c_a v_{ak}^t + \sum_{i \in N^w} W_{ik}^t \right)$

6.4 Computational results

In this section, we present the computational study based on the model (6.19), (6.24), and acceleration techniques presented in Section 6.3.3. We start by describing the details of the experiment used to show the application of the proposed method. Then, we present the details of the network design results in Section 6.4.2, which is followed by the comparison of the computational performance of the solver, the classic Benders implementation, and the enhanced Benders techniques described in Section 6.4.3. Finally, we discuss the results of the sensitivity analysis on two important parameters in the model, namely, the available fleet of buses $\bar{B}$ and vehicles $\bar{F}$ in Section 6.4.4, which is followed by the comparison of the performance of optimized existing transit system and proposed integrated system in Section 6.4.5.

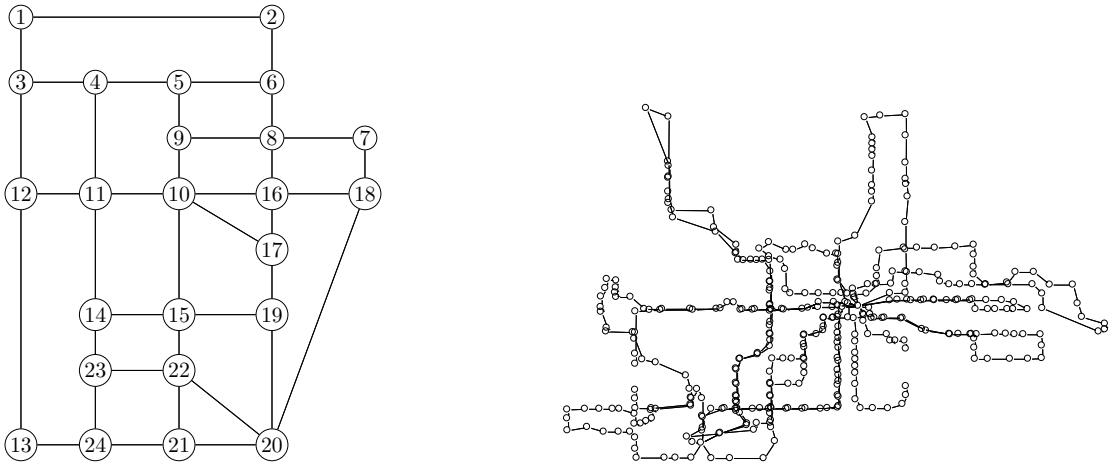
6.4.1 Experiment details

The computational experiments are based on the Sioux Falls road and transit network. The road network has 24 nodes, whereas the static transit network has 384 stops. A walking distance of 0.5 miles is used to create walking links. An illustration of the two networks is shown in Figure 6.2 and the number of different types of links in the network is given in Table 6.2. There are 12 candidate transit routes in the transit network. We consider the set of possible frequencies as $\Theta = \{2, 3, 4, 6, 12\}$ buses/hr to be assigned to any candidate transit route and possible vehicle fleet size to be assigned to any zone as $\Omega = \{0.01, 50, 100, 200, 500\}$. The vehicle fleet size value 0.01 is a dummy element to represent that no vehicles are assigned in a zone and that zone can be served by transit service only. A time-based fare of \$0.21/ min and a base fare of \$0.8 is assumed for the MoD service, whereas the transit fare is assumed to be a fixed value of \$2. To convert the monetary costs into time units, the value of travel time equal to 23\$/hr is used. The

Table 6.2: Number of different types of links in the Sioux Falls multimodal network

Link type	**Number of links**
Access	243
Egress	243
Road	76
Transit	398
Transit transfer	368
Mode transfer	152

value of parameter $\mathcal{A}$ used in the wait time computations of MoD service is assumed to be equal to 0.0017 for all the zones [167]. The available number of buses $\bar{B}$ and vehicles $\bar{F}$ are assumed to be 70 and 3,000 respectively. There are 576 O-D pairs in the network with a total number of trips equal to 36,060. All implementations are coded in Python 3.8 using Gurobi 9.0.1 as the optimization solver. The tests were executed on Intel(R) i7-7700 CPU running at 3.6 GHz with 32 GB RAM under a Windows operating system.



(a) Road network (b) Transit network

Figure 6.2: Sioux Falls network

6.4.2 Network design results

We solve the network design problem (6.19) for the instance explained in Section 6.4.1. The selected transit routes with their optimal frequency are given in Table 6.3. Out of 12 candidate routes, 6 routes are kept operating. The transit network with active and inactive routes is shown in Figure 6.3. We observe that most of the routes are located in the central region of the network. All the routes have been assigned the highest frequency, i.e., 12 buses/hr, except route 8, which has been assigned a frequency of 3 buses/hr. To provide this service, 69 buses are required. The average number of vehicles deployed in each zone is given in Table 6.4. In the optimal allocation of vehicles, it is decided not to deploy any vehicles in 10 zones out of 24 zones. Most of the zones have been allocated 200 vehicles providing an average wait time of 3 minutes. We further observe that the vehicles are deployed in the outskirt zones of the network where transit routes are not located.

Table 6.3: Selected transit routes with their optimal frequency

Route	**Located?**	**Optimal frequency** (buses/hr)	**Average wait time** (min)
1	Yes	12	5
2	No	-	-
3	Yes	12	5
4	Yes	12	5
5	No	-	-
6	No	-	-
7	No	-	-
8	Yes	3	20
9	No	-	-
10	No	-	-
11	Yes	12	5
12	Yes	12	5

The total time spent in the system is equal to 11,301 passenger-hours including

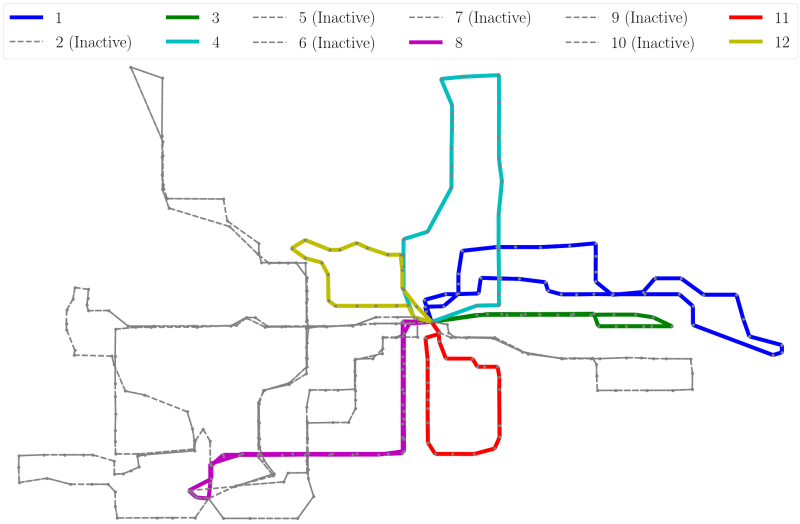


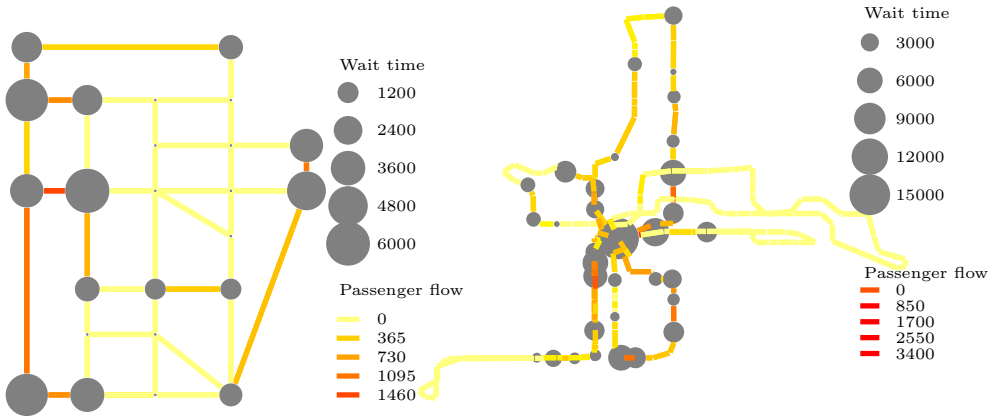
Figure 6.3: Transit routes (inactive routes are shown by dashed gray color)

Table 6.4: Vehicle allocation to different zones

Zone	Vehicles	Avg. Wait time (min)	Zone	Vehicles	Avg. Wait time (min)
1	200	3	13	200	3
2	100	6	14	200	3
3	200	3	15	200	3
4	200	3	16	-	-
5	-	-	17	-	-
6	-	-	18	200	3
7	200	3	19	200	3
8	-	-	20	200	3
9	-	-	21	-	-
10	-	-	22	-	-
11	200	3	23	-	-
12	500	1.2	24	200	3

8,673 passenger-hours of travel time on various links, 1,881 passenger-hours of wait time spent on the transit network, and 747 passenger-hours of wait time spent on the road network. We found that more passengers take transit than MoD service. The share

of passengers using the road, transit, and multimodal service are 23 %, 61 %, and 16 % respectively. The passenger flow on various links and wait time on various nodes of the road and transit networks (resp.) are visualized in Figure 6.4(a) and (b) respectively. We observe that most of the passenger trips in the central zones are made using transit network, whereas the trips on the outskirts of the network are made using both MoD and multimodal service. The figures further show that the congestion in the central zones is significantly improved with the resulting network design.



