Developing the Analysis Methodology and Platform for Behaviorally Induced System Optimal Traffic Management

by

XIANBIAO HU

A Dissertation Submitted to the Faculty of the

DEPARTMENT OF CIVIL ENGINEERING AND ENGINEERING MECHANICS

In Partial Fulfillment of the Requirements

For the Degree of

DOCTOR OF PHILOSOPHY

In the Graduate College

THE UNIVERSITY OF ARIZONA

2013
THE UNIVERSITY OF ARIZONA
GRADUATE COLLEGE

As members of the Dissertation Committee, we certify that we have read the dissertation prepared by Xianbiao Hu, titled Developing the Analysis Methodology and Platform for Behaviorally Induced System Optimal Traffic Management and recommend that it be accepted as fulfilling the dissertation requirement for the Degree of Doctor of Philosophy.

______________________________________________ Date: (July 19th, 2013)
Yi-Chang Chiu

______________________________________________ Date: (July 19th, 2013)
Mark D. Hickman

______________________________________________ Date: (July 19th, 2013)
Larry Head

______________________________________________ Date: (July 19th, 2013)
Jorge Villalobos

Final approval and acceptance of this dissertation is contingent upon the candidate’s submission of the final copies of the dissertation to the Graduate College.

I hereby certify that I have read this dissertation prepared under my direction and recommend that it be accepted as fulfilling the dissertation requirement.

______________________________________________ Date: (July 19th, 2013)
Dissertation Director: Yi-Chang Chiu
STATEMENT BY AUTHOR

This dissertation has been submitted in partial fulfillment of the requirements for an advanced degree at the University of Arizona and is deposited in the University Library to be made available to borrowers under rules of the Library.

Brief quotations from this dissertation are allowable without special permission, provided that an accurate acknowledgement of the source is made. Requests for permission for extended quotation from or reproduction of this manuscript in whole or in part must be granted by the author.

SIGNED: Xianbiao Hu
First of all, my deepest gratitude goes to Dr. Yi-Chang Chiu, my advisor and committee chair. He was the main reason that I came to the University of Arizona and spent several wonderful years of my life in Tucson. I have benefited tremendously from him not only academically, but also from outside of research: the attitude towards work and life, the leadership and innovation he showed us all the way. Without his consistent encouragement, direction, patience and warm support, this dissertation would not have been possible.

I would like to express my sincere appreciation and gratitude to my dissertation committee: Dr. Mark Hickman, Dr. Larry Head and Dr. Jorge Villalobos for their instruction and guidance in this dissertation.

I am very grateful to all of my colleagues in the University of Arizona. I am truly lucky to have a group of wonderful friends, especially the members of the DynusT lab. They are Hong Zheng, Eric Nava, Ye Tian, Shuo Wang, Brenda Bustillos, Lei Zhu, Yohan Chang, Shin-Yu Lin, Michael Lin, Jason Freitas, Paul Hoffer. You all are part of my wonderful memories in UofA.

I would like to extend my deep appreciation to my parents for their emotional encouragement, unconditional love and support. They have always been there for me. My achievement wouldn’t be possible without all your support!

Last but not least, I would like to thank my wife, Jinxiang Wu for her love, understanding, and support. No other words would express my feeling for you better than “I love you” and “Thank you”, you are the one I want to cherish for all my life. I would like to dedicate this dissertation to her.
# TABLE OF CONTENTS

**LIST OF FIGURES** ........................................................................................................ 7

**ABSTRACT** ................................................................................................................ 8

1 **INTRODUCTION** ................................................................................................... 11

1.1 Problem Statement ................................................................................................ 11

1.2 Research Objectives ............................................................................................. 13

1.3 Uniqueness and Contributions of the Research .................................................. 14

1.4 Literature Review .................................................................................................. 16

1.4.1 ATDM Related Research ............................................................................... 16

1.4.2 Implementation Cases .................................................................................... 18

1.4.3 Incentives in ATDM ...................................................................................... 21

1.4.4 OD Calibration Research .............................................................................. 24

1.4.5 K Shortest Path Research ............................................................................. 25

1.4.6 System Optimal Research ............................................................................ 27

1.5 Summary of Literature ......................................................................................... 29

2 **PRESENT STUDY** .................................................................................................. 32

2.1 Dissertation Outline ............................................................................................. 32

2.1.1 Research Methodology for Simulation Model Calibration ......................... 32

2.1.2 Research Methodology for System Optimal Approach ............................. 33

2.1.3 Research Methodology for Traveler Route Choice Model ....................... 34
# TABLE OF CONTENTS - Continued

2.2 Comprehensive Case Study .............................................................................................................. 36

2.2.1 Case Study Framework .................................................................................................................. 36

2.2.2 Case Study Result Analysis .......................................................................................................... 39

2.3 Conclusions ...................................................................................................................................... 45

REFERENCES ......................................................................................................................................... 47

APPENDICES ......................................................................................................................................... 54

APPENDIX A: A DECOMPOSITION FRAMEWORK AND METHOD FOR CALIBRATING DYNAMIC ORIGIN-DESTINATION MATRICES FOR DYNAMIC TRAFFIC ASSIGNMENT UNDER CONGESTED TRAFFIC CONDITIONS ............................................. 55

APPENDIX B: DEVELOPMENT OF A BEHAVIORALLY INDUCED SYSTEM OPTIMAL MODEL FOR ACTIVE TRAFFIC AND DEMAND MANAGEMENT SYSTEM ........................................................................................................... 107

APPENDIX C: A CONSTRAINTED TIME-DENDPENT K SHORTEST PATHS ALGORITHM ADDRESSING OVERLAP AND TRAVEL TIME DEVIATION .............. 169
LIST OF FIGURES

Figure 2-1 Work flow of the integrated framework ................................................................. 37
Figure 2-2 Baseline Scenario Heat Map – NB (Baseline versus comprehensive scenario) .......... 40
Figure 2-3 Baseline Scenario Heat Map – SB (Baseline versus comprehensive scenario) .......... 40
Figure 2-4 Travel behavior changes ..................................................................................... 41
Figure 2-5 Total system travel time comparison ..................................................................... 43
ABSTRACT

Traffic congestion has been imposing a tremendous burden on society as a whole. For decades, the most widely applied solution has been building more roads to better accommodate traffic demand, which turns out to be of limited effect. Active Traffic and Demand Management (ATDM) is getting more attention recently and is considered here, as it leverages market-ready technologies and innovative operational approaches to manage traffic congestion within the existing infrastructure.

The key to a successful Active Traffic and Demand Management strategy is to effectively induce travelers’ behavior to change. In spite of the increased attention and application throughout the U.S. or even the world, most ATDM strategies were implemented on-site through small-scale pilot studies. A systematic framework for analysis and evaluation of such a system in order to effectively track the changes in travelers’ behavior and the benefit brought about by such changes has not been established; nor has the effect of its strategies been quantitatively evaluated.

In order to effectively evaluate the system benefit and to analyze the behavior changes quantitatively, a systematic framework capable of supporting both macroscopic and microscopic analysis should be established. Such system should be carefully calibrated to reflect the traffic condition in reality, as only after the calibration can the baseline model be used as the foundation for other scenarios in which alternative design or management strategies are incorporated, so that the behavior changes and system benefit can be computed accurately by comparing the alternative scenarios with the baseline scenario.

Any effective traffic management strategy would be impossible if the traveler route choice behavior in the urban traffic network has not been fully understood. Theoretical research assumes
all users are homogeneous in their route choice decision and will always pick the route with the shortest travel cost, which is not necessarily the case in reality. Researchers in Minnesota found that only 34% of drivers strictly traveled on the shortest path. Drivers’ decision is made usually based on several dimensions, and a full understanding of the travel route choice behavior in the urban traffic network is essential.

The existence of most current Advanced Traveler Information Systems (ATIS) offer the capability to provide pre-trip and/or en route real time information, allowing travelers to quickly assess and react to unfolding traffic conditions. The basic design concept is to present generic information to drivers, leaving drivers to react to the information their own way. This “passive” way of managing traffic by providing generic traffic information has difficulty in predicting outcome and may even incur adverse effect, such as overreaction (aka herding effects). Furthermore, other questions remain on how to utilize the real-time information better and guide the traffic flow more effectively towards a better solution, and most current research fails to take the traveler's external cost into consideration.

Motivated by those concerns, in this research, a behaviorally induced system optimal model is presented, aimed at further improving the system-level traffic condition towards System Optimal through incremental routing, as well as establishing the analysis methodology and evaluation framework to calibrate quantitatively the behavior change and the system benefits. In this process, the traffic models involved are carefully calibrated, first using a two-stage calibration model which is capable of matching not only the traffic counts, but also the time dependent speed profiles of the calibrated links. To the best of our knowledge, this research is the first with a methodology to incorporate the use of field observed data to estimate the Origin-Destination (OD) matrices departure profile. Also proposed in this dissertation is a Constrained K Shortest Paths algorithm (CKSP) that addresses route overlap and travel time deviation issues. This proposed
algorithm can generate K Shortest Paths between two given nodes and provide sound route options to the drivers in order to assist their route choice decision process. Thirdly, a behaviorally induced system optimal model includes the development of a marginal cost calculation algorithm, a time-dependent shortest path search algorithm, and schedule delay as well as optimal path finding models, is present to improve the traffic flow from an initial traffic condition which could be User Equilibrium (UE) or any other non-UE or non-System-Optimal (SO) condition towards System Optimal. Case studies are conducted for each individual research and show a rather promising result.

The goal of establishing this framework is to better capture and evaluate the effects of behaviorally induced system optimal traffic management strategies on the overall system performance. To realize this goal, the three research models are integrated in order to constitute a comprehensive platform that is not only capable of effectively guiding the traffic flow improvement towards System Optimal, but also capable of accurately evaluating the system benefit from the macroscopic perspective and quantitatively analyzing the behavior changes microscopically. The comprehensive case study on the traffic network in Tucson, Arizona, has been conducted using DynusT (Dynamic Urban Simulation for Transportation) Dynamic Traffic Assignment (DTA) simulation software; the outcome of this study shows that our proposed modeling framework is promising for improving network traffic condition towards System Optimal, resulting in a vast amount of economic saving.
1 INTRODUCTION

1.1 Problem Statement

Traffic congestion has been imposing a tremendous burden on society as a whole. In the US, congestion costs were about $115 billion in 439 urban areas in the year of 2010, according to a Texas Transportation Institute (TTI) report (Schrank et al. 2011). The European UNITE project (de Palma and Lindsey 2011) estimated the costs of traffic congestion in the UK, France and Germany to be respectively 1.5%, 1.3% and 0.9% of GDP. In Australia, the Bureau of Infrastructure, Transport and Regional Economics has estimated that urban congestion alone will cost nearly AU$20 billion by 2020 (Low and Odgersb 2012). A researcher in Korea found the socio-economic cost caused by traffic congestion is now approximately $45.9 billion per year plus 336,000 job losses (Jun 2011).

For decades, the most widely applied solution has been building more roads to better accommodate traffic demand, which turns out to be of limited effect: infrastructure construction and maintenance are very costly, and yet more demand is induced, and that added demand quickly saturates the newly built highways. The increasingly considered strategy called Active Traffic and Demand Management (ATDM), on the contrary, aims at a better balance between the need to travel a particular route at a particular time and the capacity of available facilities to handle this demand efficiently. ATDM is being considered by the Federal Highway Administration (FHWA) (Luten et al. 2004) as market-ready technologies and innovative operational approaches that are becoming available for managing traffic congestion within the existing infrastructure.

ATDM refers to the wide-range and diversified characteristics of traffic management strategies deployed by transportation agencies that focus on using advanced technologies, available
information, and valid methodologies, in order to monitor and manage the dynamic traffic conditions actively, and to influence people's need and intention to travel as well as their associated travel pattern, for the purpose of promoting efficient use of existing roadway systems and better handling of the vehicle demand. ATDM can be considered the combination of travel demand management (TDM) and active traffic management (ATM) (Zheng et al. 2011). ATDM offers significant potential for reducing traffic congestion without the need to build additional lanes or infrastructure. The vision for ATDM research is to allow transportation agencies to increase traffic flow, improve travel time reliability, and optimize available capacity throughout the transportation network (Cronin and Sheehan 2012).

The key to a successful Active Traffic and Demand Management is to effectively induce people’s travel behavior change through various available traffic management or demand management strategies, and thereby in the end achieve a better traffic condition and reduce the congestion. Some of the commonly seen ATDM strategies include pricing, dynamic parking management, freeway access control, dynamic information, incentive-based strategy and so on. Those strategies produce their effect by changing trip-makers' travel behavior, prior to departure or en route, in various respects, such as people's destination choice, departure time, transportation mode, routing decision, etc., to reach the goal of evenly using the existing transportation facility and better handling the traffic demand within the traffic network.

In spite of the increased attention and application throughout the U.S. or even the world, most ATDM strategies were implemented on-site through small-scale pilot studies. A systematic framework for analysis and evaluation of such system to effectively track the changes in people’s behavior and the benefit brought about by such changes has not been established, nor the effect of its strategies quantitatively evaluated. Although there is no doubt that the existence of such
framework can play a critical role in the policy-makers’ decision making process, and help evaluate and adopt the most cost-effective strategies under various traffic management goals and priorities.

In order to evaluate the system benefit effectively and analyze the behavior changes quantitatively, a systematic framework capable of supporting both macroscopic and microscopic analysis should be established. In this framework, the traffic models involved need to be carefully calibrated to reflect the traffic condition in reality and accurately capture the behavior changes and system benefit. A full understanding of the travel route choice behavior in the urban traffic network is essential for a successful strategy to be implemented, and such strategy should be able to effectively guide the improvement of traffic flow from baseline scenario to a better traffic condition.

1.2 Research Objectives

The goal of establishing this framework is to better capture and evaluate the effects of behaviorally induced system optimal traffic management strategies on the overall system performance. In order to realize this goal, the research objectives can be further divided into the following aspects:

- To better calibrate an offline simulation model to be used as the foundation for other scenarios in which alternative design or management strategies are incorporated. Such calibration model should be capable of calibrating not only the traffic counts but also the congestion temporal profile pattern at the bottleneck of interest.

- To take better advantage of the available real time traffic information provided by ATIS, further improve the system level traffic condition from UE, or any other non-UE or non-SO condition, towards System Optimal, and avoid passively managing traffic which may incur adverse effect such as herding effects.
• To better understand and model the heterogeneous travel behavior and assist drivers’ route choice decision process by providing sound route options, during which the overlap and travel time deviation issues between the K paths need to be considered.

• To develop a comprehensive platform which is not only capable of effectively improving the traffic flow towards system optimal condition, but also capable of accurately evaluating the system benefit from the macroscopic perspective and quantitatively analyzing the behavior changes microscopically.

1.3 Uniqueness and Contributions of the Research

The originality of the contribution of this dissertation rests in the development of this Behaviorally Induced System Optimal analysis framework and methodology. More specifically, this dissertation, which is proposed to satisfy a requirement for the degree of Doctor of Philosophy, makes an original contribution in the following four respects:

• Propose a bi-level OD calibration algorithm to build and calibrate the offline simulation model, which is not only able to match the traffic counts but the time-dependent speed profile as well in a large network.

This research presents a bi-level OD Calibration model that is capable of calibrating time-dependent origin-destination matrices in order to match not only the traffic counts but also the time-dependent speed profile of the calibrated links. The method is based on understanding the demand-supply relationship at the bottleneck and the utilization of shockwave and travel time propagation between origin and bottleneck. To the best of our knowledge, this research is the first with a methodology to incorporate the use of field observed data to estimate the OD matrices departure profile. The numerical results in a complicated network demonstrate the potential of the proposed methodology.
• **Build a Model with the aim of improving the traffic condition towards System Optimal through incremental routing change that considers marginal impact brought by the additional vehicles.**

A model approach is presented in this dissertation to improve the traffic from UE or any other non-UE or non-SO condition towards System Optimal. The originality of this part of the research is to consider System Optimal as the objective from the overall standpoint of the system as a whole, which distinguishes itself from most of the other ATDM researches that did not explicitly consider the marginal cost to the system when assigning multiple drivers to the system. The incremental routing mechanism in this model is also able to avoid the commonly seen herding effects, which are results of the “passive” way of managing traffic by providing generic traffic information to all drivers and which make it difficult to predict outcomes of such strategy. A numerical analysis is conducted on the Tucson I-10 corridor and the outcome of the case study shows that the benefit of our proposed algorithm includes significant traffic congestion alleviation, travel time and monetary saving with a relative low cost.

• **Develop a Constrained Time-Dependent K-Shortest Path (CKSP) algorithm to find K-Shortest Paths (KSP) between origin and destination nodes within a given network that exhibit a certain similarity; the result is capable of intelligently avoiding finding highly overlapped routes or unreasonably detoured routes, or routes with high travel time deviation.**

The K-Shortest paths of interest should be neither highly overlapped nor unreasonably detoured; otherwise, such paths will fail to satisfy the objective of providing sound alternative routes for the individual driver. The travel time of the paths should not deviate too much either, as otherwise the paths except for the first one would be of limited usefulness to the drivers. Addressing the overlap and travel time deviation issue when finding KSP for the individual driver are together
the major originality in the contribution of this part of the research. A Constrained K Shortest Paths algorithm will be presented in this research to find k time-dependent shortest paths between two given nodes with the reasonable similarity and travel time deviation degree, where the term "similarity" is defined to show the overlapping degree between paths. The outcome of the case study over the Tucson network shows the result of the proposed algorithm is quite satisfactory.

- Integration of the three research modules into a comprehensive platform to be able to better capture and evaluate the effects of behaviorally induced system optimal traffic management strategies on the overall system performance.

The goal of establishing this framework is to better capture and evaluate the effects of behaviorally induced system optimal traffic management strategies on the overall system performance. To realize this, the three research modules are integrated to be a comprehensive platform, which is not only capable of effectively guiding the traffic flow improvement towards System Optimal, but also capable of accurately evaluating the system benefit from the macroscopic perspective and quantitatively analyzing the behavior changes microscopically. The comprehensive case study on the traffic network in Tucson, Arizona, has been conducted using DynusT (Dynamic Urban Simulation for Transportation) Dynamic Traffic Assignment (DTA) simulation software and demonstrates a satisfactory result.

1.4 Literature Review

1.4.1 ATDM Related Research

The general sense of a need for Active Traffic and Demand Management has a long history which can be traced back to the 1970s and 1980s and stems from legitimate desires to provide alternatives to single-occupancy commuter travel – for reasons of energy conservation, improvement of air quality, and reduction of peak-period congestion. Today, the need to manage
travel demand has broadened to encompass the desire to optimize transportation system performance for both commute and non-commute types of trips, and for both recurring as well as non-recurring events (Luten et al. 2004).

In 2004, the Federal Highway Administration (FHWA) proposed the idea that the demand-side strategies should be designed to better balance the need to travel a particular route at a particular time with the capacity of available facilities to handle this demand efficiently (Luten et al. 2004). The U.S. Department of Transportation further defined ATDM as market-ready technologies and innovative operational approaches for managing traffic congestion within the existing infrastructure (Cronin and Sheehan 2012). In another FHWA report, ATDM is defined as dynamic management, control, and influence of travel demand, traffic demand, and traffic flow on transportation facilities (Battelle et al. 2012). The purview of ATDM is to combine travel demand management (TDM) and active traffic management (ATM), to exert influence actively on the need to travel, as well as on the associated travel pattern, for the purpose of promoting efficient use of the roadway system that is handling the vehicle demand (Zheng et al. 2011). The vision for Active Traffic and Demand Management research is to allow transportation agencies to increase traffic flow, improve travel time reliability, and optimize available capacity throughout the transportation network (Cronin and Sheehan 2012).

Active management of transportation and demand can include multiple approaches, spanning demand management, traffic management, parking management, and efficient utilization of other transportation modes and assets (Battelle et al. 2012).

The dimension of demand-side strategies includes: *Mode choice* such as from driving alone to carpool, vanpool, public transit or other alternatives – researches in this category include (COMSIS 1996; Loukopoulos et al. 2004; Garling and Schuitema 2007); *Route choice* such as alternative roadway routes, alternative mode routes assisted with real-time route information, in-
vehicle navigation, web-based route-planning tools, etc. – relevant researches are (Small and Yan 2001; Yildirim and D.W.Hearn 2005); Departure time choices to change the time of day or day of week to travel, the typical strategies include worksite flex-time and coordinated event or shift scheduling (Ben-Akiva et al. 1986; Arnott et al. 1990); and trip reduction choices taking advantage of telework or compressing work week schedules (Eriksson et al. 2010).

On the traffic management side, various strategies have been applied to address both recurrent and non-recurrent congestions. Some of the most commonly seen strategies include: Speed harmonization on a freeway to actively manage the network and delay the onset of congestion under normal operating conditions (Geza Pesti et al. 2008; Waller et al. 2009), temporary shoulder use to increase capacity temporarily during peak hour (Curren 1995; Geistefeldt 2009), queue warning messages about the presence upstream of queues via use of dynamic traffic detection (Khan 2007), and dynamic merge control for dynamic metering or for closing specific upstream lanes (McCoy and Pesti 2001; Kang et al. 2006). Other common strategies include construction site management, truck restrictions, dynamic rerouting and driver information, dynamic lane marking, automated enforcement and so on. Those strategies are believed to be capable of moving the United States toward comprehensive active traffic management to manage congestion (Mirshahi et al. 2007).

1.4.2 Implementation Cases

For a long time, government agencies, organizations or even companies have been seeking ways to alleviate traffic congestion by actively adjusting drivers’ behavior through different strategies. Several examples will be listed in this section, but it is impossible to enumerate all implementation cases.
Lee County, FL used variable Bridge Tolls to manage the traffic congestion, and was able to spread traffic away from the peak period. Over the peak hours, the tolls on two principal bridges were raised from 75 cents to one dollar, and for the off-peak time a 50-cent discount was provided. With the $9.7 million grant from the Federal Highway Administration and another $7 million "emergency revenue reserve", 5% of traffic was shifted from peak to off-peak time periods, although as many as half of respondents indicated they always or sometimes considered the discounts prior to making a trip across one of the bridges (Burris 2001). Similar dynamic pricing strategy can also be found in Seattle, Los Angeles, San Diego, Atlanta, where dynamic pricing of High Occupancy/Toll (HOT) lanes and incentives for High Occupancy Vehicle (HOV) usage are provided (Texas Transportation Institute 2010).

The University of Washington at Seattle engaged in the U-Pass program, which started in the 1990s and aimed to alleviate the traffic problem both around the campus and within the city. The program provided an array of transportation options, including transit service, preferential parking, consumer discounts, and rideshare matching. The benefits of this program have been significant: the building of 3,600 new parking spaces has been avoided, 86% user satisfaction has been reported, and a reduction of 33% has been seen in parking permit purchases, plus 18% less in traffic counts during morning peak hour in 2002, compared with the data in 1983 (University of Washington 2002).

The reconstruction of I-15 aimed to resolve traffic congestion and provide an efficient corridor for the 2002 Olympic Games in Salt Lake City, Utah. Besides adding two general lanes and two high occupancy vehicle lanes, demand-side strategies were used to maintain traffic during reconstruction, and an Intelligent Transportation System was installed to monitor the impact of reconstruction on the traffic, to respond to traffic accidents in a faster way, and to communicate and coordinate with the drivers. Due to the I-15 corridor improvement, peak hour freeway volume
was increased by 20% during the Olympic Games period, and intersections delay decreased by 27% (FHWA 2006).

SmartRide is a software tool developed by Georgia Power to provide a user friendly way to track carpool, vanpool and transit use information and direct the information to its employees internally, in order to get as many cars off the roads as possible during the 1996 Olympic Games. The company also provided fleet vehicles to its employees to run errands or go to meetings during the day, so those who used alternative transportation to commute to and from work would not feel trapped in the office during the day. This program successfully reduced 1.2 million Vehicle Miles Traveled (VMT) each month, with 13% of employees carpooling or vanpooling, 15% working compressed or flex-time schedules, 5% using transit and another 5% doing telework (FHWA 2006).

The millions of visitors to Zion National Park brought traffic congestion and parking problems. Started in 2000, the propane-powered, shuttle system vehicles became the only vehicles permitted in the park during the peak season. Visitors are encouraged to park outside and use the shuttle free of charge to access the national park. Shuttle boarding reported for year 2002 was 2.35 million, and about 75% of the visitors ride the Zion Canyon shuttle. Besides the benefits from traffic and parking aspects, the local government benefits from this program as well, not only from the reduction in infrastructure cost but also from the economic growth (NPS 2003).

The transit-oriented development in Hillsboro, in the Portland Metropolitan area in Oregon, has made it one of the most liveable communities in the U.S. It features a pedestrian friendly environment and mixed-use development in the station area which made it possible for the residents to walk to the light rail station. New tenants are offered a one-year transit pass free of charge. Due to this program 83% of residents reported they used transit in May 1999, and transit usage for commuting had increased 22%. It was also reported that 85% of respondents stated they had reduced needs to drive because of the program (Podobnik 2002).
With the advances in wireless communication and computer technology, new ATDM applications have been emerging rapidly during the last several years, especially the applications based on the popular Smartphone platform. For example, the dynamic ridesharing iOS application in Cork, Ireland, brings drivers and passengers together in real time, extending the public transportation network using private automobiles, with the features of dynamic ride-matching to match available seats en-route and the integration of an electronic wallet to support the pricing mechanism (Transportation for America 2010). TELE-Bus in Krakow, Poland, provides on-demand transit service, based on the Dynamic Dispatch and Routing platform, in order to attract new transit passengers and help build more effective bus fleet management, – and it ended up with a 600% increase of the transit passengers after a half year's implementation (AENEAS 2010). The MITTENS (Messaging Infrastructure for Travel Time Estimates to a Network of Signs) in San Francisco provides real-time freeway and CalTran travel-times information to motorists in order to induce en-route mode shifts; the internal algorithm is deployed to determine travel time by using available information and internal logic (Sharafsaleh et al. 2011). In the San Francisco area, the Predict-a-TripSM was built based on the popular 511 Driving TimesSM service by using historical information on freeway traffic speeds and driving times to provide point-to-point forecasts for about 90 percent of the Bay Area freeway network; drivers can access this info via the phone, web or freeway message signs. It is helpful for motorists to obtain route information not only for planning trips, but also for them to consider taking public transit instead of driving, or to use the ride-matching tool for prospective carpools (Goodwin 2007).

1.4.3 Incentives in ATDM

Road pricing has been argued by the transport economists as one of the most efficient strategies to alleviate congestion externalities, although the practice is controversial and its
behavior implications are not well understood (Ben-Elia and Ettema 2011). Recently, researchers have begun to reverse the idea and use incentives (rather than a penalty) to influence driver behavior. The effectiveness of rewards to reinforce a desirable behavior is supported by a large volume of empirical evidence (Kreps 1997; Berridge 2001); however, the implementation and relevant research on incentives in the transportation area has a relatively short history.

An experiment with respect to 43 drivers was carried out during an 8-day temporary freeway closure in Osaka, Japan. Of the drivers, 23 were offered a one-month free bus ticket, but the free ticket was not given to the other 20 drivers in a control group. The result shows that drivers who received a one-month free bus ticket in the experiment used the bus more frequently after the intervention. The increase was 20% higher than the frequency of bus use before, although it turns out in this experiment that these behavior changes are not permanent (Fujii and Kitamura 2003).

A longitudinal study in Germany investigated the effects of an intervention on increased bus use among college students by offering the incentive of a pre-paid bus ticket. The intervention was found to influence attitudes toward bus use, subjective norms, and perceptions of behavioral control. The bus usage has increased by 35% and trips finished by car decreased by 13%, which indicates the incentive given to the drivers can influence their travel behavior as compared to their previous habitual behavior (Damberg et al. 2002).

In Melbourne, Australia, an early bird ticket program was proposed to alleviate the rail overcrowding issue during the peak hours. Free rail fares were provided for rail drivers completing their trip before 7:00am as incentives to shift demand from the peak of the peak to relieve the overcrowding problem. The program cost $6 million mainly in lost fares, and it ended up that 23% of ticket users (2,000-2,600 each week) had shifted from the peak hour by an average of 42 minutes; the ticket use was increased by 1.7% while overall rail usage was stable. After introduction of the early bird program, pre-7:00a.m loads increased 41%, reducing pressures for purchase/operations
of new peak trains saving 2.5–5.0 trains (2008) to a forecast saving of 8.05 trains in 2038. Financial analysis suggested the savings would substantially cover the financial costs of the free fares and that benefits would increase over time (Currie 2011).

In a 13-week field study conducted in The Netherlands, 340 participants were provided with daily rewards – monetary and in-kind, in the second half of 2006, in order to encourage them to avoid driving during the morning rush-hour. Participants could earn a reward (either money or credits to keep a Smartphone handset) by driving to work earlier or later, by switching to another mode or by teleworking. Results provided evidence of substantial behavior changes in response to the rewards, with commuters shifting to earlier and later departure times and more use of public transport and alternative modes or working from home. The researcher also found the choice as to how to change behavior was influenced by additional factors as well, including education, scheduling, habitual behavior, attitudes, and travel information availability (Ben-Elia et al. 2011; Ben-Elia and Ettema 2011).

Stanford University also used the idea of incentives to manage the parking problem in Palo Alto, California, through Dynamic Parking Pricing. The project was called Congestion and Parking Relief Incentives (CAPRI). Credits for avoiding peak parking hours were awarded to the drivers who enter and exit the main Stanford University campus by car at designated off-peak hours Monday through Friday. The credits can be used to win random cash rewards from $2 to $50 over and over again. In the end, the rewards accumulated will be disbursed monthly via Stanford's payroll or through bank deposits (Stanford University 2012). The problem of this program, however, is that it cannot distinguish the users who actually shifted from peak hour from those who were already accustomed to travel during the off-peak time periods.
1.4.4 OD Calibration Research

There is a rich body of research into the OD calibration problem. Most of the earlier studies focus on estimating an OD matrix with a constant proportion or splitting rate matrix assuming free-flow traffic condition (Carey et al. 1981; Bell 1983; Bierlaire and Toint 1995). Another category of research considers congestion by incorporating traffic assignment to determine the OD proportion or splitting rate as related to the equilibrium condition (Yang et al. 1992; Florian and Chen 1995; Chang and Tao 1999; Tavana 2001; Zhou et al. 2003). A commonly seen model formulation is the least-square bi-level formulation, in which the upper level problem is to minimize some weighted measures of deviation from the target matrix and from the observed counts (Yang et al. 1992; Cascetta et al. 1993; Florian and Chen 1995; Sherali and Park 2001; Tavana 2001; Zhou et al. 2003; Chiu et al. 2007). Most of the above studies focus on the development of model formulation and solution algorithm, but none of which was tested on a large real-life network with hundreds or thousands of zones.