(a) Flow and wait time of passengers in the road network

(b) Flow and wait time of passengers in the transit network

Figure 6.4: Flow and wait time (pass-min) of passengers in the network

6.4.3 Computational performance

In this section, we compare the computational performance of various models and implementation techniques. We consider the following approaches to compare:

1. Solving model (6.19) using Gurobi solver
2. Solving model (6.24) using Gurobi solver
3. Solving model (6.24) using classic Benders decomposition (Algorithm 8)

4. Solving model (6.24) using Benders decomposition with clique/cover cuts (Section 6.3.3)

5. Solving model (6.24) using Benders decomposition with multiple cuts via multiple solutions (Section 6.3.3)

6. Solving model (6.24) using Benders decomposition with both clique/cover and multiple cuts via multiple solutions

7. Solving model (6.24) using Benders decomposition with disaggregated cuts (Section 6.3.3)

8. Solving model (6.24) using Benders decomposition with disaggregated and clique/cover cuts

9. Solving model (6.24) using Benders decomposition with disaggregated and multiple cuts via multiple solutions

10. Solving model (6.24) using Benders decomposition with disaggregated, clique/cover, and multiple cuts via multiple solutions

To solve the bilinear model (6.19), we set the Gurobi parameter `NonConvex = 2`. For Benders decomposition with multiple cuts via multiple solutions, we set the Gurobi parameters `PoolSolutions = 2`, `PoolGap = 0.01`, `PoolSearchMode = 2`. For all above tests, the maximum time limit was set to 3 hours.

The computational performance of every method is shown in Table 6.5. The iterations are counted as the number of times RMP is solved, the computational time is recorded in seconds, and Gap is defined as $(UB - LB) * 100/UB$. The bilinear model (6.19) is hard to solve, and Gurobi took 3 hours to reach the optimality gap of 13.3

Table 6.5: Computational performance

Method	**Iterations**	**Computational time (s)**	**Gap (%)**
Gurobi bilinear	-	Timed out*	13.3
Gurobi	-	Timed out*	0.62
Classic	734	Timed out*	0.16
Classic + Clique/Cover	510	7,440	0
Classic + Multiple	500	6,524	0
Classic + Clique/Cover + Multiple	423	5,566	0
Disaggregate	31	381	0
Disaggregate + Clique/Cover	30	347	0
Disaggregate + Multiple	28	535	0
Disaggregate + Multiple + Clique/Cover	27	498	0

*Note: Maximum time limit = 3 hours

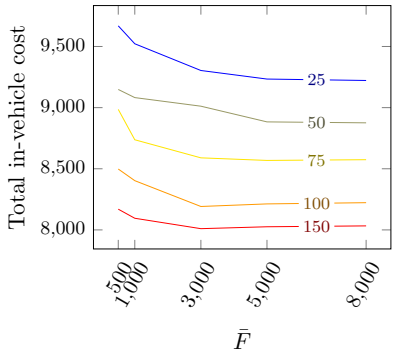
%. The rest of the results are discussed for the optimization model (6.24). Other than Gurobi and classic Benders decomposition, all the methods coverage to the optimal solution. Gurobi and classic Benders decomposition reached an optimality gap of 0.62% and 0.16% respectively. This means both methods reached very close to the optimal solution in 3 hours. The hybrid approach of classic Benders decomposition with both clique/cover cuts and multiple cuts via multiple solutions outperforms the classic Benders decomposition with clique/cover cuts or multiple cuts via multiple solutions only. The disaggregated Benders decomposition is computationally more efficient than any classic Benders approach with cut improvements. The disaggregated cuts with other cuts show further improvement in the solution time and the number of iterations to converge to the optimal solution. The Benders decomposition using disaggregated, clique/cover, and multiple cuts via multiple solutions outperforms other methods in terms of the number of iterations to converge to an optimal solution, whereas Benders decomposition with disaggregated and clique/cover cuts outperform other methods in terms of computational time. This may be because the multiple cuts are generated by solving several subproblems, which takes more computational time, but the generated

cuts may not be as effective. Overall, the experiments performed for this section show that the computational methods presented in this study are quite efficient in solving the current problem exactly.

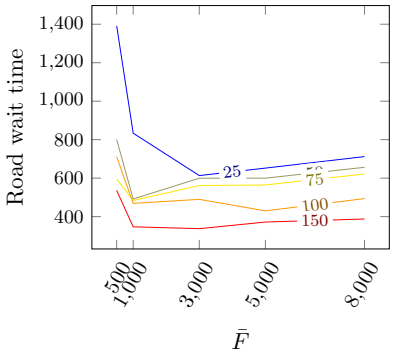
6.4.4 Sensitivity analysis on parameters

The availability of buses and vehicles can result in different network design results. Hence, we choose to perform a sensitivity analysis on the available bus fleet $\bar{B}$ and vehicle fleet $\bar{F}$. We solve the model (6.24) with varying bus fleet size of 25, 50, 75, 100, and 150 and varying vehicle fleet size of 500, 1000, 3000, 5000, 8000. Figure 6.5 and 6.6 show the sensitivity analysis results based on contour plots. The x-axis shows the varying vehicle fleet sizes, and contours represent varying bus fleet sizes. Figure 6.5(a), (b), (c), and (d) show the in-vehicle cost, average road wait time, average transit wait time, and total expected travel cost in passenger-hours respectively. We can observe that the in-vehicle cost decreases with the increase in the number of available vehicles. The effect of increasing the number of buses is more than the increase in the number of vehicles. Moreover, the in-vehicle cost is not affected by increasing the number of vehicles to more than 5,000. The average road wait time decreases with the increase in the number of available vehicles. It also decreases with the increase in the number of available buses due to mode shift. The passenger-hours spent as the wait time in the transit network increases with the increase in the number of available vehicles as well as buses. This is because more passengers take the transit mode as more buses are made available. The overall expected travel cost also reduces with the increase in the bus and vehicle fleet. However, the effect of an increase in the vehicle fleet size of more than 5,000 is negligible. Figure 6.6 shows the mode share as a function of the available bus and vehicle fleet size. As expected, the transit and MoD share increase with the increase in the available bus and vehicle fleet size respectively. The share of multimodal

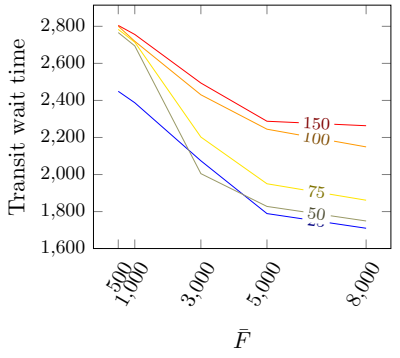
service increases with the number of vehicles and buses up to 5,000 and 75 respectively but declines after that. The decline in multimodal share is because of the reduced wait time for both services, which drives passengers to use single mode rather than multiple modes.