Another area of concern with OD estimation is real-time traffic estimation and prediction. An OD matrix calibration of DYNASMART-X was performed using multi-day peak-hour freeway traffic counts collected in Irvine, CA based on the adjustment of the multi-day peak-hour OD using a least-square formulation (Zhou 2003) and Kalman Filtering (Zhou and Mahmassani 2007). Several studies have focused on the calibration of DynaMIT (Ben-Akvia et al. 1998; Park et al. 2006; Wen et al. 2006), in which both the traffic flow models and OD demand were adjusted in real time according to the incoming traffic data.

In addition to OD calibration, Mahut et al. (2004) discussed the calibration of DTASQ (now called DYNAMEQ) on a suburban network of Calgary, Canada. The calibration effort focused on a technique to expand the Traffic Analysis Zone (TAZ) to enhance zonal representation of the study area, calibration of the peak-hour OD matrix, car-following and lane-changing model
parameters, and the fine-tuning of intersection turning movements. Apart from the above studies of relatively small networks, Ziliaskopoulos et al. (Ziliaskopoulos et al. 2004) attempted to address the implementation of large-scale DTA applications. They addressed the manipulation of network and demand data, the modeling of turning movements, the efficient computation of link travel times, and the handling of complex path data. Although this study did not explicitly address the model calibration procedure, the discussions provided valid and useful insights into the application of the DTA model to long-term planning.

The studies described above highlight that the DTA models, with modest effort, can be reasonably calibrated for project-level peak-hour applications. To date however, few attempts have been made to apply DTA in a long-range planning process context, due to computational intractability for long-period simulation and assignment, and to lack of OD calibration methods effective for large-scale networks. It should also be noted almost all prior studies use traffic counts as the calibration target so as to determine the model formulation, which could be problematic.

1.4.5 K Shortest Path Research

Various literature can be found in the K-Shortest Paths research area. Martins published his Shortest Paths Ranking algorithm as early as 1983 (Martins 1983) on how to find the K-Shortest Paths between an initial and a terminal node in a network and how then to rank them according to their cost. Martins continued to be devoted into research in the ranking algorithm field (Azevedo et al. 1990; Azevedo et al. 1993; Deazevedo et al. 1994) to improve the computation efficiency. Researches about the Node-Disjoint Shortest Paths can be found in several studies (Frish 1967; Steiglite 1971; Suurballe 1974), where nodes are not allowed to repeat in the different paths found. Researches about the Edge-Disjoint Shortest Paths can be found in a number of studies (Y. Perl 1978; Suurballe 1982; Suurballe and Tarjan 1984), with the constraint of no repeating links. Yen
(Yen 1971) also proposed a K-Shortest Path algorithm to find K loopless paths; the significance of his algorithm was in having computation of the upper bound increase only linearly with the value of K. Other literature reports work by others (Rosen et al. 1991; Chen 1993; Chen and Hung 1994; Chen 1994), who studied how to find the K-Shortest paths when the capacities constraint exists on the links between two nodes. Eppstein (Eppstein 1994) developed an algorithm to find K shortest path through a shortest path tree and a graph to represent all possible deviations from the shortest path, and later Jimenez (Jimenez and Marzal 2003) modified this algorithm to improve its practical performance. Many other researchers also studied K-Shortest Paths algorithms from different perspectives; see (Dreyfus 1969; Wongsenashote 1976; Horne 1980; Katoh et al. 1982; Perko 1986).

Modeling of route choice is an area that receives increased attention and involves KSP research and application, and can usually be found in Stochastic User Equilibrium, traffic assignment, OD calibration and other research fields (Prashker and Bekhor 2004). A straightforward and conventional way of finding these paths is to compute a sufficiently large number of overall shortest paths, i.e., the path enumeration method, and select ones that satisfy the constraints or delete the ones that do not satisfy the constraints. With reference to route choice, the most commonly seen options include Multinomial Logit (MNL) and probit model (Ben Akiva and S.R. 1985; Sheffi 1985; Cascetta 1990; Ortuzar and L.G. 1990). Cascetta proposed a modified specification of the logit model (C-Logit) which overcomes the main shortcoming of MNL, i.e., the overlap issue, by introducing a "commonality factor" in the logit utility function (Cascetta et al. 1996). However, as noted by the author, the application of the C-Logit model requires the explicit enumeration of alternative paths and the algorithm might work better if a limited number of reasonable paths, i.e., paths of comparable costs, is generated.
However, path enumeration approaches are time consuming and CPU expensive, researchers have been trying to deal with the route choice model while avoid excessive computational burden. The Probit Stochastic Method (SAM) proposed by Maher and Hughes (1997) does not suffer from the path enumeration by its assumption of a 'Markovian' routing strategy, and the capacity restraint was also considered in the SAM model. Castillo and other researchers (Castillo 2008; Castillo et al. 2008) used Bayesian network in his OD calibration model and traffic flow prediction model also without path enumeration. Zijpp and Catalano (2003) proposed a method to find the feasible shortest paths directly and in combination with predefined constraints, by using the ordinary shortest path computation as its elementary operation. Other non-Enumeration approach can be referred to (Bell 1995; Akamatsu 1996; Damberg et al. 2002; Qin et al. 2005).

1.4.6 System Optimal Research

Wardrop stated his first principle of choice "The journey times in all routes actually used are equal and less than those which would be experienced by a single vehicle on any unused route," which is generally referred to as "User Equilibrium" (UE), and his second principle "At equilibrium the average journey time is minimum," which is usually referred to as "system optimal" (SO) (Wardrop 1952). On the one hand, the UE condition is achieved when all the drivers have perfect information about traffic conditions and make the optimal decision independently, based on their own interests. On the other hand, SO would require that a system have control over all vehicles' routing decisions, and make decisions to minimize the system total travel time as a whole, which is more difficult to implement but superior to UE flow.

Some of the most famous research on SO includes the M-N model by Merchant and Nemhauser (Merchant and Nemhauser 1978; Merchant and Nemhauser 1978), Ziliaskopoulos's linear programming model for the Single Destination System Optimal Dynamic Traffic
Assignment (SODTA) Problem (Ziliaskopoulos 2000) and Daganzo's research based on the Cell Transmission Model (CTM) (Daganzo 1994; Daganzo 1995), as well as plenty of others, such as the deterministic queuing assignment model proposed by Ghali and Smith (Ghali and Smith 1995), Peeta and Mahmassani's formulation of two dynamic network traffic assignment models incorporating a traffic simulation model within an overall iterative search framework (Peeta and Mahmassani 1995), and so on.

A rather wide and diversified body of research about Marginal Cost exists in the literature, studying various aspects including theory, calculation, application, etc. In fact, depending upon the research objective, research on Marginal Cost exists in several research areas, such as Congestion Pricing and System Optimal Dynamic Traffic Assignment (SODTA), Transportation Policy and Decision Making, e.g., consideration of a Vehicle Miles Traveled tax. One of the most popular areas in Marginal Cost research looks into the marginal travel time contribution of an additional unit of flow into the system. Ghali and Smith (Ghali and Smith 1992) defined four levels of Marginal Cost from the perspective of queuing delays. The first level is the travel time on that link that an additional driver experiences; the second level includes the additional delay experienced by all vehicles that traverse that link after that vehicle. The third level extends the definition of the second level to include the additional delay to all vehicles (on all links) whose path travel times are increased due to that vehicle, and the fourth level defines the global link marginal travel time brought about by that additional vehicle. Peeta (Peeta 1994) also has a similar definition for marginal travel time in this category from several perspectives: static or time-dependent; global or local, and link-level or path-level.
1.5 Summary of Literature

Calibrating a traffic simulation model is a necessary step in a modeling exercise and evaluation process. Only after the calibration can the baseline model be used as the foundation for other scenarios in which alternative design or management strategies are incorporated. Almost all the prior studies use traffic counts as the calibration target in order to determine the model formulation, which could become problematic, particularly in failing to estimate OD matrices so as to produce a plausible simulated congestion pattern that matches the field data. This deficiency stems from the fact that count data is not the ideal descriptor of congestion.

Travel demand, as an aggregated terminology, emerges from individual decisions. Any effective ATDM strategy would be impossible if the driver route choice behavior in the urban traffic network has not been fully understood. The travel decision is made usually based on several dimensions, such as drivers’ individual objectives, constraints, preferences, experiences and knowledge about the network and the traffic condition. Moreover, drivers' decisions are both heterogeneous and evolutionary. Theoretical research assumes all users are homogeneous in their route choice decision and will always pick the route with the shortest travel cost, which is not necessarily the case in reality. Researchers in Minnesota found that only 34% of drivers strictly traveled on the shortest path; even if the standard was relaxed to allow 10% deviation spatially, the number was increased merely to 40% (Zhu and Levinson 2012).

It has been long recognized that, following Wardrop’s second principle, the total system travel cost under System Optimal (SO) would be lower than under a User Equilibrium (UE) flow. Minimizing the travel cost from the individual driver's standpoint versus minimizing the total system cost brought about by the additional individual drivers is the essential difference between Wardrop's two principles. However, almost all current Advanced Traveler Information System
(ATIS) versions provide users with the path with lowest instantaneous travel time (cost) without explicitly considering the marginal impact of this routing action and, do so even when many other similar requests follow. Further, this “passive” way of managing traffic by providing generic traffic information makes it difficult to predict outcome and may even incur adverse effect, such as overreaction (aka herding effects).

Motivated by those concerns, in this research, a behaviorally induced system optimal model is presented, aimed at further improving the system-level traffic condition towards System Optimal through incremental routing, as well as establishing the analysis methodology and evaluation framework to calibrate quantitatively the behavior change and the system benefits. In this process, the traffic model involved is carefully calibrated, first using a two-stage calibration model that is capable of matching not only the traffic counts but also the time dependent speed profiles of the calibrated links. To the best of our knowledge, this research is the first with a methodology to incorporate the use of field observed data to estimate the OD matrices departure profile.

After the network setup and model calibration, a constrained K shortest paths algorithm (CKSP) that addresses route overlap and travel time deviation issues is also proposed in this dissertation. This proposed algorithm can generate K Shortest Paths between two given nodes and provide sound travel options to the travelers to assist their travel decision making process. Lastly, a behaviorally induced system optimal model includes the development of a marginal cost calculation algorithm, a time-dependent shortest path search algorithm, and schedule delay as well as optimal path finding modules, is presented to improve the traffic flow from an initial traffic condition which could be User Equilibrium (UE) or any other non-UE or non-System Optimal (SO) condition towards System Optimal. The goal of establishing this framework is to
better capture and evaluate the effects of behaviorally induced, system optimal traffic management strategies on the overall system performance.
2 PRESENT STUDY

2.1 Dissertation Outline

With the goal of exploring and establishing the modeling framework to better capture and evaluate the effects of behaviorally induced, system optimal traffic management strategies, this dissertation is decomposed into three main research parts.

2.1.1 Research Methodology for Simulation Model Calibration

Calibrating a traffic simulation model is a necessary step in a modeling exercise, in which a baseline model needs first to be established and calibrated to ensure that the model properly depicts the macroscopic observations from the field data. Only after the calibration can the baseline model be used as the foundation for other scenarios in which alternative design or management strategies are incorporated.

Almost all the prior studies use traffic counts as the calibration target so as to determine the model formulation. Using counts as the calibration target is reasonable for a static problem in which counts are observed over a long period, and the OD matrices of interest are also static. However, calibrating the time-dependent OD matrices using count data from a congested network could become problematic, particularly in failing to estimate OD matrices so as to produce a plausible simulated congestion pattern that matches the field data. This deficiency stems from the fact that count data is not the ideal descriptor of congestion. Learning from the traffic flow fundamental diagrams, one can see that a given flow may be the result from either a light or heavy congestion situation. The traditional count-minimizing approach has no way of telling if a link with low count experiences light or heavy congestion. Mistakenly increasing OD trips simply because simulated counts are lower than the observed count will cause the model to add more trips to zone
pairs that have already been oversaturated, making the simulated link more congested with even lower counts, deviating still further from the field data.

In this dissertation, a two-stage model is built in that is capable of calibrating the offline simulation model in such a way that the congestion temporal profile pattern at the bottleneck of interest is properly depicted, which means not only the traffic counts, but also the time dependent speed profiles will match the Ground Truth Data. The upper-level problem of this two-stage problem is formulated as a minimizing link count deviation problem and the lower-level problem is the time-dependent user equilibrium traffic assignment (TDUETA) problem, which aims to match the calibrated result with the field observed traffic condition. The beauty of this bi-stage model lies not only in that the calibrated result can match both link traffic counts and travel speed, but also because it utilizes the field observed data in the demand calibration. To the best of our knowledge, this research is the first with a methodology to incorporate the use of field observed data to estimate the OD matrices departure profile.

The method is based on understanding the demand supply relations at the bottleneck and the utilization of shockwave and travel time propagation between origin and bottleneck. The two-stage model was implemented in a SBDTA model DynusT and tested in a case study in Tucson, Arizona. The testing results demonstrate the effectiveness and robustness of the proposed method under situations in which initial demand matrices deviate from the true matrices with varying degrees of deviations.

2.1.2 Research Methodology for System Optimal Approach

Real-Time Traveler Information Systems or Advanced Traveler Information Systems (ATIS) provide pre-trip and/or en route information allowing drivers to quickly assess and react to unfolding traffic conditions. The basic design concept is to present generic information to drivers,
leaving drivers to react to the information their own way. This “passive” way of managing traffic by providing generic traffic information makes it difficult to predict outcome and may even incur adverse effect such as overreaction (aka herding effects).

For those ATIS that come with path finding functionality the goal is often to provide users with the path with lowest instantaneous travel time (cost) without explicitly considering the marginal impact of this routing action and even of many others making similar requests and following that same routing information. However, it is generally known that the marginal cost of adding a driver to the traffic network includes not only his/her experienced travel time in the network, but also the delays this trip imposes on to other vehicles in the vicinity or departing afterward. The research interest in this paper is how to provide users with a route that aims to minimize the system marginal impact that will result from this routing.

In this paper, a behaviorally-induced system optimal model is presented aiming to further improve the system level traffic condition towards System Optimal through incremental routing. The proposed methodology includes the development of a marginal cost calculation algorithm, a time-dependent shortest path search algorithm, and schedule delay as well as optimal path finding modules. Both analytical derivation and numerical analysis have been conducted on a hypothetical network resembling the traffic network structure in Tucson, Arizona. The outcome of this study shows that our proposed modeling framework is promising for improving network traffic condition towards System Optimal, resulting in an economic saving of 159 million USD per year.

2.1.3 Research Methodology for Traveler Route Choice Model

Travel demand, as an aggregated terminology, emerges from individual decisions. Any effective ATDM strategy would be impossible if the traveler route choice behavior in the urban traffic network has not been fully understood. Drivers make decisions according to their individual
objectives, constraints, preferences, experiences and knowledge about travel, and their final route choice would be affected by a few factors such as traffic congestion, fuel consumption, route distance, activity location constraints including dropping off or picking up kids, preference over certain roads and so on.

Moreover, drivers' decisions are both heterogeneous and evolutionary. Drivers' previous decisions provide them with unique experience and the latest knowledge about the transportation network and traffic condition, thus influencing their subsequent decisions.

Theoretical research assumes all users are homogeneous in their route choice decision and will always pick the route with the shortest travel cost, which is not necessarily the case in reality. As noted above, researchers in Minnesota found that only 34% of drivers strictly traveled on the shortest path. The development of a constrained time-dependent K shortest paths algorithm to find K Shortest Paths between two given nodes is presented in this dissertation. The goal of this research is to provide sound route choice options to the drivers in order to assist their route choice decision process, during which the overlap and travel time deviation issues between the K paths need to be considered. The proposed algorithm balancing overlap and travel time deviation is developed in this research. A numerical analysis is conducted on the Tucson I10 corridor, The outcome of the case study shows that our proposed algorithm is able to find the different shortest paths with a reasonable degree of similarity and close travel time, which indicates the result of the proposed algorithm is satisfactory.

The detailed methodology and case study for each main research parts in this dissertation is presented in the appendices. Research efforts on the two stage calibration model is shown in Appendix A, Appendix B describes the system optimal approach, and Appendix C presents the constrained KSP algorithm.
2.2 Comprehensive Case Study

The development of the analysis methodology and platform for behaviorally induced system optimal traffic management is composed of three proposed main modules. In the appendices, the methodology of each model is presented with its case study result shown in the individual papers. In this section, the three models are integrated to form together a systematic framework. Also in this section, a comprehensive case study scenario is created that will include the above proposed Two-Stage OD calibration model, the Constrained K-Shortest Path algorithm and the System Optimal model.

This integrated case study is conducted on the Tucson I-10 network using DynusT DTA simulation software, with all three models coded in Java programming language. The goal of this case study is to demonstrate the capability of integrating its main components into the single framework and to show the performance of the whole framework.

2.2.1 Case Study Framework

Figure 2-1 illustrates the work flow of the proposed framework, which integrates the three main modules, i.e., OD Calibration model, System Optimal Model, and Constrained K Shortest Path algorithm. Each box enclosed in dashes in the picture below stands for one of the three main research modules in this dissertation.
In the proposed framework, the OD demand will first be calibrated in order to set up the simulation environment. As shown in the orange box at the top, a bi-level OD Calibration model is
proposed that is capable of calibrating time-dependent origin-destination matrices in order to match not only the traffic counts, but also the speed profile of the calibrated links. This is an essential step to set up the offline simulation environment, as only after the calibration can the baseline model be used as the foundation for other scenarios in which alternative design or management strategies are incorporated.

After the demand calibration and baseline scenarios setup, the system optimal model will be implemented to guide the driver decision making process, in order to improve the traffic from User Equilibrium to System Optimal flow. As shown in the blue box at the left side, the key to this approach is to calculate the optimal path for the driver from the overall system perspective, instead of from the individual driver's point of view. The general idea is that the system Marginal Cost of adding a driver to the traffic network will include not only his/her experienced travel time in the network, but also the delays he/she brought to the other vehicles after him/her. In other words, from the system optimization standpoint, drivers will be guided in such a way that the overall system cost can be minimized.

Theoretical research assumes all users are homogeneous in their route choice decision and will always pick the route with the shortest travel cost, which is not necessarily the case in reality. In order to affect drivers’ behavior effectively and guide the traffic flow improvement in an orderly manner, the Constrained K Shortest Path Algorithm (black box at the right side) is developed to provide drivers with multiple travel choices that are both reasonably overlapped and with low travel time deviation. The result of the CKSP algorithm will be fed into the TDSP search module in the system optimal model as the travel choice for the drivers.

In the end, with the updated travel choice decision for the drivers in the traffic network, the simulation model will run one shot using the latest vehicle and its path information to generate the
system output for further analysis. Both the system performance as a whole and individual drivers’ behavior changes will be analyzed in the following section.

### 2.2.2 Case Study Result Analysis

The proposed work flow in section 2.2.1 is coded in Java programming language and integrated with DynusT DTA simulation software. In this comprehensive case study, it is assumed 10% of drivers are willing to change their travel behavior and shift their departure time by as much as 15 minutes, meaning 90% drivers are considered as background traffic that will stay on the same route and depart at the same time, and the other 10% drivers have the flexibility both in their departure time choice and route choice so as to minimize their total travel cost. Those 10% experimental vehicles are randomly chosen out of the total population.

The other parameters of interest are taking the standard values as suggested in each individual research reported in the appendices. Except for the system integration part, this comprehensive case study scenario is identical to the case study scenario in Section 4.2 of appendix B.

The system performance and behavior changes will be analyzed in detail in the next sections here.

#### 2.2.2.1 Traffic Congestion Alleviation

To show the traffic congestion alleviation after the behavior changes, the heat maps of I-10 Northbound and Southbound before and after are drawn and shown below. X axis stands for the time and Y axis represents the physical road segment along the corridor, and the color in the heat map describes the congestion levels in the traffic network, with blue as free flow condition and red as severely congested.
In Figure 2-2 and Figure 2-3, the left side shows the heat map on the freeway for the baseline scenario which is User Equilibrium flow, and the right side shows the heat map for the comprehensive case scenario after the behavior change. It can be observed from the heat maps on the right side that the color of the congested area becomes lighter and the size of the colored area...
also shrinks. The difference is especially noticeable in the freeway southbound corridor, which indicates the traffic congestion along the freeway corridor has been alleviated noticeably.

2.2.2.2 Peak Hour Demand Reduction

Further analysis of the dataset reveals more in-depth insight into the result, the first observation is on the peak hour demand reduction. The analysis shows that, compared with the baseline scenario, the traffic demand in the peak hour is reduced by 8.6% in the case study result.

Those drivers who changed their travel behavior are divided further into several categories depending on their behavior change. Figure 2-4 shows the number of drivers from different categories.

![Figure 2-4 Travel behavior changes](image)

Recall that it is assumed that 10% of the total drivers are willing to change their departure time and/or routes. Among these 10% drivers 47% have found to change their departure time but stayed on the same route. Further breakdown shows 41% chose to leave earlier and 6% depart later than their original departure time.
Only a small number (7 divers) of drivers chose to take a different route but still leave at the same time, most likely due to the reason that peak hour traffic was so congested and since drivers had flexibility as to departure time, leaving at a different time will yield greater benefits.

The number of drivers who not only changed departure time but also traveled on a different route accounts for 3.8% of the whole population.

Drivers who changed their departure time not only removed themselves from the congestion, but also contributed to reduce the peak hour traffic demand, and in the end the whole system benefits from their behavior changes. The following section will analyze the benefit each driver category gets quantitatively.

2.2.2.3 Travel Time Saving

Due to the fact that a certain percentage of drivers changed their travel behavior by either shifting the departure time or changing the route they take, we know from the Section 2.2.2.1 that the traffic congestion is alleviated in the peak hour, which means drivers in the network are benefiting from their behavioral change by saving travel time.

The system travel time comparison between the baseline scenario case and after the behavior change is shown in Figure 2-5, the total system travel time decreases by 13.4% (drops from 131,639 to 114,056 hours). The total travel time saving after the behavior change is 17,583 hours for the simulation period.
If we divide the drivers into several categories according to their behavior changes, the travel time saving stats will be computed for the different categories respectively.

- **Group 1** - drivers who changed their departure time but stayed on the same route can reduce their own travel time by 7.5 (16.6%) min on average, 9.8 (23.6%) min for leaving earlier and 2.4 (4.5%) min for leaving later.

- **Group 2** - drivers who changed the route they took but not the departure time can reduce travel time by 1.1 (2.1%) min on average.

- **Group 3** - drivers who changed both departure time and their route choice can reduce travel time by 2.6 (6.5%) min on average.

- **Group 4** - other drivers who didn’t make any change to either departure time or route choice can benefit by 3.9 min (7.8%) of travel time reduction, in spite of taking no action.

![Figure 2-5 Total system travel time comparison](image)
2.2.2.4 Monetary Saving and Cost

If we translate the travel time saving into monetary value, the saving brought about by the small group of users’ behavior change can be significantly huge. Here, the value of time from the calculations of the Texas Transportation Institute (TTI) will be used to calculate the monetary saving.

The 2011 Annual Urban Mobility report produced by (TTI) estimated the congestion cost to be $16.79 per hour of person travel (TTI 2011). If we use that recommended value, the total annual monetary saving will be:

\[(131,639 - 114,056) \times 2 \text{ peak hours/day} \times 16.79$/hour \times 260\text{workdays/year} \]

\[= 146.3 \text{ million USD/year} \]

The cost involved in this system varies and depends on the specific strategy to be used to induce the change in travel behavior, but generally speaking it will be relatively cheap. For example, in the Spitsmijden experiment in Europe (Knockaert et al. 2007), the cost of persuading travelers to make behavior changes is between $3.99 per hour to leave earlier and $4.24 per hour to leave later, which translates to a cost of $4.2 million / year.

\[(\text{Number of drivers shifting earlier} \times 4.24$/hour + \text{Number of drivers shifting later} \times 3.99$/hour) \times 15\text{min}/(60\text{min/hour}) \times 2\text{peak hours/day} \times 260\text{workdays/year} \]

\[= 4.2 \text{ million USD/year} \]

A similar experiment was also carried out in the US and the cost involved to persuade drivers to make travel behavior changes is $13.33/hour for leaving earlier and $12.62 for leaving later (Leblanc and Walker 2013), which gives us $13.3 million in total cost:
(Number of drivers shifting earlier * 13.33 $ / hour + Number of drivers shifting later
  * 12.62 $ / hour) * 15min/(60 min / hour) * 2 (peak hours) / day
  * 260 weekdays / year

  = 13.3 million USD/year

Compared with the total monetary saving, either calculation of the cost of methods used to
foster changed behavior yields a good Return on Investment (ROI), which indicates the proposed
framework is able both to reduce the traffic congestion greatly and save a huge amount of money
with relative low cost.

2.3 Conclusions

This research explored and established the modeling framework to better capture and
evaluate the effects of behaviorally induced system optimal traffic management strategies. The
limitations of the previous studies have been addressed by

1. proposing a two-stage OD calibration model that is capable of matching both traffic counts
   and time-dependent speed profile in a large network. To the best of our knowledge, this
   research is the first with a methodology to incorporate the use of field observed data to
   estimate the OD matrices departure profile. The numerical results in a complicated network
demonstrate the potential of the proposed methodology.

2. building a model with the aim of improving the traffic condition towards system optimal
   and avoiding the herding effect, the proposed model considers the marginal cost to the
   system when assigning multiple drivers to the system, and the incremental routing
   mechanism is also applied. The outcome of the case study shows that the benefit of our
   proposed model includes significant traffic congestion alleviation, travel time saving and
   monetary saving with a relative low cost.
3. developing a Constrained Time-Dependent K-Shortest Path algorithm to generate the route choices for drivers. Addressing the route overlapping and travel time deviation issue when searching KSP for the individual driver is the major originality in the contribution of this part of this research. The outcome of the case study over the Tucson network shows the result of the proposed algorithm is quite satisfactory.

4. developing an comprehensive analysis methodology and platform for behaviorally induced system optimal traffic management that integrates the proposed three main modules. These modules are integrated together to constitute a systematic framework in this research; further, a comprehensive case study scenario is also created which includes the above proposed Two-Stage OD calibration model, Constrained K-Shortest Path algorithm and System Optimal model.

The contributions of this research lie not only in the establishment of the mathematical models whose capability and performance have been proved in the case studies, but more importantly in the integration of the different modules and the formulation of a systematic analysis methodology and platform for behaviorally induced system optimal traffic management. The integrated platform is not only capable of effectively guiding the traffic flow to improve towards System Optimal, but is also capable both of accurately evaluating the system benefit from the macroscopic perspective and quantitatively analyzing the behavior changes microscopically.
REFERENCES


Waller, S. T., Ng, M., Ferguson, E., Nezamuddin, N. and Sun, D. D. (2009). Speed Harmonization and Peak-Period Shoulder Use to Manage Urban Freeway Congestion Center for Transportation Research, The University of Texas at Austin, 3208 Red River, Suite 200, Austin, TX 78705-2650


APPENDICES
APPENDIX A: A DECOMPOSITION FRAMEWORK AND METHOD FOR CALIBRATING DYNAMIC ORIGIN-DESTINATION MATRICES FOR DYNAMIC TRAFFIC ASSIGNMENT UNDER CONGESTED TRAFFIC CONDITIONS

Paper is pending publication to Transportation Research Part B
ABSTRACT

Calibrating a traffic simulation model is a necessary step in a modeling exercise and has been an active research area, yet almost all the prior studies use traffic counts as the calibration target so as to determine the model formulation. It is, however, a widely accepted understanding that count data are not the ideal descriptor of congestion. Learning from the traffic flow fundamental diagrams, one can see that a given flow may result from either a light or heavy congestion situation.

A two-stage model is built in this research to calibrate the Time-Dependent, Origin-Destination (O-D) demand, with the upper-level problem being to minimize link count deviation and the lower-level problem being time-dependent, user equilibrium traffic assignment (TDUETA), which aims to match the calibrated result with the field observed traffic condition. The beauty of this bi-stage model lies not only in that the calibrated result can match both link traffic counts and time dependent travel speed, but also because the approach utilizes the field observed data in the demand calibration. To the best of our knowledge, this research is the first with a methodology to incorporate use of field observed data to estimate the O-D matrices departure profile.

The method is based on understanding the demand-supply relationship at the bottleneck and utilization of shockwave and travel time propagation between origin and bottleneck. The two-stage model was implemented in a simulation-based dynamic traffic assignment (SBDTA) model DynusT and tested in a case study in Tucson, Arizona. The testing results demonstrated the effectiveness and robustness of the proposed method under situations in which initial demand matrices deviated by varying degrees from the true matrices.

Keywords: O-D calibration; Simulation Model Calibration; dynamic traffic assignment; long-range planning; method of isochronal vehicle assignment; one-norm formulation
1. BACKGROUND

Traffic simulation models have been increasingly applied to analyzing a traffic network in cases where depicting the congestion pattern is often the primary concern. Traffic simulation models can be generally categorized into microscopic and mesoscopic types based on: (1) length of simulation time intervals, and (2) level of details applied to car-following and lane-changing and other relevant decision rules. Regardless of the details in the simulation logic utilized by both microscopic and mesoscopic models, most of these simulation models generate vehicles from time-dependent origin-destination (O-D) matrices. These matrices depict the amount of trips traveling between each zone pair during the period covered by the simulation. The definition of what constitutes a zone is a modeling choice. A zone could be defined to represent a parking lot or a traffic analysis zone (TAZ). In simulation, each generated trip/vehicle needs to be assigned with a path connecting the origin and the destination, and various kinds of simulation logics could be applied to each vehicle’s journey from its origin to destination. Each vehicle’s journey through the network and its interaction with many other vehicles form the traffic dynamics that are usually observed using link-based or network-based macroscopic measures, such as average flow rates, densities, speeds and/or queues.