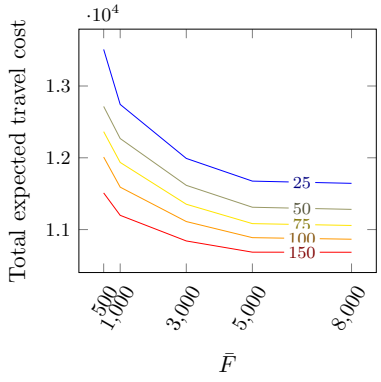
(a) In-vehicle cost (pass-hrs)



(b) Average road wait time (pass-hrs)

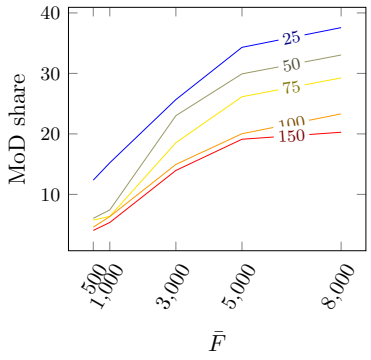


(c) Average transit wait time (pass-hrs)

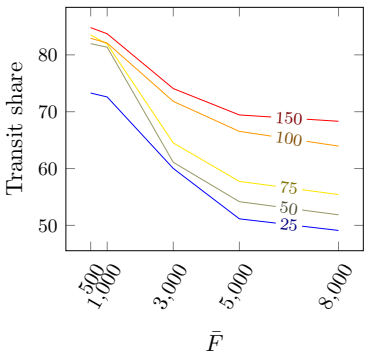


(d) Total expected travel cost (pass-hrs)

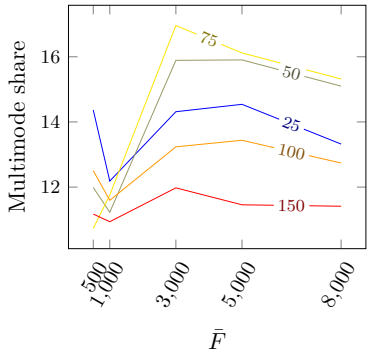
Figure 6.5: Sensitivity of parameters $\bar{F}$ and $\bar{B}$ on different costs (contour represents varying bus fleet sizes)



(a) MoD share



(b) Transit share



(c) Multimode share

Figure 6.6: Sensitivity of parameters $\bar{F}$ and $\bar{B}$ on mode share (contour represents varying bus fleet sizes)

6.4.5 Comparison of optimized base transit system with proposed integrated system

In this section, we present a comparison of the operation of the "optimized base transit system" corresponding to the existing transit system with optimized frequencies versus the design of the integrated system evaluated in Section 6.4.2. For the optimized base case, we solve the optimization program (6.26) for the instance described in Section 6.4.1. The results of optimized frequencies of various routes are given in Table 6.6. The network provides an average wait time of 18 minutes to the passengers.

$$\underset{v,W,y}{\text{minimize}} \quad \sum_{k\in D}\left(\sum_{a\in A} c_a v_{ak} + \sum_{i\in N_T^w} W_{ik}\right) \tag{6.26a}$$

$$\text{subject to} \quad \sum_{l\in L}\sum_{f\in\Theta} \mathcal{B}(l,f)\times y_{lf} \le \bar{B} \tag{6.26b}$$

$$\sum_{a\in FS(i)} v_{ak} = \sum_{a\in BS(i)} v_{ak} + g_{ik}, \forall i \in N, \forall k \in D \tag{6.26c}$$

$$v_{ak} \le \left(\sum_{f\in\Theta} f y_{l(a)f}\right) W_{ik}, \forall a \in FS(i) : a \in A, \forall i \in N^w, \forall k \in D \tag{6.26d}$$

$$v_{ak} \ge 0, \forall a \in A, \forall k \in D \tag{6.26e}$$

$$W_{ik} \text{ free }, \forall i \in N_T^w, \forall k \in D \tag{6.26f}$$

$$y_{lf} \in \mathfrak{B}, \forall f \in \Theta, \forall l \in L \tag{6.26g}$$

Table 6.6: Routes with their optimal frequency (optimized base case)

Route	**Optimal frequency** (buses/hr)	**Average wait time** (min)
1	4	15
2	2	30
3	6	10
4	12	5
5	3	20
6	2	30
7	3	20
8	3	20
9	3	20
10	2	30
11	12	5
12	6	10

The results comparing the performance of the optimized base transit system and integrated system are provided in Table 6.7. The base transit system has 12 two-way bus services operated by a bus fleet of 69 buses, whereas the new integrated system has 6 two-way bus services operated by 69 buses. Along with 69 buses, the new integrated

system deploys 3,000 vehicles to serve the demand. The deployment of these extra vehicles can be costly to the transportation agencies. However, they provide several benefits. First, the optimized base transit system is not able to serve 13 % of the demand due to the non-availability of transit service in 2 zones in the network. On the other hand, the first mile and last mile of these zones are covered by vehicles in the integrated system. Second, the average in-vehicle travel time of passengers using the integrated system is only 14.43 minutes in comparison to the 21.16 minutes for passengers using the base transit system. However, the average wait time of the integrated system users is increased slightly in comparison to the base transit system. This is due to the increased number of transfers to access MoD and transit service.

Table 6.7: Comparison of optimized base transit system and integrated system

	Optimized base transit system	**Integrated system**
Number of active routes	12	6
Number of buses used	69	69
Number of vehicles used	0	3,000
Satisfied demand (%)	87	100
Average in-vehicle time (min/passenger)	21.16	14.43
Average wait time (min/passenger)	2.88	4.37

6.4.6 Managerial insights for implementing such service

For implementing the proposed model in practice, we need to follow the following procedure. First, we divide the region into zones. Second, we collect the peak hour demand data in the region. Third, solve the proposed design model for varying fleet sizes. This step will be similar to the sensitivity analysis given in Section 6.4.4. This analysis will help us decide the optimal fleet size of buses and vehicles for our service. It will also provide us with the allocation of vehicles and buses for different zones and bus routes respectively. This allocation is designed for peak hours. We can reduce the vehicle

operation in non-peak hours. For the bus service, scheduling needs to be performed to publish a schedule for the service.

Chapter 7

Conclusions and Future Work

The current dissertation develops modeling frameworks to predict passenger behavior and design integrated transit networks. It develops routing algorithms to describe the strategic behavior of transit and park-and-ride passengers using online information when traveling in a stochastic and time-dependent network. Using the transit passenger routing algorithm, it further develops a schedule-based transit assignment model with online information for uncapacitated and capacitated networks. Then, it employs the strategic frequency-based assignment models for the design of long-term planning of transit networks as well as integrated Mobility-on-Demand and transit systems. This chapter describes the summary of the research presented in this dissertation, draws conclusions, and outlines suggestions for the future research.

7.1 Summary of results and conclusions

The modeling contributions made in this dissertation start with Chapter 3, which develops routing algorithms to predict passenger route choice behavior in presence of online information. In transit networks, the online information about bus arrival time induces

an adaptive behavior where a passenger who faces failed transfer due to early or late arrival of buses considers alternative bus routes to minimize their expected cost to destination. In the multimodal network of park-and-ride mode, the decision to whether, where, and when to park a car and take transit is made adaptively based on the realized state of traffic on a freeway and wait time for transit transfer. We model these phenomena using stochastic shortest path problems in a network with time-dependent and stochastic link travel times. To solve the corresponding Bellman equation, it proposes two solution algorithms, namely, value iteration and label correcting algorithms. Two case studies are presented to show the application of the proposed models. In the case of transit networks, a case study based on the schedule-based transit network given in [2] is presented. It shows how the presence of online information changes passenger routing behavior. Specifically, passengers consider alternative routes to travel in case of missed transfers which provides a lesser expected cost than the *a priori* shortest path. We also observed that the label correcting algorithm outperforms the value iteration algorithm. In the case of the park-and-ride network, a case study of freeway I-394 in Minneapolis is presented. Results computed the time and state of the realized traffic when park-and-ride mode becomes an attractive mode. For example, the park-and-ride mode provides faster access to Downtown Minneapolis during severe congestion between 7:00-9:45 A.M. We expect to see further improvement in the travel time if more frequent service of buses is provided by Metro Transit on I-394. In terms of attractiveness among park-and-ride locations, Co. Rd. Park-and-Ride and Plymouth Road Transit Center were found to be more attractive parking locations than others. The standard deviation of travel time obtained by using the adaptive policy varies between 20-222 seconds. Finally, a sensitivity analysis of the parking cost and transit fare reveals that the higher parking cost and lower transit fare make the park-and-ride mode more attractive to

commuters. With $2 increase in the transit fare at $15 parking cost, we see a 2% decline in the share of the park-and-ride option in the optimal policy. Similarly, with $15 increase in the parking cost at $2 transit fare, we see a 4% increase in the share of the park-and-ride option in the optimal policy. We conclude that by using adaptive policy, a commuter will save around 36 hours every year.