Calibrating a traffic simulation model is a necessary step in a modeling exercise. A baseline model needs to be first established and then calibrated to ensure that the model properly depicts the macroscopic observations from the field data. Only after the calibration can the baseline model be used as the foundation for other scenarios in which alternative design or management strategies are incorporated.

Calibration of a traffic simulation model has been an active research area. The calibration efforts can be generally categorized, but not limited, into several dimensions, such as: (1) model parameters, and (2) O-D matrices. The model parameter adjustment is aimed at tuning parameters
associated with various driving behavior rules that govern information gathering, decision making and execution mechanisms during continuous driving maneuvers. Exemplary studies include (He and Ran 2000; Toledo et al. 2004; Balakrishna et al. 2005; Cools et al. 2010).

Adjustment of O-D flows to improve simulation realism also has received wide attention in the past and a rich body of research on the O-D estimation problem has been created. Most of the earlier studies focus on estimating O-D matrix with either one constant rate or split into segments each with their own constant rate and assuming a free-flow traffic condition (Carey et al. 1981; Bell 1983; Bierlaire and Toint 1995). Later research considers congestion by incorporating traffic assignment to determine the O-D proportion or splitting rate as related to the equilibrium condition (Yang et al. 1992; Florian and Chen 1995; Chang and Tao 1999; Tavana 2001; Zhou et al. 2003). A commonly seen model formulation is the least-square bi-level formulation in which the upper level problem is to minimize some weighted measures of deviation from the target matrix and from the observed counts, and lower level problem is usually formulated as linear or non-linear constraints or traffic assignment problem (Yang et al. 1992; Cascetta et al. 1993; Florian and Chen 1995; Sherali and Park 2001; Tavana 2001; Zhou et al. 2003; Chiu et al. 2007). Most of the above studies focus on the development of model formulation and solution algorithm, but none of which was tested on large real-life networks that include hundreds or thousands of zones.

Moreover, almost all the prior studies use traffic counts as the calibration target in determining the model formulation; research with calibration targets other than traffic counts are very limited. Using counts as the calibration target is reasonable for a static problem in which counts are observed over a long period and the O-D matrices of interest are also static. However, calibrating the time-dependent O-D matrices using count data from a congested network could become problematic, particularly in the estimation of O-D matrices that produce a simulated congestion pattern that matches the target data. The deficiency lies in the fact that count data are
not the ideal descriptor of congestion. Learning from the traffic flow fundamental diagram, one can see that a given flow may be the result from either a light or heavy congestion situation. The traditional count-minimizing approach has no way of telling if a link with low count experiences light or heavy congestion. Mistakenly increasing O-D trips simply because simulated counts are lower than the observed count will cause the model to add more trips to zone pairs that have already been oversaturated, making the simulated link more congested with even lower counts, further deviating from the field data (Hu and Chiu 2011).

In this research, we propose a new concept and approach that incorporates speed profile data along with count data to derive an estimate of “arrival” or “demand” information for the bottleneck of interest based on the concept of shockwaves. Such information, through proper temporal and spatial mapping, enables the model to infer departure pattern (curve) at various origins from which the bottleneck traffic is originated. Consequently, the simulated results properly capture the traffic demand/arrival at the bottleneck and, thus, match the observed congestion (speed profile).

A two-stage model is built in this research to calibrate the Time-Dependent O-D demand, with the upper-level problem being to minimize link count deviation and the lower-level problem being the time-dependent user equilibrium traffic assignment (TDUETA) which aims to match the calibrated result with the field observed traffic condition. The beauty of this bi-stage model lies not only in that the calibrated result can match both link traffic counts and travel speed, but also because this approach utilizes the field observed data in the demand calibration. To the best of our knowledge, this research is the first with a methodology to incorporate use of field observed data to estimate the O-D matrices departure profile. The numerical results demonstrate the potential of the proposed methodology.
The rest of this chapter is structured as follows: Section 2 presents the basic calibration methodology. Section 3 discusses the calibration methodology and detailed procedure. Numerical analysis results are presented and discussed in Section 4. Section 5 concludes this research.

2. LITERATURE REVIEW

2.1 General overview

Several early studies have extended the User Equilibrium (UE) problems to the so-called dynamic user equilibrium (DUE) problem (Wie et al. 1987; Ran and Shimazaki 1989; Boyce et al. May 1995). One category of dynamic traffic assignment (DTA) formulations commonly seen in the literature assumes that the time-dependent O-D trip departure pattern over the simulation (analysis) period is known a priori, thereby removing the departure time choice dimension. This assumption leads to time-dependent user equilibrium traffic assignment (TDUETA) models. The objective of this type of model is to obtain a TDUE flow pattern by equilibrating time-varying experienced trip times given time-varying O-D trips within a modeling period.

One of the first contributions to the TDUETA problem was a heuristic proposed by Yagar (Yagar 1975), which was one of the earliest approaches to recognize the importance of adequately capturing queuing phenomena in this problem. Since the late 1980s, the TDUETA modeling approaches have begun to branch into analytical and simulation-based methods. These two approaches differ in their network loading as well as in the associated assignment solution algorithms. Within the class of general analytical DTA problems, a wide range of studies can be found in the literature. Most of these approaches relied on link exit functions to represent traffic congestion. Several studies specified functional forms for the link exit function and most of these assumed certain mathematical properties (Wie et al. 1987). Ran and Boyce (Ran and Boyce 1994) formulated a continuous TDUETA model in which the link outflows were treated as a set of control variables, rather than as functions, in order to overcome difficulties posed by the non-linearity of
the link exit function for multiple origin-destination networks. A common challenge associated with analytical TDUETA models is that the First-In-First-Out (FIFO) property, used as a proxy for model tractability, is problematic and/or is not explicitly addressed in most analytical TDUETA models.

Existing time-dependent TDUETA formulations mostly fall into two categories based on how the temporal dimension is treated: discrete time mathematical programming formulations and continuous time optimal-control formulations. Friesz et al. (Friesz et al. 1989) proposed a time-dependent generalization of Beckmann’s equivalent optimization problem (for a static UE) in the form of an optimal control problem. Following Ran and Shimazaki (Ran and Shimazaki 1989), Boyce et al. (Boyce et al. May 1995) formulated a convex optimal-control model for TDUETA by defining inflows and outflows on links to be control variables. They discussed a methodology to solve the discretized version of the problem using the Frank-Wolfe algorithm and an expanded time-space network representation. No implementation or illustration of the procedure has been reported. In addition, the use of static link performance functions may preclude adequate modeling of the dynamics of congested traffic behavior.

Because of the limitations of analytical performance functions in capturing FIFO and traffic congestion, researchers have developed simulation-based approaches. Mahmassani and Jayakrishnan (Mahmassani and Jayakrishnan 1991) computed a stochastic DUE in a corridor network where traffic performance was represented with a traffic simulation model solved by the Method of Successive Averages method. In their model, the simulation-based algorithm consisted of an iterative procedure in which a mesoscopic traffic simulation model, DYNASMART, was used to represent the traffic interactions in the network, thereby evaluating the performance of the system under a given assignment (Peeta 1994; Peeta and Mahmassani 1995). The use of a traffic simulator to evaluate the objective function and model system performance circumvented the need
for link performance functions and link exit functions, ensured that FIFO was met, captured link interactions, and precluded unintended holding of traffic ensuring consistency with realistic traffic behavior. The procedure assigned vehicles to various paths directly, obviating the need to infer a path assignment from the solution to a link-based formulation. An excellent review of DTA models was given by Peeta and Ziliaskopoulos (Peeta and Ziliaskopoulos 2001).

Due to its modeling flexibility, the simulation-based DTA has become the paradigm for most commercially or academically available simulation-based transportation planning software tools, such as INRO’s DYNAMEQ (INRO 2005), VTG’s VISTA (VTG 2007), DYNASMART-P (Chiu and Mahmassani 2000; Mahmassani et al. 2007), and DynusT (Chiu et al. 2008; Chiu et al. 2008). Due to computational burdens, most of the existing transportation planning projects using DTA are limited to corridor based analysis for peak hours only (Mahut et al. 2002; Ziliaskopoulos and Chang 2004; Ziliaskopoulos et al. 2004; Balakrishna et al. 2005; Mahut et al. 2005; Zhou et al. 2008). No regional traffic assignment over a long (daily) period was found in literature.

2.2 O-D calibration

There is a rich body of research into the O-D estimation problem. Most of the earlier studies focus on estimating an O-D matrix with constant proportion or splitting rate matrix assuming free-flow traffic condition (Carey et al. 1981; Bell 1983; Bierlaire and Toint 1995). Another category of research considers congestion by incorporating traffic assignment to determine the O-D proportion or splitting rate as related to the equilibrium condition (Yang et al. 1992; Florian and Chen 1995; Chang and Tao 1999; Tavana 2001; Zhou et al. 2003). A commonly seen model formulation is the least-square bi-level formulation in which the upper level problem is to minimize some weighted measures of deviation from the target matrix and from the observed counts, and lower level problem is usually formulated as linear or non-linear constraints or traffic assignment
problem (Yang et al. 1992; Cascetta et al. 1993; Florian and Chen 1995; Sherali and Park 2001; Tavana 2001; Zhou et al. 2003; Chiu et al. 2007). Most of the above studies focus on the development of model formulation and a solution algorithm, but none of which was tested on a large real-life network with hundreds or thousands of zones.

Another less seen formulation for the least-squares problem (or quadratic) is the one-norm formulation in which the objective function is the absolute value of the deviation rather than the least-square or two-norm formulation (Sherali et al. 1997). The one-norm formulation is more computationally effective and solvable on large real-life networks than the least-square formulation because of its linear model structure, which makes it solvable by most existing LP solvers.

Frederix, Viti et al. proposed to use Marginal Computation (MaC) to derive the relationship between O-D flows and link flows; they took into account that link flows are non-separable. They used MaC to calculate the sensitivity of all link flows to every O-D flow, determined the gradient and performed a line search to determine the new estimated O-D matrix (Frederix et al. 2011). In other research, they proposed a hierarchical approach for decomposing and simplifying the dynamic O-D estimation procedure for a large-scale congested network. The network was divided into multiple hierarchical levels, and then O-D estimation was performed on each level, starting with the highest level to the lowest level. One of the advantages is that different complexity can be used for different parts of the network (Frederix et al. 2010).

Another area concerned with O-D estimation is real-time traffic estimation and prediction. An O-D matrix calibration of DYNASMART-X was performed using multi-day, peak-hour freeway traffic counts collected in Irvine, CA, based on the adjustment of the multi-day, peak-hour O-D using a least-square formulation (Zhou 2003) and Kalman Filtering (Zhou and Mahmassani 2007). Several studies have focused on the calibration of DynaMIT (Ben-Akvia et al. 1998; Park
et al. 2006; Wen et al. 2006), in which both the traffic flow models and O-D were adjusted in real time according to the incoming traffic data.

In addition to O-D calibration, Mahut et al. (Mahut et al. 2004) discussed the calibration of DTASQ (now called DYNAMEQ™) on a suburban network of Calgary, Canada. The calibration effort focused on a technique to expand the traffic analysis zone (TAZ) to enhance zonal representation of the study area, the calibration of the peak-hour O-D matrix, car-following and lane-changing model parameters, and the fine-tuning of intersection turning movements. Apart from the above studies of relatively small networks, Ziliaskopoulos et al. (Ziliaskopoulos et al. 2004) attempted to address the implementation of large-scale DTA applications. They addressed the manipulation of network and demand data, the modeling of turning movements, the efficient computation of link travel times, and the handling of complex path data. Although this study did not explicitly address the model calibration procedure, the discussions provided valid and useful insights into the application of the DTA model to long-term planning.

The studies described above highlight the fact that the DTA models, with modest effort, can be reasonably calibrated for project-level peak-hour applications. To date, however, due to computational intractability for long-period simulation and assignment, and lack of O-D calibration methods effective for large-scale networks, few attempts have been made to apply DTA in the context of a long-range planning process. The following two sections present research efforts that are aimed at overcoming these barriers.

In this research, a new concept and approach is proposed that incorporates speed profile data along with count data to derive an estimate of “arrival” or “demand” information for the bottleneck of interest based on the concept of shockwaves. Such information, through proper temporal and spatial mapping, enables the model to infer departure pattern (curve) at various origins from which the bottleneck traffic is originated. Consequently, the simulated results properly capture
the traffic demand/arrival at the bottleneck so that the simulated demand/arrival matches the observed congestion (speed profile.)

To the best of our knowledge, this research is the first with a methodology for incorporating use of field observed data to estimate the O-D matrices departure profile. The numerical results demonstrate the potential of the proposed methodology.

3. METHODOLOGY

SBDTA model calibration can be generally focused on three areas: traffic simulation, assignment and time-varying O-D matrices. Several prior studies discussed the calibrations of traffic simulation models (Hourdakis et al. 2003; Bayarri et al. 2004; Ben-Akiva et al. 2004; Dowling et al. 2004; Mahanti 2004). Although the majority of these studies were concerned with microscopic traffic simulation models, the general procedure and methodology apply also to SBDTA models. Thus, generally speaking, network attributes that represent the actual physical configurations (e.g., number of lanes, turning bays) need to match the real-world setting. Capacity inputs – such as the maximum flow rate or intersection saturation (discharge) flow rates – can be adjusted/calibrated, but need to be maintained in a reasonable range that agrees with traffic engineering standards such as the Highway Capacity Manual (FHWA 2000). Traffic flow dynamics need to represent field observation by adjusting the model parameters.

With the simulation model’s traffic dynamic parameters calibrated using actual field data and the traffic control data properly placed into the model, any discrepancy between the observed and simulated link statistics can be attributed to assignment and O-D factors. Although one can speculate with respect to various assignment formulations and various behavioral assumptions about route choice, there is no consensus or proven concept on how the assignment principle should be calibrated. In contrast, for calibration of the O-D matrices, a widely accepted practice exists for
matching the simulated and field observed measure of effectiveness (MoE) of interest, because a certain amount of errors may be introduced during the process of trip generation, distribution and mode choice. The primary MoE used in simulation-based traffic assignment (SBDTA) calibration is the percent error of the total screen-line counts, allowing error compensation of individual links in each screen line. Since DTA models generate additional MoEs such as speed, queue, and density, additional MoEs may be considered in the calibration. The following discussions focus on developing the scalable O-D calibration methodology that can be applied to real-life networks with a large number of TAZs.