Chapter 4 presents two transit assignment models based on whether or not the limited capacity of transit vehicles is considered. The uncapacitated assignment model is useful for transit systems with low ridership but unreliable service, whereas the capacitated assignment model is useful for transit systems with high ridership and unreliable service. In both cases, it proposes that passengers adopt strategies to travel and use the stochastic shortest path as a modeling tool to characterize passenger hyperpaths. Under restrictive assumptions, a linear program is developed to perform the uncapacitated assignment. On the other hand, the capacitated assignment is observed to be more complex than the uncapacitated assignment. This is because the strategic behavior of passengers is observed not only because of online information but also due to the limited capacity of transit vehicles. For this purpose, the chapter formulates the capacitated assignment problem as a variational inequality problem, which is solved using an MSA-based heuristic algorithm. The algorithm runs the shortest path as well as a loading procedure to incorporate realistic passenger behavior. The MSA algorithm shows good convergence performance in the conducted experiments. We present a case study based on the schedule-based transit network given in [2]. The result evaluated the departure times of various groups and passenger flows on various trips on transit routes. The computational time required to perform both uncapacitated and capacitated assignments was within 10 minutes. The limited capacity results in high-dimensional strategies and more complex behavior.

The dissertation further proposes optimization models to design networks while incorporating strategic passenger behavior through frequency-based transit assignment. For this purpose, Chapter 5 proposes a bi-level transit network design model. It incorporates both passenger's and operator's perspectives in the same model by considering measures of passenger travel time, limited bus fleet, operating cost of buses, bus capacity, road capacity, and the number of transfers experienced by passengers. Several important results related to the current problem, such as characterization of bi-level feasible design and feasibility of the lower level problem are proved in the study. These results help develop an efficient branch-and-benders cut algorithm to solve the problem. Numerical experiments are performed on various network instances to flashlight on different aspects of network design and the efficacy of the solution method. We observe that the current model creates a network where passengers do not face denied boarding, and only a few routes are needed to serve the whole demand. The sensitivity analysis on passenger weight β shows that a higher value of β results in the allocation of more buses to the frequency design, which results in lower passenger cost and higher operational cost. Furthermore, we observe that including the operator's perspective is important to obtain an economical design with an acceptable level of service to passengers. The sensitivity analysis on fleet size shows that increasing the fleet size allocates more buses to the routes, which results in lesser passenger costs and more operating costs. The model also considers different transfer constraints, namely soft and hard transfer constraints. The numerical experiments show that soft capacity constraints are more effective in controlling the number of transfering passengers. We conclude that the proposed method with enhancements such as cut-set-inequalities and normalizing infeasibility cuts results in the best performance among all the strategies. Finally, higher demand in the same network leads to higher computational time to solve the same design model.

Chapter 6 develops a mixed-integer non-linear program (MINLP) to design an intermodal system, where the first mile and last mile of transit trips are served using the MoD service. The MINLP was relaxed to a mixed-integer linear program with the help of McCormick relaxations. To solve the resulting MILP model efficiently, it proposes the Benders decomposition method with several enhancements. These enhancements include the use of disaggregated cuts, clique/cover cuts, and multiple cuts via multiple solutions. The numerical results show that disaggregated cuts with clique/cover cuts and multiple cuts via multiple solutions are efficient techniques to solve the current problem. Furthermore, the experiment results on the Sioux Falls network show that the congestion in the city center is improved with such design as most of the passengers were found to take the transit in that region. The sensitivity analysis on bus fleet size and vehicle fleet size reveals that the passenger hours spent in the system as in-vehicle time and wait time reduces with an increase in the number of available buses and vehicles. For the city of Sioux Falls, the share of multimodal service was observed to be highest for the vehicle fleet size and bus fleet size of 3,000 and 75 respectively. We also compared the proposed integrated system with the optimized base-case transit system. We found that the integrated system can be costly due to the deployment of vehicles, but it reduces the passenger in-vehicle time and serves more demand than the optimized base case.

7.2 Recommendations for future research

The current dissertation opens new avenues for research in the area of modeling and design of integrated transit systems. This section discusses how to relax the assumptions made in the presented models and extend these ideas for other types of transportation

solutions. It is discussed as follows:

First, in the routing algorithms and transit assignment presented in 3 and 4, passengers are assumed to be expected-cost-minimizers. In reality, they may consider other factors, such as risk related to close transfers, preferences towards different modes, etc., as part of their utility. One way to address this issue is to consider an adaptive routing policy based on a discrete choice model. The proposed models provide a flexible framework to incorporate choice probabilities. For example, one can use logit-based route choice probabilities to achieve a stochastic user equilibrium.

Second, one of the issues in solving the capacitated assignment problem presented in Chapter 4 for large networks is the explosion of state space due to the incorporation of the availability vector in it. It causes the high computational time required to solve the corresponding stochastic shortest path (SSP) problem. For the assignment, one needs to solve the SSP several times, which makes it challenging to produce assignment results for large-scale transit systems. Future research should focus on proposing techniques to solve this problem faster. This could be achieved using approximate dynamic programming algorithms. The exact solution methods presented in the dissertation will help evaluate the accuracy of the approximation algorithms.

Third, for the park-and-ride passenger routing in Chapter 3, we considered only one-way routing of a commuter, but in reality, nearly all commuters have to return home. The park-and-ride policy found for one-way routing may not provide an optimal tour. The problem can be addressed using the network transformation proposed by [103] for finding the multimodal multi-destination tour. The development of an efficient algorithm for solving such a problem needs further investigation.

Fourth, more research is needed on finding strong valid inequalities specific to the bi-level transit network design problem presented in Chapter 5. This is evident from the numerical experiments, where we observe that most of the time in solving the current problem using the proposed method is spent in closing the gap after finding the optimal solution. This is because the linear relaxation of the current single-level reformulation is weak. We observe that cut-set-based inequalities improve the initial lower bound. Therefore, an intelligent way of generating high-dimensional cut-set-based inequalities could further improve the lower bound. Other cuts such as metric inequalities should be developed for optimal strategy transit assignment with side constraints on the flow of passengers.

Fifth, for the bi-level transit network design problem, it was assumed that the transit agency (leader) is optimistic. This means if there are multiple solutions to the lower level problem, the leader selects the lower level solution that benefits it the most. However, in real-life, different lower-level (passenger behavior) solutions could take place. This would result in proposing the design based on the incorrect prediction of passenger behavior. The uniqueness of the lower-level solution can be checked by solving the high point problem with an extra constraint restricting the upper-level objective value equal to the current optimal value and minimizing the upper-level objective with $\beta = 1$. If the objective value for this new problem turns out to be less than the computed lower-level objective value, then the uniqueness does not hold. In all our numerical experiments, we found that the lower-level solution was unique. However, to avoid risking the incorrect prediction of passenger behavior, one should solve the pessimistic bi-level network design. A significant research effort is needed to characterize the optimal solution and develop an efficient solution algorithm for this variant.

Sixth, ridepooling can be incorporated into the design of integrated transit systems. This requires further investigations on the ways to include the matching of passengers for ridepooling, which will reduce the size of the vehicle fleet required to provide the service. This investigation will have an impact on the planning of an integrated AV and transit system.

Seventh, a congested frequency-based transit assignment should be incorporated in the design models presented in both Chapter 5 and Chapter 6. For example, one can use the transit assignment model proposed by [168]. It will require developing an efficient algorithm for solving their assignment model and using that for the solution method of the design problem.

Eighth, a model-free reinforcement learning model can be developed to predict passenger behavior in presence of online information. The calibration of such a model using travel behavior data (e.g., Automatic Fare Collection data) will require a significant effort.

Ninth, finding the optimal park-and-ride facility locations in the network considering adaptive routing will be another interesting research problem to investigate. It can be formulated as a bi-level problem.

Tenth, Chapter 4 developed a transit assignment model with online information. However, the idea can extend to multimodal transportation systems.