A SBDTA two-stage calibration model is built in this research. Figure 3-1 shows the overall framework of this model. The method presented in this research can be regarded as the bi-level formulation classified by Lundgren and Peterson (Lundgren and Peterson 2008), with the upper-level problem being the link count deviation minimization problem and the lower-level problem being the TDUETA problem. The proposed formulation departs from the literature in that the upper-level problem seeks to minimize the absolute difference between the estimated and actual link counts via a one-normal formulation instead of the two-norm (least-square) formulation. The lower-level problem is a TDUETA problem that seeks to obtain the equilibrium assignment matrix or the route choice proportion information, which matches the calibrated link speed with the real world traffic data. The DTA model used in this study is DynusT, i.e., Dynamic Urban Simulation for Transportation, (Chiu et al. 2008) with Method of Isochronal Vehicle Assignment (MIVA) implementation capable of extended time period simulation and assignment (Nava and Chiu 2012).
3.1. Stage one methodology – Minimizing link count deviation

The objective of the first stage model is to minimize the Link Count Deviation between the simulated and field observed data. The proposed formulation departs from the literature in that the upper-level problem seeks to minimize the absolute difference between the estimated and actual link counts via a one-normal formulation instead of the two-norm (least-square) formulation. Moreover, unlike other prior formulations that minimize the weighted measures of deviation from the base matrix and from the observed counts (Yang et al. 1992; Sherali et al. 1997; Sherali and Park 2001; Chiu et al. 2007; Lundgren and Peterson 2008), the deviations between the calibrated and the base O-D matrices were constrained in the constraint set instead of being specified in the upper-level objective function. These constraints include user-specified tolerable deviations for zone pairs, as well as for total trips. This formulation strategy achieves minimal link count discrepancies while maintaining a tolerable deviation between the calibrated and the base O-D
matrices. Most importantly, this strategy facilitates a transformed linear programming (LP) upper-level formulation that can be solved effectively for a large network.

It is noteworthy that the proposed O-D calibration method includes simultaneous calibration of truck and auto O-D matrices. Many planning agencies generate separate auto and truck O-D matrices, and autos and trucks are known to have different spatial and temporal distribution patterns. Further, in calibrating the time-dependent O-D matrices for autos and trucks, the proportion of total trips allocated to each O-D demand time interval was assumed to follow those in the base O-D matrices. Doing so ensured that the problem size was manageable as the temporal pattern followed that in the base O-D matrices.

The mathematical notations and model of the O-D calibration problem are discussed as follows:

Notations:

\[ N: \] set of O-D pairs to which the screen-line counts relate; this is determined by retrieving the volume from the screen-line links after the initial equilibrium procedure and tracing their paths back to their respective O-D matrix

\[ M: \] set of screen-line links

\[ d^a_{k,n}: \] auto vehicle counts on screen-line link \( k \) from O-D pair \( n \), determined by the DTA model at the equilibrium

\[ d^t_{k,n}: \] truck vehicle counts on screen line link \( k \) from O-D pair \( n \), determined by the DTA model at the equilibrium

\[ r^a_n: \] number of daily auto trips for O-D pair \( n \) in the initial O-D matrix
The proposed formulation modeling process starts from the algebraic expression of the one-norm linear problem as stated in objective function (1):

\[
\text{minimize} \quad \sum_{n} \sum_{t} \sum_{l} \left( r_{n}^{t,c} - x_{n}^{c} \right) + \sum_{n} \sum_{t} \sum_{l} \left( r_{n}^{t,a} - x_{n}^{a} \right)
\]

subject to

\[
\sum_{n} \sum_{t} r_{n}^{t,a} = g_{m}^{a}, \quad \sum_{n} \sum_{t} r_{n}^{t,c} = g_{m}^{c}, \quad \sum_{n} \sum_{t} r_{n}^{t,a} + r_{n}^{t,c} = g_{m}
\]

where

- \( r_{n}^{t,a} \): number of auto trips for O-D pair \( n \) in O-D interval \( t \) in the initial O-D matrix
- \( r_{n}^{t,c} \): number of truck trips for O-D pair \( n \) in O-D interval \( t \) in the initial O-D matrix
- \( r_{n}^{t,a,l} \): number of auto trips for O-D pair \( n \) in O-D interval \( t \) in the initial O-D matrix estimated at iteration \( l \), \( l = 0 \) is for initial zonal O-D trips
- \( r_{n}^{t,c,l} \): number of truck trips for O-D pair \( n \) in O-D interval \( t \) in the initial O-D matrix estimated at iteration \( l \), \( l = 0 \) is for initial zonal O-D trips
- \( x_{n}^{a} \): number of estimated auto trips for O-D zone pair \( n \), decision variable
- \( x_{n}^{c} \): number of estimated truck trips for O-D zone pair \( n \), decision variable
- \( g_{m}^{a} \): Field observed auto counts on link \( m \)
- \( g_{m}^{c} \): Field observed truck counts on link \( m \)
- \( g_{m} \): Field observed total counts on link \( m \)
- \( \alpha^{a}, \alpha^{c} \): user-specified tolerable O-D zone pair trip deviation percentage for auto and truck respectively
- \( \beta^{a}, \beta^{c} \): user-specified tolerable total trip deviation percentage for auto and truck respectively
- \( \lambda \): passenger car equivalent for truck
Minimize \[ \sum_{m=1}^{\mid M \mid} \left\{ \left| \sum_{n=1}^{\mid N \mid} \left( \frac{d_{mn}^a}{r_n^a} x_{n}^a \right) - g_m^a \right| + \left| \sum_{n=1}^{\mid N \mid} \left( \frac{d_{mn}^c}{r_n^c} x_{n}^c \right) - g_m^c \right| \right\} \] \tag{1}

The first term of objective function (1) is the absolute value of the auto count deviation and the second is the count deviation for trucks. In the case that a truck O-D matrix is available but not the link count (e.g., a typical permanent count station may not produce separate auto and truck counts), the objective function (1) may be revised as:

Minimize \[ \sum_{m=1}^{\mid M \mid} \left\{ \left| \sum_{n=1}^{\mid N \mid} \left( \frac{d_{mn}^a}{r_n^a} x_{n}^a \right) \right| + \lambda \cdot \sum_{n=1}^{\mid N \mid} \left| \frac{d_{mn}^c}{r_n^c} x_{n}^c \right| \right\} - g_m \] \tag{2}

The one-norm minimization objective function (2) can be reformulated to be objective function (3) plus four constraints (4)-(7) by introducing slack variables \( h_m^a \) and \( h_m^c \) as shown below.

Minimize \[ \sum_{m=1}^{\mid M \mid} (h_m^a + h_m^c) \] \tag{3}

\[ \sum_{n=1}^{\mid N \mid} \left( \frac{d_{mn}^a}{r_n^a} x_{n}^a \right) - g_m^a \leq h_m^a \ \forall \ m = 1, ..., \mid M \mid \] \tag{4}

\[ -\left[ \sum_{n=1}^{\mid N \mid} \left( \frac{d_{mn}^a}{r_n^a} x_{n}^a \right) - g_m^a \right] \leq h_m^a \ \forall \ m = 1, ..., \mid M \mid \] \tag{5}

\[ \sum_{n=1}^{\mid N \mid} \left( \frac{d_{mn}^c}{r_n^c} x_{n}^c \right) - g_m^c \leq h_m^c \ \forall \ m = 1, ..., \mid M \mid \] \tag{6}

\[ -\left[ \sum_{n=1}^{\mid N \mid} \left( \frac{d_{mn}^c}{r_n^c} x_{n}^c \right) - g_m^c \right] \leq h_m^c \ \forall \ m = 1, ..., \mid M \mid \] \tag{7}

Next, from equations (4) and (6), two new slack variables \( v_m^a \) and \( v_m^c \) are introduced to yield equations (8) and (9).

\[ \sum_{n=1}^{\mid N \mid} \left( \frac{d_{mn}^a}{r_n^a} x_{n}^a \right) - g_m^a - h_m^a + v_m^a = 0 \ \forall \ m = 1, ..., \mid M \mid \] \tag{8}
\[
\sum_{n=1}^{N} \left( \frac{d_{mn}^c x_n^c}{r_n^c} \right) - g_m^c - h_m^c + \nu_m^c = 0 \forall m = 1, \ldots, |M| 
\]

(9)

Substitute \( h_m^a \) and \( h_m^c \) in equations (4)-(7) with \( h_m^a = \sum_{n=1}^{N} \left( \frac{d_{mn}^a x_n^a}{r_n^a} \right) - g_m^a + \nu_m^a \) and \( h_m^c = \sum_{n=1}^{N} \left( \frac{d_{mn}^c x_n^c}{r_n^c} \right) - g_m^c + \nu_m^c \) from (8) and (9), Equations (3)-(7) become (10)-(12) and (18).

The final complete model is presented in equations (10)-(20). Equation (10) is the transformed linear objective function equivalent to the original one-norm formulation. Equations (13) and (14) include constraints that ensure that the estimated number of trips for each O-D pair does not deviate from the user-specified ratio \( \alpha \). Equations (15) and (16) represent the constraints ensuring that the estimated total auto trips and truck trips do not deviate from a user-defined ratio \( \beta \). Equations (17)-(18) are non-negativity constraints. Equation (19) indicates that \( \alpha \) or \( \beta \) values are between 0.0 and 1.0. Equation (20) represents a TDUETA process that maps the O-D to the screen-line link counts.

Minimize \[ \sum_{m=1}^{M} \left[ \sum_{n=1}^{N} \left( \frac{d_{mn}^a x_n^a}{r_n^a} \right) - g_m^a + \nu_m^a \right] + \left[ \sum_{n=1}^{N} \left( \frac{d_{mn}^c x_n^c}{r_n^c} \right) - g_m^c + \nu_m^c \right] \] \[ \sum_{m=1}^{M} \left[ \sum_{n=1}^{N} \left( \frac{d_{mn}^a x_n^a}{r_n^a} \right) - g_m^a + \nu_m^a \right] \leq 0 \forall m = 1, \ldots, |M| \] \[ \sum_{n=1}^{N} \left( \frac{d_{mn}^c x_n^c}{r_n^c} \right) - g_m^c + \nu_m^c \leq 0 \forall m = 1, \ldots, |M| \] \[ (1 - \alpha^a) r_n^a \leq x_n^a \leq (1 + \alpha^a) r_n^a \forall n = 1, \ldots, |N| \] \[ (1 - \alpha^c) r_n^c \leq x_n^c \leq (1 + \alpha^c) r_n^c \forall n = 1, \ldots, |N| \]
(1 - β^a) \sum_{n=1}^{|N|} r_n^a \leq \sum_{n=1}^{|N|} x_n^a \leq (1 + β^a) \sum_{n=1}^{|N|} r_n^a \quad (15)

(1 - β^c) \sum_{n=1}^{|N|} r_n^c \leq \sum_{n=1}^{|N|} x_n^c \leq (1 + β^c) \sum_{n=1}^{|N|} r_n^c \quad (16)

x_n^a, x_n^c \geq 0 \quad \forall n = 1, \ldots, |N| \quad (17)

v_m^a, v_m^c \geq 0 \quad \forall m = 1, \ldots, |M| \quad (18)

1.0 \geq \alpha \geq 0, 1.0 \geq \beta \geq 0 \quad (19)

G = \varphi(x_n^a, x_n^c, \forall n \in N) \quad (20)

It should be noted that the problem (10)-(20) determines the optimal zonal O-D adjustment for the entire analysis period. This adjustment is allocated to each time-varying O-D matrix by distributing \( x_n^a \) following the temporal distribution of each O-D matrix, that is, \( r_n^{t,a} = x_n^a \left( r_n^{a,t,l=0} / r_n^a \right) \) where \( r_n^a = \sum_l r_n^{a,t,l=0} \). This means that the temporal patterns of the time-varying O-D matrices are maintained. Calibrating the temporal pattern to match the field observed speed profile or density will be the focus of Stage 2 calibration model.

\section{3.2. Stage one calibration procedure}

The stage-one calibration procedure consists of solving for the master one-norm problem, as well as solving a sub TDUEDTA problem. The overall algorithmic steps are briefly discussed as follows.

\textit{Step 0:} Set iteration counter \( l = 0 \)

Initialization, preparing all input data for DynusT and link count \( G \)
Step 1: Set iteration counter \( l = l + 1 \)

Utilize DynusT to perform TDUETA using the base auto and truck O-D matrices. Obtain the output (vehicles and their associated path O-D pair).

Step 2: Convergence check. Stop if the maximum number of iterations is reached or the convergence criterion is met; otherwise, proceed to Step 3.

Step 3: Prepare all matrix transformations to the standard forms shown in problems (10)-(20).

Step 4: Utilize the optimization solver to solve problems (10)-(20) to obtain the estimated \( x^a_n, x^c_n \) for both auto and trucks.

Step 5: Obtain the estimated O-D matrices using \( x^a_n \) and \( x^c_n \). Update new time-dependent zonal auto trips to be

\[
 r_n^{t,a,l+1} = x^a_n \left( r_n^{a,t,l=0} / r_n^a \right),
\]

and time-dependent zonal truck trips to be

\[
 r_n^{t,c,l+1} = x^c_n \left( r_n^{c,t,l=0} / r_n^c \right).
\]

Step 6: Go to Step 1.

3.3. Stage two methodology – Minimizing speed profile deviation

3.3.1. Basic concept

It is widely known from traffic flow theories that when the demand for a roadway segment is no greater than the roadway capacity, the demand equals the observed traffic data. However, when demand is higher than capacity, the observed flow may be equal to or lower than capacity due to congestion; and in this case, congestion is formed and represented as queues or slow-moving traffic. As shown in Figure 3-2(a), the temporal pattern of the demand profile exceeds the capacity between hour 1 and 2; however, the excessive demand cannot be observed from the output of the system as observed in Figure 3-2(b), which shows the result from the congestion. The key to a
successful demand profile calibration is to be able to reconstruct the demand (arriving) curve at the bottleneck location as shown in the demand curve in Figure 3-2. Once the demand curve at the bottleneck is constructed, this demand curve can be mapped back to the origin of the demand by shifting the demand curve by the travel time from the origin to the bottleneck location.

More specifically, the demand curve is constructed using the input-output $N$-Curve, which is discussed further in the following section.

![Figure 3-2: Demand and Observed Traffic at Bottleneck (Khisty 1990)](image)

Before discussing the overall methods, the following assumptions apply:

1. Field data exist for the place/locations of interest over the period of interest.
2. The total simulated counts and field observed counts are equal or differ within an acceptable range (e.g., 10-15%). In other words, the calibration methods discussed herein assume that the total simulated and observed link counts are in a rather similar range and the difference in congestion and speed profile pattern arise from a difference in departure profile only.
3.3.2. Construction of bottleneck demand (arriving) curve

Let \( N_d^t \) be the cumulative number of vehicles that have passed the bottleneck (denoted as location \( d \)) between time \([0, t]\) and \( N_u^t \) be the cumulative number of vehicles that have passed an immediate upstream location that is the farthest upstream extent of the slow moving vehicles caused by the shockwave (denoted as \( u \)) between time \([0, t]\). In other words, location \( u \) is in the free-flow condition not subject to the slow-moving traffic caused by the downstream bottleneck. The relationship between these two curves is shown in Figure 3-3.