References

[1] W. Y. Szeto and Y. Jiang. Transit route and frequency design: Bi-level modeling and hybrid artificial bee colony algorithm approach. *Transportation Research Part B: Methodological*, 67:235–263, 2014.

[2] C O Tong and A J Richardson. A computer model for finding the time-dependent minimum path in a transit system with fixed schedules. *Journal of Advanced Transportation*, 18(2):145–161, 1984.

[3] Héctor Cancela, Antonio Mauttone, and María E. Urquhart. Mathematical programming formulations for transit network design. *Transportation Research Part B: Methodological*, 77:17–37, 2015.

[4] W. Y. Szeto and Y. Jiang. Transit route and frequency design: Bi-level modeling and hybrid artificial bee colony algorithm approach. *Transportation Research Part B: Methodological*, 67:235–263, 2014.

[5] Christopher Mandl. *Applied Network Optimization*. Academic Press, London, 1980.

[6] Janette Sadik-Khan and Seth Solomonow. Fear of Public Transit Got Ahead of the Evidence, 2020.

[7] Heinz Spiess and Michael Florian. Optimal strategies: A new assignment model for transit networks. *Transportation Research Part B*, 23(2):83–102, 1989.

[8] Ravindra K Ahuja, Thomas L Magnanti, and James B Orlin. *Network Flows: Theory, Algorithms, and Applications*. Pearson; 1 edition, 1st edition, 1988.

[9] Alireza Khani, Mark Hickman, and Hyunsoo Noh. Trip-based path algorithms using the transit network hierarchy. *Networks and Spatial Economics*, 15(3):635–653, 2015.

[10] Alireza Khani, Sanggu Lee, Mark Hickman, Hyunsoo Noh, and Neema Nassir. Intermodal Path Algorithm for Time-Dependent Auto Network and Scheduled Transit Service. *Transportation Research Record: Journal of the Transportation Research Board*, 2284(1):40–46, 2012.

[11] Pitu B Mirchandani and Hossein Soroush. Optimal paths in probabilistic networks: A case with temporary preferences. *Computers & operations research*, 12(4):365–381, 1985.

[12] Alireza Khani and Stephen D Boyles. An exact algorithm for the mean–standard deviation shortest path problem. *Transportation Research Part B: Methodological*, 81:252–266, 2015.

[13] Yufeng Zhang and Alireza Khani. An algorithm for reliable shortest path problem with travel time correlations. *Transportation Research Part B: Methodological*, 121:92–113, 2019.

[14] S. Travis Waller and Athanasios K. Ziliaskopoulos. On the Online Shortest Path Problem with Limited Arc Cost Dependencies. *Networks*, 40(4):216–227, 2002.

[15] John S Croucher. A note on the stochastic shortest-route problem. *Naval Research Logistics Quarterly*, 25(4):729–732, 1978.

[16] Randolph W. Hall. The Fastest Path through a Network with Random Time-Dependent Travel Times. *Transportation Science*, 20(3):182–188, 1986.

[17] Giovanni Andreatta and Luciano Romeo. Stochastic shortest paths with recourse. *Networks*, 18(3):193–204, 1988.

[18] Harilaos N Psaraftis and John N Tsitsiklis. Dynamic shortest paths in acyclic networks with Markovian arc costs. *Operations Research*, 41(1):91–101, 1993.

[19] George H Polychronopoulos and John N Tsitsiklis. Stochastic shortest path problems with recourse. *Networks: An International Journal*, 27(2):133–143, 1996.

[20] Elise D. Miller-Hooks and Hani S. Mahmassani. Least Expected Time Paths in Stochastic, Time-Varying Transportation Networks. *Transportation Science*, 34(2):198–215, 2003.

[21] Elise Miller-Hooks. Adaptive least-expected time paths in stochastic, time-varying transportation and data networks. *Networks*, 37(1):35–52, 2001.

[22] Song Gao and Ismail Chabini. Optimal routing policy problems in stochastic time-dependent networks. *Transportation Research Part B: Methodological*, 40(2):93–122, 2006.

[23] J. Scott Provan. A Polynomial-Time Algorithm to Find Shortest Paths with Recourse. *Networks*, 41(2):115–125, 2003.

[24] Stephen D Boyles and Tarun Rambha. A note on detecting unbounded instances of the online shortest path problem. *Networks*, 67(4):270–276, 2016.

[25] Song Gao, Emma Frejinger, and Moshe Ben-Akiva. Adaptive route choice models in stochastic time-dependent networks. *Transportation Research Record*, 2085(1):136–143, 2008.

[26] S Gao. *Optimal adaptive routing and traffic assignment in stochastic time-dependent networks*. PhD thesis, Massachusetts Institute of Technology, 2004.

[27] Jing Ding-Mastera. *Adaptive Route Choice in Stochastic Time-Dependent Networks: Routing Algorithms and Choice Modeling*. PhD thesis, University of Massachusetts Amherst, 2016.

[28] Song Gao and He Huang. Real-time traveler information for optimal adaptive routing in stochastic time-dependent networks. *Transportation Research Part C: Emerging Technologies*, 21(1):196–213, 2012.

[29] Claude Chriqui and Pierre Robillard. Common bus lines. *Transportation science*, 9(2):115–121, 1975.

[30] Yang Liu, Sebastien Blandin, and Samitha Samaranayake. Stochastic on-time arrival problem in transit networks. *Transportation Research Part B: Methodological*, 119:122–138, 2019.

[31] Yu Marco Nie and Xing Wu. Shortest path problem considering on-time arrival probability. *Transportation Research Part B: Methodological*, 43(6):597–613, 2009.

[32] Tarun Rambha, Stephen D. Boyles, and S. Travis Waller. Adaptive Transit Routing in Stochastic Time-Dependent Networks. *Transportation Science*, 50(3):1043–1059, 2016.

[33] Alireza Khani. An online shortest path algorithm for reliable routing in schedule-based transit networks considering transfer failure probability. *Transportation Research Part B: Methodological*, 126:549–564, 2019.

[34] S. Nguyen and S. Pallottino. Equilibrium traffic assignment for large scale transit networks. *European Journal of Operational Research*, 37(2):176–186, 1988, /dx.doi.org/10.1016/0377-2217(88)90327-X.

[35] Jia Hao Wu, Michael Florian, and Patrice Marcotte. Transit Equilibrium Assignment: A Model and Solution Algorithms. *Transportation Science*, 28(3):193–203, 1994.

[36] Joaquin De Cea and Enrique Fernández. Transit assignment for congested public transport systems: an equilibrium model. *Transportation science*, 27(2):133–147, 1993.

[37] Roberto Cominetti and José Correa. Common-lines and passenger assignment in congested transit networks. *Transportation science*, 35(3):250–267, 2001.

[38] Manuel Cepeda, Roberto Cominetti, and Michael Florian. A frequency-based assignment model for congested transit networks with strict capacity constraints: characterization and computation of equilibria. *Transportation research part B: Methodological*, 40(6):437–459, 2006.

[39] Fumitaka Kurauchi, Michael G.H. Bell, and Jan Dirk Schmöcker. Capacity Constrained Transit Assignment with Common Lines. *Journal of Mathematical Modelling and Algorithms*, 2(4):309–327, 2003.

[40] Agostino Nuzzolo, Umberto Crisalli, and Luca Rosati. A schedule-based assignment model with explicit capacity constraints for congested transit networks. *Transportation Research Part C: Emerging Technologies*, 20(1):16–33, 2012.

[41] Sang Nguyen, Stefano Pallottino, and Federico Malucelli. A modeling framework for passenger assignment on a transport network with timetables. *Transportation Science*, 35(3):238–249, 2001.

[42] C O Tong and S C Wong. A stochastic transit assignment model using a dynamic schedule-based network. *Transportation Research Part B: Methodological*, 33(2):107–121, 1999.

[43] M H Poon, S C Wong, and C O Tong. A dynamic schedule-based model for congested transit networks. *Transportation Research Part B: Methodological*, 38(4):343–368, 2004.

[44] Umberto Crisalli. Dynamic transit assignment algorithms for urban congested networks. *WIT Transactions on The Built Environment*, 44, 1970.

[45] Otto Anker Nielsen. A large scale stochastic multi-class schedule-based transit model with random coefficients. In *Schedule-based dynamic transit modeling: theory and applications*, pages 53–77. Springer, 2004.