![Figure 3-3: N-Curve at two freeway locations](image)

In Figure 3-3, one can see that:

- At time \( t_1 \), the difference of \( N_u^{t_1} \) and \( N_d^{t_1} \) is the number of vehicles present between \( u \) and \( d \) at time \( t_1 \).

- For a given cumulative count \( N \), the time difference between \( N_u \) and \( N_d \) curves is the experienced travel time from \( u \) to \( d \).

- In a congested situation, \( N_u^{t_1} \) and \( N_d^{t_1} \) curves are most likely to be not parallel, unless the traffic always maintain exactly the same congestion pattern between the two locations.
The relationship between flow rate, cumulative counts and segment density can be expressed as:

\[ N^t_n = \int_{t_0}^T q^t_n dt \]  \hspace{1cm} (21)

\[ \bar{k}_a^t = \frac{N^t_u - N^t_d}{L_a} \] \hspace{1cm} (22)

\[ N^t_u = N^t_d + \bar{k}_a^t L_a \] \hspace{1cm} (23)

Where,

- \( N^t_n \) is the cumulative counts at location \( n \) at time \( t \) where \( n \in \{o,u,d\} \).
- \( q^t_n \) is the flow rate at location \( n \) at time \( t \leq T \), where \( n \in \{o,u,d\} \).
- \( \bar{k}_a^t \) is the average density of segment \( a \) between locations \( u \) and \( o \).
- \( L_a \) is the length of segment \( a \).

Eq. (22) leads immediately to Eq. (23), and Eq. (23) reveals an important property, i.e., that the \( N \)-curve at the bottleneck upstream can be readily estimated using data from the downstream bottleneck. In other words, \( N^t_u \) can be obtained from the field observed data at the bottleneck location and the \( \bar{k}_a^t \) can be approximated using the density measured at bottleneck location \( d \) because it is intuitive to show that the density in the slow-moving traffic does not fluctuate much, that is:

\[ \bar{k}_a^t \approx k_d^t \] \hspace{1cm} (24)

As a result, Eq. (23) becomes:

\[ N^t_u = N^t_d + k_d^t L_a \] \hspace{1cm} (25)
Next, the wave front of the shockwave caused by the inflow and bottleneck can be expressed as Eq. (26). It is intuitive to see that the distance between the farthest extent of the shockwave and the bottleneck length $L_a$ can be calculated using Eq. (26), where the time of interest is $[t_0, T]$.

$$L_a = \max \int_{t_0}^{T} \omega_{ad}(t) dt = \max \int_{t_0}^{T} \frac{q_u^t - q_d^t}{k_u - k_d} dt$$  \hspace{1cm} (26)

Alternatively, $L_a$ can also be determined by locating an upstream detector whose speed profile does not show significant decrease and take the distance between that detector and the bottleneck.

3.3.3. Construction of departure curves at origins

The construction of the departure curves at origins relies on the estimation of travel time between the origin $o$ and the upstream location $u$ as defined in the preceding sections. Figure 3-4 offers a simplified example with one origin $o$ and one bottleneck $d$. In a more general case, multiple origins exist.

![Figure 3-4: N-curve at o, u and d](image)
The \( N \) curves for locations \( u \) and \( o \) may or may not be parallel, depending on the traffic condition between the two locations. If traffic is free flowing between locations \( o \) and \( u \), then the \( N \)-curves would appear parallel. In this case, Eq. (27) holds:

\[
N_o^t = N_u^{t + \tau_{o,u}}
\]  \( (27) \)

Where \( \tau_{o,u} \) is the free-flow travel time from location \( o \) to \( u \).

In a more general case where the experienced travel time between \( o \) and \( u \) departing at time \( t \) is time-varying as \( \tau_{o,u}^t \), then the following equation holds.

\[
N_o^t = N_u^{t + \tau_{o,u}^t}
\]  \( (28) \)

Combining Eq. (23) and either (27) or (28), the \( N \)-curve for origin \( o \) could be estimated using (29):

\[
N_o^t = N_u^{t + \tau_{o,u}^t} = N_d^{t + \tau_{o,u}^d} + k_d^{t + \tau_{o,u}^d} L_a
\]  \( (29) \)

Alternatively, Eq. (30) can also be used.

\[
N_o^t = N_u^{t + \tau_{o,u}^t} = N_d^{t + \tau_{o,u}^d} + k_d^{t + \tau_{o,u}^d} \left\{ \max \int_{t_0}^{T} \frac{q_u^t - q_d^t}{k_u^t - k_d^t} dt \right\}
\]  \( (30) \)

In the case where multiple origins exist, \( N_u^t \) will need to be disaggregated into several sub \( N \)-curves \( N_{u-o}^t \) each representing traffic coming from each origin \( o \). To construct \( N_{u-o}^t \) we would take the flow rate at location \( u \), \( q_u^t \) and distribute \( q_u^t \) to \( N_{u-o}^t \) for each origin \( o \) according to the origin distribution fraction \( \delta_{u,o}^t \) with \( \sum_o \delta_{u,o}^t = 1.0 \); that is \( q_{u-o}^t = q_u^t \cdot \delta_{u,o}^t \), where \( q_{u-o}^t = dN_{u-o}^t / dt \). The term \( \delta_{u,o}^t \) cannot be directly measured in the field, but can be estimated by scanning the simulated vehicle trajectories from the simulation model. This approach is similar to the assignment matrix in traditional O-D estimation methods. With \( q_{u-o}^t \) being estimated, the sub \( N \)-
Curve $N_{u \rightarrow o}^{t}$ can be obtained using an integral, namely, $N_{u \rightarrow o}^{t} = \int_{0}^{t} q_{u \rightarrow o}^{t} \, dt$. The departure curve for origin $o$ can therefore be expressed as:

$$N_{o}^{t} = N_{u \rightarrow o}^{t + t_{u}}$$  \hspace{1cm} (31)

3.4. Stage two calibration procedure

Using the methods described in Section 3.3, the departure curves at the various origins that generate trips traversing the bottleneck can be estimated following the calibration procedure as illustrated in Figure 3-5. In this work flow, the calibration procedure is structured into several major steps:
3.4.1. Step A: Check if speed profile is satisfactory

Compare the calibrated and observed speed profiles at the calibrated bottleneck. If several calibrated links exist, sum them up to calculate the summation deviation. At the beginning of stage 2 calibration, the calibrated flow, speed and density are the calibrated results from Stage 1 calibration model; after that, the outcomes of the last iteration in the Stage 2 model will be fed back to the next iteration to determine the convergence by comparing the results with field observed data.

If the deviation measure is within the pre-set threshold, stop; otherwise, go to step B.

3.4.2. Step B: Location \( d \) flow curve estimation

This step calculates the estimated flow curve \( q'_{d,OD} \) at location \( d \) at time \( t \). On the one hand, from the field data, we have the real volume \( f_{real} \) on the bottleneck link. On the other hand, from the simulation based model, we have not only the simulated volume \( f_{simulation} \), but by tracking the trajectory of each vehicle, we can also split \( f_{simulation} \) into different O-D pairs according to the vehicle origin zone O and destination zone D, i.e. we will have \( q^i_{d,OD} \).

Then, the Estimated Flow Curve for origin zone O and destination D at location \( d \) at time \( t \) can be calculated by Equation (32):

\[
q'_{d,OD} = \frac{q^i_{d,OD}}{f_{simulation}} * f_{real}
\]  

(32)

3.4.3. Step C: Location \( u \) arrival curve construction
This step constructs the demand (arrival) curve at the bottleneck upstream location \( u \) using the downstream bottleneck observed data. In other words, with available field flow rate \( q_d^t \) and speed data \( v_d^t \), density \( k_d^t \) can be obtained using \( k_d^t = q_d^t / v_d^t \).

From Eq. (25), we can know the difference between \( N_u^t \) and \( N_d^t \) is the number of vehicles in the queue. We need to split these vehicles among different O-D pairs according to their flow proportions in the simulation model. In order to calculate Estimated Arrival N-Curve for origin zone O and destination D at location \( u \) at time \( t \), Eq. (25) can be changed to be:

\[
N_{u,OD}^t = N_{d,OD}^t + k_d^t L_a * \frac{q_{d,OD}^t}{f_{simulation}} = \int_{t_0}^{T} q_{d,OD}^t dt + k_d^t L_a * \frac{q_{d,OD}^t}{f_{simulation}}
\]

(33)

3.4.4. Step D: Origin o departure curve

Changed from Eq. (28) and Eq. (31), for each O-D pair, the demand (departure) curve at the Origin location \( o \) can be calculated by Eq. (34).

\[
N_{o,OD}^t = N_{u,OD}^t + k_d^t L_a * \frac{q_{d,OD}^t}{f_{simulation}}
\]

(34)

3.4.5. Step E: Adjust origin-departure profile in the simulation model

At this step, \( N_{o,OD}^t \) has been estimated from field data and we can compare it with the simulation departure \( N \) curve. If deviation is found, changes need to be made to the simulation \( N \)-curve according to the difference proportionalities. Figure 3-6 shows that the departure \( N \)-curve in simulation at an origin (denoted as \( N_{uncalibrated} \)) is found to differ from the one estimated as \( N_{o,OD}^t \) (denoted as \( N_{calibrated} \)), in which \( N_{uncalibrated} \) first overestimates \( N_{calibrated} \) and then later underestimates. As shown in Figure 3-6, the adjustment process would focus on reducing \( N_{uncalibrated} \) before the crossover time and increasing \( N_{uncalibrated} \) after the crossover time by
changing the vehicles departure time in the simulation model. The end goal is to adjust the simulated departure curve to be similar to the $N$-Curve estimated through Eqs. (21)-(34).

![Figure 3-6: Change simulation input](image)

By adjusting the vehicle departure time, the uncalibrated departure curve will move towards the calibrated departure curve, i.e., moving along in the improving direction. In order to find an improving feasible solution, the step size alpha is also introduced ($0<\alpha<1$). The new departure $N$ curve can be calculated by Eq. (35).

$$N_{\text{new}}^t = N_{\text{uncalibrated}}^t + (N_{\text{calibrated}}^t - N_{\text{uncalibrated}}^t) \times \alpha$$  \hspace{1cm} (35)

The larger alpha is, the closer the new departure $N$ curve will be to the calibrated curve. When step size is equal to 0, no vehicle departure time will be changed, and the new departure $N$ curve is exactly the same as the uncalibrated one. When step size is equal to 1, the new departure $N$ curve will be the same as the calibrated $N$ curve.
Note: In cases in which multi-bottleneck links exist, one principle should be observed, which is to avoid changing the departure time of vehicles whose departure time has already been changed. For example, if departure times of vehicle set S have been changed when calibrating the first bottleneck link, then when the other bottleneck links are being calibrated, these departure times of vehicles in set S should not be changed again; only the other vehicles traveling through the calibrated links can be selected and changed.

3.4.6. Step F: Rerun simulation with DUE

Once the simulation departure profile is modified, we need to re-run the simulation to reach DUE within the network using the updated vehicle departure time and the vehicle path. After the simulation process, the link statistics will be updated, including the link speed, density and volume. And then go back to step A.

It is important to note that, in each subsequent instance when carrying out 3.4.3, the travel time between each origin \( o \) and \( u \tau_{o,u}^t \) needs to be updated using travel time extracted from the latest simulation run in Step F. This updating is needed as, in a general case, travel times are affected by the departure profile. The travel time \( \tau_{o,u}^t \) used in the current iteration may not be consistent with what would be produced by the departure profile produced by Step E. Iteratively updating \( \tau_{o,u}^t \) would allow the \( \tau_{o,u}^t \) inputted into and outputted from Step F to be equal, meeting the consistency requirement.

4. CASE STUDY

A case study was conducted to illustrate the performance of the proposed model. We chose the Tucson network to be our case study region. The network is constructed in DynusT (Dynamic Urban Simulation for Transportation); the resulting network consists of 395 nodes, 830 links and 80 traffic zones. Figure 4-1 shows the Tucson network in DynusT GUI.
4.1 Scenario setup

The highlighted two links are chosen as the calibrated bottleneck links in this network. Bottleneck link 1 is located on the I-10 freeway going Southbound, from node 3063 to node 3061, there is an on-ramp before node 3063 and an off-ramp starting from node 3061. Bottleneck link 2 is an arterial link going Eastbound from node 2036 to node 7075. This is a generalized network with 80 zones, and thousands of different O-D pairs.

The total simulation time is 4 hours in this case study, with 15 minutes in each time interval. In total, there are 16 time periods.

The Initial Demand Derivation (noted as the un-calibrated case) is constructed in such a way that:
1) the total travel demand is 90% of that of the Ground Truth Data (GT), and
2) the temporal distribution of travel demand is different from that of GT: the peak demand in GT is constructed between 2nd hour and 4th hour, but in the initial scenario setup, the traffic congestion happens between 1st hour and 3rd hour.

Thus, the traffic condition of the bottleneck links that are to be calibrated is significantly different from that of the base line case scenario.

At the 16 locations where the Ground Truth volume data is available, Figure 4-2 shows the total traffic demand of the un-calibrated case compared with the GT. We can see the overall demand in the un-calibrated network is lower than the baseline scenario.

![Figure 4-2 Overall demand of the network](image)

Before the calibration, the temporal profiles of the link traffic condition are also dramatically different from the Ground Truth. For example, Figure 4-3 and Figure 4-5 describe the temporal profiles with respect to speed on the freeway link and arterial link respectively, whereas Figure 4-4 and Figure 4-6 show the temporal profiles with respect to volume; one can clearly see the difference in traffic condition between the GT and Un-calibrated scenario.
Figure 4-3 Speed profile on freeway link

Figure 4-4 Volume profile on freeway link

Figure 4-5 Speed profile on arterial Link
4.2 Stage 1 results analysis

Stage 1 algorithm is coded in MATLAB and Python. Link count data are available from 16 sensors in our test area.

The Stage 1 calibration algorithm converges at the 9th iteration. After stage 1 calibration, the volume comparison result is shown in Figure 4-7. The red diamonds stand for the two bottleneck links to be further calibrated in Stage 2, while the other diamond nodes, shown in blue, stand for the other 14 sensors in the network. The link traffic counts of the Un-calibrated scenario are also shown in the figure.

- The X-axis indicates the traffic counts from the sensor, i.e. the GT data. The Y-axis is the Link traffic counts before and after stage 1 calibration. The stage 1 calibration goal is to move the nodes as close to the 45 degree line as possible, which means the calibrated result matches the GT data.
Before stage 1 calibration, the traffic counts on the calibrated links are much lower than the GT traffic counts (which was by design in the initial solution), but after stage 1 calibration, the total traffic counts on the links match much better with the GT data, which can be demonstrated by the nodes staying closer to the 45 degree line.

After stage 1 calibration, the travel demand is almost the same in the Calibrated case as in that of the GT, which satisfies the 2nd assumption of the stage 2 calibration. (Total simulated counts and field observed counts are equal or differ within an acceptable range)

Although the total travel demand has been matched quite well after stage 1 calibration, if we look at the speed and volume temporal profile again, the traffic condition is still way off, as demonstrated in Figure 4-8, Figure 4-9, Figure 4-10 and Figure 4-11.