[46] Agostino Nuzzolo, Francesco Russo, and Umberto Crisalli. A doubly dynamic schedule-based assignment model for transit networks. *Transportation Science*, 35(3):268–285, 2001.

[47] Younes Hamdouch, Patrice Marcotte, and Sang Nguyen. Capacitated transit assignment with loading priorities. *Mathematical programming*, 101(1):205–230, 2004.

[48] Younes Hamdouch and Siriphong Lawphongpanich. Schedule-based transit assignment model with travel strategies and capacity constraints. *Transportation Research Part B: Methodological*, 42, 2008.

[49] Hyunsoo Noh, Mark Hickman, and Alireza Khani. Hyperpaths in network based on transit schedules. *Transportation research record*, 2284(1):29–39, 2012.

[50] Patrice Marcotte, Sang Nguyen, and Alexandre Schoeb. A strategic flow model of traffic assignment in static capacitated networks. *Operations Research*, 52(2):191–212, 2004.

[51] Maëlle Zimmermann, Emma Frejinger, and Patrice Marcotte. A strategic markovian traffic equilibrium model for capacitated networks. *Transportation Science*, 55(3):574–591, 2021.

[52] Agachai Sumalee, Zhijia Tan, and William H K Lam. Dynamic stochastic transit assignment with explicit seat allocation model. *Transportation Research Part B: Methodological*, 43(8-9):895–912, 2009.

[53] Younes Hamdouch, H W Ho, Agachai Sumalee, and Guodong Wang. Schedule-based transit assignment model with vehicle capacity and seat availability. *Transportation Research Part B: Methodological*, 45(10):1805–1830, 2011.

[54] Stefan Binder, Yousef Maknoon, and Michel Bierlaire. Exogenous priority rules for the capacitated passenger assignment problem. *Transportation Research Part B: Methodological*, 105:19–42, 2017.

[55] Guido Gentile, Sang Nguyen, and Stefano Pallottino. Route choice on transit networks with online information at stops. *Transportation science*, 39(3):289–297, 2005.

[56] Carolina Billi, Guido Gentile, and Sang Nguyen. Rethinking the wait model at transit stops. In *Proceedings of TRISTAN V : The Fifth Triennial Symposium on Transportation Analysis*, number November, pages 1–8, Le Gosier, 2004.

[57] Peng Will Chen and Yu Marco Nie. Optimal transit routing with partial online information. *Transportation Research Part B: Methodological*, 72:40–58, 2015.

[58] Nurit Oliker and Shlomo Bekhor. A frequency based transit assignment model that considers online information. *Transportation Research Part C: Emerging Technologies*, 88:17–30, 2018.

[59] Younes Hamdouch, W Y Szeto, and Y Jiang. A new schedule-based transit assignment model with travel strategies and supply uncertainties. *Transportation Research Part B: Methodological*, 67:35–67, 2014.

[60] Yuqing Zhang, William H K Lam, Agachai Sumalee, Hong K Lo, and C O Tong. The multi-class schedule-based transit assignment model under network uncertainties. *Public Transport*, 2(1):69–86, 2010.

[61] Clara Brimnes Gardner, Sara Dorthea Nielsen, Morten Eltved, Thomas Kjær Rasmussen, Otto Anker Nielsen, and Bo Friis Nielsen. Calculating conditional passenger travel time distributions in mixed schedule-and frequency-based public transport networks using Markov chains. *Transportation Research Part B: Methodological*, 152:1–17, 2021.

[62] Mark D. Hickman and Nigel H M Wilson. Passenger travel time and path choice implications of real-time transit information. *Transportation Research Part C*, 3(4):211–226, 1995.

[63] Mark D Hickman and David H Bernstein. Transit service and path choice models in stochastic and time-dependent networks. *Transportation Science*, 31(2):129–146, 1997.

[64] Mohamed Wahba and Amer Shalaby. MILATRAS: A new modeling framework for the transit assignment problem. In *Schedule-Based Modeling of Transportation Networks*, pages 1–24. Springer, 2009.

[65] Agostino Nuzzolo, Umberto Crisalli, Antonio Comi, and Luca Rosati. A mesoscopic transit assignment model including real-time predictive information on crowding. *Journal of Intelligent Transportation Systems*, 20(4):316–333, 2016.

[66] Oded Cats and Zafeira Gkioulou. Modeling the impacts of public transport reliability and travel information on passengers' waiting-time uncertainty. *EURO Journal on Transportation and Logistics*, 6(3):247–270, 2017.

[67] Guy Desaulniers and Mark D Hickman. Chapter 2 Public Transit. In Cynthia Barnhart and Gilbert Laporte, editors, *Transportation*, volume 14 of *Handbooks in Operations Research and Management Science*, pages 69–127. Elsevier, 2007.

[68] Valérie Guihaire and Jin Kao Hao. Transit network design and scheduling: A global review. *Transportation Research Part A: Policy and Practice*, 42(10):1251–1273, 2008.

[69] Konstantinos Kepaptsoglou and Matthew Karlaftis. Transit route network design problem: Review. *Journal of Transportation Engineering*, 135(8):491–505, 2009.

[70] Anita Schöbel. Line planning in public transportation: Models and methods. *OR Spectrum*, 34(3):491–510, 2012.

[71] Reza Zanjirani Farahani, Elnaz Miandoabchi, W. Y. Szeto, and Hannaneh Rashidi. A review of urban transportation network design problems. *European Journal of Operational Research*, 229(2):281–302, 2013.

[72] Avishai Ceder. *Public transit planning and operation: Modeling, practice and behavior*. CRC press, 2016.

[73] T. L. Magnanti and R. T. Wong. Network Design and Transportation Planning: Models and Algorithms. *Transportation Science*, 18(1):1–55, 1984.

[74] S. M. Hassan Mahdavi Moghaddam, K. Ramachandra Rao, G. Tiwari, and Pravesh Biyani. Simultaneous Bus Transit Route Network and Frequency Setting Search Algorithm. *Journal of Transportation Engineering, Part A: Systems*, 145(4):04019011, 2019.

[75] Ralf Borndörfer, Martin Grötschel, and Marc E Pfetsch. A Column-Generation Approach to Line Planning in Public Transport. *Transportation Science*, 41(1):123–132, 2007.

[76] Christoph E. Mandl. Evaluation and optimization of urban public transportation networks. *European Journal of Operational Research*, 5(6):396–404, 1980.

[77] John R. Current, Charles S. ReVelle, and Jared L. Cohon. The hierarchical network design problem. *European Journal of Operational Research*, 27(1):57–66, 1986.

[78] Avishai Ceder and Nigel H M Wilson. Bus network design. *Transportation Research Part B: Methodological*, 20(4):331–344, 1986.

[79] M Hadi Baaj and Hani S Mahmassani. An AI-based approach for transit route system planning and design. *Journal of advanced transportation*, 25(2):187–209, 1991.

[80] Saeed Asadi Bagloee and Avishai Avi Ceder. Transit-network design methodology for actual-size road networks. *Transportation Research Part B: Methodological*, 45(10):1787–1804, 2011.

[81] Ahmed Tarajo Buba and Lai Soon Lee. A differential evolution for simultaneous transit network design and frequency setting problem. *Expert Systems with Applications*, 106:277–289, 2018.

[82] Leena Ahmed, Christine Mumford, and Ahmed Kheiri. Solving urban transit route design problem using selection hyper-heuristics. *European Journal of Operational Research*, 274(2):545–559, 2019.

[83] Jan Willem Goossens, Stan van Hoesel, and Leo Kroon. A branch-and-cut approach for solving railway line-planning problems. *Transportation Science*, 38(3):379–393, 2004.

[84] Konrad Steiner and Stefan Irnich. Schedule-based integrated intercity bus line planning via branch-and-cut. *Transportation Science*, 52(4):882–897, 2018.

[85] Ángel G. Marín and Patricia Jaramillo. Urban rapid transit network design: Accelerated Benders decomposition. *Annals of Operations Research*, 169(1):35–53, 2009.

[86] Arthur Mahéo, Philip Kilby, and Pascal Van Hentenryck. Benders decomposition for the design of a hub and shuttle public transit system. *Transportation Science*, 53(1):77–88, 2019, 1601.00367.

[87] Isabelle Constantin and Michael Florian. Optimizing frequencies in a transit network: a nonlinear bi-level programming approach. *International Transactions in Operational Research*, 2(2):149–164, 1995.

[88] Ziyou Gao, Huijun Sun, and Lian Long Shan. A continuous equilibrium network design model and algorithm for transit systems. *Transportation Research Part B: Methodological*, 38(3):235–250, 2004.

[89] J. F. Guan, Hai Yang, and S. C. Wirasinghe. Simultaneous optimization of transit line configuration and passenger line assignment. *Transportation Research Part B: Methodological*, 40(10):885–902, 2006.

[90] Kenetsu Uchida, Agachai Sumalee, David Watling, and Richard Connors. Study on optimal frequency design problem for multimodal network using probit-based user equilibrium assignment. *Transportation Research Record*, 1923(1):236–245, 2005.

[91] Kenetsu Uchida, Agachai Sumalee, David Watling, and Richard Connors. A study on network design problems for multi-modal networks by probit-based stochastic user equilibrium. *Networks and Spatial Economics*, 7(3):213–240, 2007.

[92] Bin Yu, Zhongzhen Yang, and Jinbao Yao. Genetic algorithm for bus frequency optimization. *Journal of Transportation Engineering*, 136(6):576–583, 2010.

[93] Bin Yu, Lu Kong, Yao Sun, Baozhen Yao, and Ziyou Gao. A bi-level programming for bus lane network design. *Transportation Research Part C: Emerging Technologies*, 55:310–327, 2015.

[94] W H Wilson. Statewide Intermodal Transportation Planning in the Less Urbanized State. *Highway Research Record*, 401(1), 1972.

[95] Hai Wang. Routing and Scheduling for a Last-Mile Transportation System. *Transportation Science*, 53(December 2018):trsc.2017.0753, 2017.

[96] Arthur Maheo, Philip Kilby, and Pascal Van Hentenryck. Benders decomposition for the design of a hub and shuttle public transit system. *Transportation Science*, 53(1):77–88, 2017.

[97] Randall Cayford and Y B Youngbin Yim. Personalized Demand-Responsive Transit Service. Technical report, California PATH Research Report, California, 2004.

[98] D Koffman. *Operational Experiences with Flexible Transit Services*. National Academies Press, 2004.

[99] Alan Lee and Martin Savelsbergh. An extended demand responsive connector. *EURO Journal on Transportation and Logistics*, 6(1):25–50, 2017.

[100] Luca Quadrifoglio, Maged M. Dessouky, and Fernando Ordóñez. A simulation study of demand responsive transit system design. *Transportation Research Part A: Policy and Practice*, 42(4):718–737, 2008.

[101] Chung-Wei Shen and Luca Quadrifoglio. Evaluation of Zoning Design with Transfers for Paratransit Services. *Transportation Research Record: Journal of the Transportation Research Board*, 2277(1):82–89, 2012.

[102] Xiugang Li and Luca Quadrifoglio. Optimal Zone Design for Feeder Transit Services. *Transportation Research Record: Journal of the Transportation Research Board*, 2111(1):100–108, 2009.

[103] Neema Nassir, Alireza Khani, Mark Hickman, and Hyunsoo Noh. Algorithm for Intermodal Optimal Multidestination Tour with Dynamic Travel Times. *Transportation Research Record: Journal of the Transportation Research Board*, 2283:57–66, 2012.

[104] Alireza Khani, Sanggu Lee, Mark Hickman, Hyunsoo Noh, and Neema Nassir. Intermodal Path Algorithm for Time-Dependent Auto Network and Scheduled Transit Service. *Transportation Research Record: Journal of the Transportation Research Board*, 2284(2284):40–46, 2012.

[105] Alexander Webb and Alireza Khani. Park-and-Ride Choice Behavior in a Multimodal Network with Overlapping Routes. *Transportation Research Record*, 2674(3):150–160, 2020.

[106] Neda Masoud, Daisik Nam, Jiangbo Yu, and R. Jayakrishnan. Promoting Peer-to-Peer Ridesharing Services as Transit System Feeders. *Transportation Research Record: Journal of the Transportation Research Board*, 2650:74–83, 2017.

[107] Mitja Stiglic, Niels Agatz, Martin Savelsbergh, and Mirko Gradisar. Enhancing urban mobility: Integrating ride-sharing and public transit. *Computers and Operations Research*, 90:12–21, 2018.

[108] Zheyong Bian and Xiang Liu. Mechanism design for first-mile ridesharing based on personalized requirements part I: Theoretical analysis in generalized scenarios. *Transportation Research Part B: Methodological*, 120:147–171, 2019.

[109] Tai Yu Ma, Saeid Rasulkhani, Joseph Y.J. Chow, and Sylvain Klein. A dynamic ridesharing dispatch and idle vehicle repositioning strategy with integrated transit transfers. *Transportation Research Part E: Logistics and Transportation Review*, 128(July):417–442, 2019, 1901.00760.

[110] Shukai Chen, Hua Wang, and Qiang Meng. Solving the first-mile ridesharing problem using autonomous vehicles. *Computer-Aided Civil and Infrastructure Engineering*, 35(1):45–60, 2020.

[111] Pramesh Kumar and Alireza Khani. An algorithm for integrating peer-to-peer ridesharing and schedule-based transit system for first mile/last mile access. *Transportation Research Part C: Emerging Technologies*, 122:102891, 2021.

[112] Krishna Murthy Gurumurthy, Kara M Kockelman, and Natalia Zuniga-Garcia. First-Mile-Last-Mile Collector-Distributor System using Shared Autonomous Mobility. *Transportation Research Record*, 0(0):0361198120936267, 2020.

[113] Michael W. Levin and Stephen D. Boyles. Effects of autonomous vehicle ownership on trip, mode, and route choice. *Transportation Research Record*, 2493:29–38, 2015.

[114] Akhil Vakayil, Wolfgang Gruel, and Samitha Samaranayake. Integrating shared-vehicle mobility-on-demand systems with public transit. Technical report, Transportation Research Board, Washington, D.C., 2017.

[115] Lucas Mestres Mendes, Manel Rivera Bennàssar, and Joseph Y.J. Chow. Comparison of light rail streetcar against shared autonomous vehicle fleet for Brooklyn–Queens connector in New York City. *Transportation Research Record*, 2650(1):142–151, 2017.

[116] Rounaq Basu, Andrea Araldo, Arun Prakash Akkinepally, Bat Hen Nahmias Biran, Kalaki Basak, Ravi Seshadri, Neeraj Deshmukh, Nishant Kumar, Carlos Lima Azevedo, and Moshe Ben-Akiva. Automated Mobility-on-Demand vs. Mass Transit: A Multi-Modal Activity-Driven Agent-Based Simulation Approach. *Transportation Research Record*, 2672(8):608–618, 2018.

[117] Yu Shen, Hongmou Zhang, and Jinhua Zhao. Integrating shared autonomous vehicle in public transportation system: A supply-side simulation of the first-mile service in Singapore. *Transportation Research Part A: Policy and Practice*,

113(March):125–136, 2018.

[118] Jian Wen, Yu Xin Chen, Neema Nassir, and Jinhua Zhao. Transit-oriented autonomous vehicle operation with integrated demand-supply interaction. *Transportation Research Part C: Emerging Technologies*, 97(January):216–234, 2018.

[119] M Salazar, F Rossi, M Schiffer, C H Onder, and M Pavone. On the Interaction between Autonomous Mobility-on-Demand and Public Transportation Systems. In *2018 21st International Conference on Intelligent Transportation Systems (ITSC)*, pages 2262–2269, nov 2018.

[120] Yang Liu, Prateek Bansal, Ricardo Daziano, and Samitha Samaranayake. A framework to integrate mode choice in the design of mobility-on-demand systems. *Transportation Research Part C: Emerging Technologies*, 105:648–665, 2019.

[121] Arthur Mahéo, Philip Kilby, and Pascal Van Hentenryck. Benders decomposition for the design of a hub and shuttle public transit system. *Transportation Science*, 53(1):77–88, 2019, 1601.00367.

[122] J. F. Campbell, A. T. Ernst, and M. Krishnamoorthy. Hub arc location problems: Part I - Introduction and results. *Management Science*, 51(10):1540–1555, 2005.

[123] J. F. Campbell, A. T. Ernst, and M. Krishnamoorthy. Hub Arc location problems: Part II - Formulations and optimal algorithms. *Management Science*, 51(10):1556–1571, 2005.

[124] P Manser. *Public Transport Network Design in a World of Autonomous Vehicles*. PhD thesis, Master thesis, 2017.

[125] Helen K.R.F. Pinto, Michael F. Hyland, Hani S. Mahmassani, and I. Ömer Verbas. Joint design of multimodal transit networks and shared autonomous mobility

fleets. *Transportation Research Part C: Emerging Technologies*, 113(June):2–20, 2020.

[126] Konrad Steiner and Stefan Irnich. Strategic Planning for Integrated Mobility-on-Demand and Urban Public Bus Networks. *Transportation Science*, 54(6):1616–1639, 2020.

[127] Jan-dirk Schmöcker, Hiroshi Shimamoto, and Fumitaka Kurauchi. Generation and Calibration of Transit Hyperpaths. *Procedia - Social and Behavioral Sciences*, 80:211–230, 2013.

[128] Jiangshan Ma and Daisuke Fukuda. A Hyperpath-based Network Generalized Extreme-value Model for Route Choice under Uncertainties. *Transportation Research Procedia*, 7:44–58, 2015.

[129] Mohammad Nurul Hassan, Taha Hossein Rashidi, S. Travis Waller, Neema Nassir, and Mark Hickman. Modeling transit user stop choice behavior: Do travelers strategize? *Journal of Public Transportation*, 19(3):98–116, 2016.

[130] N. Nassir, Mark Hickman, and Zhen Liang Ma. A strategy-based recursive path choice model for public transit smart card data. *Transportation Research Part B: Methodological*, 126:528–548, 2019.

[131] Neema Nassir, Mark Hickman, and Zhenliang Ma. Statistical inference of transit passenger boarding strategies from farecard data. *Transportation Research Record*, 2652(1):8–18, 2017.

[132] Alireza Khani, Mark Hickman, and Hyunsoo Noh. Trip-Based Path Algorithms Using the Transit Network Hierarchy. *Networks and Spatial Economics*, 15(3):635–653, 2014.

[133] Google. General Transit Feed Specification, 2005.

[134] Stephen Riter and Jan McCoy. Automatic vehicle location—An overview. *IEEE Transactions on Vehicular Technology*, 26(1):7–11, 1977.

[135] Dimitri P Bertsekas. *Dynamic Programming and Optimal Control*. Athena Scientific, Belmont, Massachusetts, fourth edition, 2012.

[136] Raymond K Cheung. Iterative methods for dynamic stochastic shortest path problems. *Naval Research Logistics (NRL)*, 45(8):769–789, 1998.

[137] Minnesota Department of Transportation. Mn/DOT Traffic Data, 2019.

[138] Peter Belenky. Revised Departmental Guidance on Valuation of Travel Time in Economic Analysis. Technical report, U.S. Department of Transportation, 2011.

[139] Heinrich von Stackelberg. *Market structure and equilibrium*. Springer Science & Business Media, 2010.

[140] R. Cominetti and J. Correa. Common-lines and passenger assignment in congested transit networks. *Transportation Science*, 35(3):250–267, 2001.

[141] Fabien Leurent, Ektoras Chandakas, and Alexis Poulhès. A traffic assignment model for passenger transit on a capacitated network: Bi-layer framework, line sub-models and large-scale application. *Transportation Research Part C: Emerging Technologies*, 47(P1):3–27, 2014.

[142] Mervat Chouman, Teodor Gabriel Crainic, and Bernard Gendron. Commodity Representations and Cut-Set-Based Inequalities for Multicommodity Capacitated Fixed-Charge Network Design. *Transportation Science*, 51(2):650–667, 2017.

[143] Jonathan F Bard. *Practical bilevel optimization: algorithms and applications*, volume 30. Springer Science & Business Media, 2013.

[144] J. F. Benders. Partitioning procedures for solving mixed-variables programming problems. *Numerische Mathematik*, 4(1):238–252, 1962.

[145] Paul A. Rubin. Benders Decomposition Then and Now, 2011.

[146] Egon Balas, Sebastián Ceria, and Gérard Cornuéjols. Mixed 0-1 Programming by Lift-and-Project in a Branch-and-Cut Framework. *Management Science*, 42(9):1229–1246, 1996.

[147] Matteo Fischetti, Domenico Salvagnin, and Arrigo Zanette. A note on the selection of Benders' cuts. *Mathematical Programming*, 124(1-2):175–182, 2010.

[148] Public Transit Networks for Research. Transit Network Design Instances for Research, 2021.

[149] Michael Laris. Uber and Lyft concede they play role in traffic congestion in the District and other urban areas, 2019.

[150] Jim Motavalli. Who Will Own the Cars That Drive Themselves?, 2020.

[151] Daniel J Fagnant, Kara M Kockelman, and Prateek Bansal. Operations of Shared Autonomous Vehicle Fleet for Austin, Texas, Market. *Transportation Research Record*, 2563(1):98–106, 2016.

[152] Michael W. Levin, Kara M. Kockelman, Stephen D. Boyles, and Tianxin Li. A general framework for modeling shared autonomous vehicles with dynamic network-loading and dynamic ride-sharing application. *Computers, Environment and Urban Systems*, 64:373–383, 2017.

[153] Md Aftabuzzaman, Graham Currie, and Majid Sarvi. Evaluating the Congestion Relief Impacts of Public Transport in Monetary Terms. *Journal of Public Transportation*, 13(1):1–24, 2015.

[154] T. Donna Chen and Kara M. Kockelman. Management of a shared autonomous electric vehicle fleet: Implications of pricing schemes. *Transportation Research Record*, 2572(2572):37–46, 2016.

[155] Baichuan Mo, Zhejing Cao, Hongmou Zhang, Yu Shen, and Jinhua Zhao. Dynamic Interaction between Shared Autonomous Vehicles and Public Transit: A Competitive Perspective. 2020, 2001.03197.

[156] OECD. Urban Mobility System Upgrade: How shared self-driving cars could change city traffic. *Corporate Partnership Board Report*, pages 1–36, 2015.

[157] Liteng Zha, Yafeng Yin, and Hai Yang. Economic analysis of ride-sourcing markets. *Transportation Research Part C: Emerging Technologies*, 71:249–266, 2016.

[158] George W Douglas. Price Regulation and Optimal Service Standards: The Taxicab Industry. *Journal of Transport Economics and Policy*, 6(2):116–127, 1972.

[159] Richard C Larson and Amedeo R Odoni. *Urban operations research*. Prentice-Hall, Englewood Cliffs, 2nd editio edition, 1981.

[160] Guy Desaulniers and Mark D. Hickman. Chapter 2 Public Transit. *Handbooks in Operations Research and Management Science*, 14(C):69–127, 2007.

[161] A M Geoffrion. Generalized Benders decomposition. *Journal of Optimization Theory and Applications*, 10(4):237–260, 1972.

[162] M Conforti, G Cornuejols, and G Zambelli. *Integer Programming*. Graduate Texts in Mathematics. Springer International Publishing, 2014.

[163] Georgios K.D. Saharidis and Marianthi G. Ierapetritou. Improving benders decomposition using maximum feasible subsystem (MFS) cut generation strategy. *Computers and Chemical Engineering*, 34(8):1237–1245, 2010.

[164] Lixin Tang, Wei Jiang, and Georgios K.D. Saharidis. An improved Benders decomposition algorithm for the logistics facility location problem with capacity expansions. *Annals of Operations Research*, 210(1):165–190, 2013.

[165] T L Magnanti and R T Wong. Accelerating Benders Decomposition: Algorithmic Enhancement and Model Selection Criteria. *Operations Research*, 29(3):464–484, 1981.

[166] N. Beheshti Asl and S. A. MirHassani. Accelerating benders decomposition: multiple cuts via multiple solutions. *Journal of Combinatorial Optimization*, 37(3):806–826, 2019.

[167] Yafeng Yin. Macroscopic modeling of ridesourcing systems - Regulations and Fundamental Diagram, 2019.

[168] Esteve Codina and Francisca Rosell. A heuristic method for a congested capacitated transit assignment model with strategies. *Transportation Research Part B: Methodological*, 106:293–320, 2017.