![Figure 4-8 Speed profile on freeway Link after stage 1 calibration](image)

![Figure 4-9 Volume profile on freeway Link after stage 1 calibration](image)
4.3 Stage 2 results analysis

Stage 2 is coded in JAVA programming language. The calibration model converges at the 96th iteration. Comparing the bottleneck speed profile between the GT and the stage 2 calibration result, noticeable improvement can be observed from Figure 4-12 and Figure 4-13.
Figure 4-12: Freeway Link Speed after stage 2 calibration

![Freeway Link Speed graph](image)

Figure 4-13: Arterial Link Speed after stage 2 calibration

We can see the speed profile difference between the GT case and Un-calibrated case. For the Freeway Link, the speed drop in the Un-calibrated case happened around the 20th minute, and even after we managed to match the total traffic counts on this link after stage 1 calibration, the speed profile is still very different - the speed drop is about 1 hour before the drop occurs in the GT case. For the Arterial Link, heavy congestion happened from the 1st hour, although in the real case the traffic is light. However, from the above figure, one can easily tell the huge speed profile improvement after stage 2 calibration. The speed temporal distribution after the two-stage calibration is actually very close to the GT case.

We can also compare the bottleneck volume time-varying profile curves by visual inspection. The results in Figure 4-14 and Figure 4-15 consistently show that the Calibrated curve is able to replicate the pattern of the GT case very well. These Figures also show the discrepancies of the result from stage 1, i.e. even if we can manage to match the link traffic counts as closely as possible, the difference in temporal statistics for the link can still be huge. That is why the proposed stage 2 calibration is necessary and important.
The departure profile can be observed in this controlled experiment and be used to examine the performance of the proposed procedure. From Figure 4-16 and Figure 4-17, one can clearly see the difference in departure profile between the GT and Un-calibrated case. The Calibrated curves become much closer to the GT curve. This result reveals an important property, i.e., although this problem potentially has multiple solutions, our proposed method is able to converge to the true solution as shown in the GT and Un-calibrated case comparison.
The vehicle cumulative departure \( N \)-curve after Stage 1 is also examined in this case. It can be shown for this case that the initial departure profile \( N \)-curve, i.e., after Stage 1, differs significantly from that of the GT case. After calibration, with Stage 2 completed, the resulting departure profile \( N \)-curve becomes very similar to the true profile, reiterating the ability of our proposed method to find the true solution in spite of the large initial deviation. This indicates rather sound robustness of our proposed method under a wide range of initial error situations.
The MAE (Mean Absolute Error) of speed from different simulation iterations is also analyzed to show the convergence property of the algorithm. From Figure 4-20 one can clearly see the MAE value drops rapidly for both calibrated links in the first several iterations and becomes stable after that, demonstrating that the calibration model is able to converge very quickly.

### 4.4 Testing an alternative scenario

This scenario is created to test if we start with another initial demand, whether the result will converge to the same result. The total simulation time is also 4 hours in this case study, with 15 minutes in each time interval.
The Initial Demand Derivation (noted as the Un-calibrated case) is constructed in such a way that:

1) the total travel demand is 95% of that of the Ground Truth Data (GT), and

2) the temporal distribution of travel demand is different from that of GT: the peak demand in GT is constructed between 2nd hour and 4th hour, but in the initial scenario setup, the traffic congestion happens between 1st hour and 3rd hour.

Thus, the traffic condition of the bottleneck links to be calibrated is significantly different from that of the base line case scenario.

4.4.1 Calibration result

After two stages of calibration, we find the proposed calibration algorithm is also able to calibrate the traffic condition on the links of interest to match the GT scenario quite well. The temporal speed and volume profile are as follows:

Figure 4-21: Freeway Link Speed after stage 2 calibration
4.4.2 Convergence property

In this section, the calibrated result in 4.3 (scenario 1) will be compared with the calibrated result in 4.4.1 (scenario 2) and the comparison will be analyzed to determine whether, if starting from a different initial solution, the calibrated demand will converge to the same solution. In this
section, the temporal demand generation profile from each zone was analyzed, and the MOE was calculated as follows:

\[ \text{MOE} = \frac{|(\text{demand}_{\text{scenario1}} - \text{demand}_{\text{scenario2}})|}{\text{demand}_{\text{scenario1}}} \]

Figure 4-25 shows the cumulative MOE curve, from which we can find that about 80% of zones are within 20% of the Ground Truth Data demand range, and we can safely conclude the calibrated demand converges to the same solution. X-axis denotes the error level, and Y denotes the MOE values.

![Cumulative MOE curve to test convergence](image)

Temporal demand for O-D pair was also analyzed; Figure 4-26 shows the time dependent traffic flow, temporal profile from zone 80 to zone 78, and Figure 4-27 shows the traffic demand from zone 78 to zone 57. Both examples show that, although starting from a different initial solution, the calibrated traffic demand is able to converge to a close result.
5. DISCUSSIONS AND CONCLUDING REMARKS

This paper documents the research effort in developing a unique multi-modal calibration procedure to calibrate the time-dependent origin-destination (O-D) matrices’ departure profile using data observed from the bottleneck location. The upper level model is based on a LP O-D calibration formulation transformed from the one-norm minimal link count deviation formulation, and the lower level problem is formulated as a TDUETA problem that seeks to obtain the equilibrium assignment matrix which matches the calibrated link speed with the real world traffic data. The primary research interest was to overcome the deficiency of the traditional count-minimizing approach that is ill-behaved due to the lack of incorporating congestion indicator data.
into the calibration process. By incorporating the observed flow, speed and density data at the bottleneck and utilizing shockwave theories and demand/supply concepts, the proposed method has been shown to produce satisfactory calibration results regardless of the initial error margins in the examined numerical cases.

The applied case study in Tucson, Arizona, demonstrates the feasibility of incorporating DTA into the long-range planning process using the proposed approaches. The calibration results indicate that the convergence rates are satisfactory, along with the total demand being only slightly modified. First, the screen-line link volume errors met or surpassed existing commonly accepted standards. Using the existing standard, the obtained errors are close to 0% in almost all screen lines in high volume groups, even using a more stringent measure, errors obtained in most screen lines and volume groups are 10-15%. Second, the speed profile of the bottleneck links of interest is able to be matched quite well after the calibration, which is rarely seen in the existing literature, as the majority of the research focuses on matching the link counts only. Also, starting from a different initial solution, the detailed convergence test conducted at the zone level proves that, regardless of the difference in the initial solution, the calibrated result is able to converge to an acceptable level.

From this study, it is learned that O-D calibration should be performed only after conducting a careful examination and debugging of network link and node attributes, simulation model parameters, signal timing, capacity settings and vehicle loading, etc. Incorrect simulation setting may create artificial and misleading congestion, without which carefully correcting the O-D calibration would not be meaningful. Another important caution for applying the proposed model is that minimizing on link counts may create misleading results under a congested and limited time period (e.g., peak hours) situation, as low traffic counts may result from either light traffic or severely congested situations. The proposed model will not be able to differentiate either case, unless other traffic data such as speed or density is also considered in the formulation. Nonetheless,
this is an issue of lesser concern for the proposed model when applying it to 24-hour assignment, as traffic volume conservation is maintained over a 24-hour period.

A further related question for applying the base year calibration in future planning years is how one should extrapolate the new calibrated demand to future years. This is a new issue compared with existing practice, in which the calibrated link attributes (e.g., link capacity in the BPR function) allow transportation planners to apply uniformly the same “calibrated” factors to future year networks, while using the forecasted future demand. Applying the calibrated O-D demand to future years requires additional considerations. One possible strategy is to apply the O-D percentage change from the original demand to the calibrated demand and apply it to the future year demand. This method should be exercised with extreme care, as population and economic growth, location of future attractors, and many other factors need to be examined before applying them. It is suggested that a more robust method may be to take the calibrated demand back to the demand generation and/or trip distribution stages to calibrate the factors within those models and apply the corrected parameters to future years to generate the future demand. However, this remains a future research subject.

6. REFERENCES


Validation. 83rd Annual Meeting of Transportation Research Board, Washington D.C., Transportation Research Board.

Transportation Science 17(2): 198-217.


PII TASK 4, prepared for Federal Highway Administration. Austin, Texas, University of Texas at Austin.


APPENDIX B: DEVELOPMENT OF A BEHAVIORALLY INDUCED SYSTEM OPTIMAL MODEL FOR ACTIVE TRAFFIC AND DEMAND MANAGEMENT SYSTEM
ABSTRACT

Most existing real-time traveler information systems use travel time when searching for the shortest path without explicitly considering the marginal cost to the system when assigning multiple travelers to the system. This paper documents the research effort in developing a Behaviorally Induced System Optimal model to improve the system performance towards System Optimal. The proposed approach includes the development of a marginal cost calculation algorithm, a time-dependent shortest path search algorithm, and schedule delay as well as optimal path finding modules. Both analytical derivation and numerical analysis on a hypothetical network resembling the traffic network structure in Tucson, Arizona, have been conducted, and the benefit analysis focuses on both the individual and system-level perspectives. The effects of different characteristics of changes in behavior are analyzed in the case study. The case study results show the benefit of the proposed methodology in producing significant traffic congestion alleviation, reduced travel time and substantial monetary saving.

Keywords: System Optimal Model, Travel Behavior, Behaviorally Induced System Optimal, Active Traffic and Demand Management (ATDM), Dynamic Traffic Assignment, Traffic Simulation, Traffic Modeling
1 INTRODUCTION

Traffic congestion has been imposing a tremendous burden on society as a whole. In the US, congestion costs were about $115 billion in 439 urban areas in the year 2010, compared to $113 billion (in constant dollars) in 2006, according to a Texas Transportation Institute (TTI) report (Schrank et al. 2011). The European UNITE project (de Palma and Lindsey 2011) estimated the costs of traffic congestion in the UK, France and Germany to be respectively 1.5%, 1.3% and 0.9% of GDP. In Australia, the Bureau of Infrastructure, Transport and Regional Economics has estimated that urban congestion alone will cost nearly AU$20 billion by 2020 (Low and Odgers 2012)(Low and Odgersb 2012). A researcher in Korea found the socio-economic cost caused by traffic congestion is approximately $45.9 billion per year plus 336,000 job losses (Jun 2011).

Active Traffic and Demand Management (ATDM) – defined by the Federal Highway Administration (FHWA) (Luten et al. 2004) as better balancing of the need to travel a particular route at a particular time with the capacity of available facilities to handle this demand efficiently – encompasses market-ready technologies and innovative operational approaches for managing traffic congestion within the existing infrastructure. The purview of ATDM is to combine travel demand management (TDM) and active traffic management (ATM), to actively influence the need to travel, as well as the associated travel pattern, for the purpose of promoting efficient use of the roadway system handling the vehicle demand (Zheng et al. 2011). The vision for Active Traffic and Demand Management research is to allow transportation agencies to increase traffic flow, improve travel time reliability, and optimize available capacity throughout the transportation network (Cronin and Sheehan 2012).

Real-Time Traveler Information Systems or Advanced Traveler Information Systems (ATIS) provide pre-trip and/or en route information allowing travelers to quickly assess and react to unfolding traffic conditions. The basic design concept is to present generic information to travelers, leaving drivers to react to the information their own way. This “passive” way of managing traffic by providing generic traffic
information makes it difficult to predict outcome and may even incur adverse effect, such as overreaction (aka herding effects).

For those ATIS that come with path finding functionality, the goal is often to provide users with the path with lowest instantaneous travel time (cost) without explicitly considering the marginal impact of this routing action and even of many others following similar requests. The research interest in this paper is how to provide users with a route that aims to minimize the marginal system impact that results from this routing. In other words, when one makes a trip, the true Marginal Cost (MC) would include not only the one traveler’s experienced travel time in the network, but also the delays this trip imposes upon other vehicles in the vicinity or departing afterward.

In this paper, a behaviorally induced, system optimal model is presented aiming to further improve the system level traffic condition towards System Optimal through incremental routing. Both analytical derivation and numerical analysis have been conducted on a hypothetical network resembling the traffic network structure in Tucson, Arizona. The outcome of this study shows that our proposed modeling framework is promising for improving network traffic condition towards System Optimal, resulting in economic saving of 159 million USD per year.

This paper is organized as follows. Section 2 reviews the relevant past literature. Section 3 discusses the modeling framework and approach in detail, and Section 4 presents the case study results using the DTA modeling platform DynusT. Section 5 concludes this research.

2 LITERATURE REVIEW

Wardrop stated his first principle of choice "The journey times in all routes actually used are equal and less than those which would be experienced by a single vehicle on any unused route," which is generally referred to as "User Equilibrium" (UE), and his second principle "At equilibrium the average journey time is minimum," which is usually referred to as "system optimal" (SO) (Wardrop 1952).

The UE condition is achieved when all the drivers have perfect information about traffic conditions and make the optimal decision independently based on their own interests. Many calculation techniques for
this assignment have been developed and researched already; some of the most famous examples include the Frank-Wolfe algorithm introduced by Dafermos (1969), Method of Successive Averages (Sheffi 1985), gradient projection and projected gradient method, but research on UE or based on UE is abundant. Another thing worth noting is that almost all commercial or non-commercial traffic assignment simulation software programs are built based on UE, such as TransCAD by Caliper, Visum by PTV, DynusT by the University of Arizona, and so on.

In contrast, SO would require a system to have control over all vehicles' routing decisions, and to make decisions to minimize the system total travel time as a whole, which is more difficult to implement. Some of the most famous research on SO includes the M-N model by Merchant and Nemhauser (Merchant and Nemhauser 1978; Merchant and Nemhauser 1978), Ziliaskopoulos's linear programming model for the Single Destination SO-DTA Problem (Ziliaskopoulos 2000) and Daganzo's research based on the Cell Transmission Model (CTM) (Daganzo 1994; Daganzo 1995), as well as plenty of others, such as the deterministic queuing assignment model proposed by Ghali and Smith (Ghali and Smith 1995), Peeta and Mahmassani's formulation of two dynamic network traffic assignment models incorporating a traffic simulation model within an overall iterative search framework (Peeta and Mahmassani 1995), and so on.

A rather wide and diversified body of research about Marginal Cost exists in the literature, studying various aspects including theory, calculation, application, etc. In fact, research on Marginal Cost exists in several research areas, such as Congestion Pricing and System Optimal Dynamic Traffic Assignment (SO-DTA), Transportation Policy and Decision Making, e.g., consideration of a Vehicle Miles Traveled (VMT) Tax. Depending upon the research objective, these Marginal Cost-related research efforts could be generally classified in two categories: 1) External Marginal Cost and 2) Marginal cost in terms of travel time.

External Marginal Cost usually refers to the various externalities the driver brings to the existing network, such as the congestion, infrastructure deterioration, emissions, safety impacts and so on (Zhang and Lu 2012). Mayeres et al. (Mayeres et al. 1996) described, in their research, a methodology to compute the multimodal external marginal cost, where congestion, accidents, air pollution and noise were included
in the marginal external cost. Verhoef built a framework to analyze the environmental external effects (Verhoef 1994); it was concluded in that paper that most of the studies carried out in that field will, by definition, provide an underestimation of road transport's external costs. Ferrari and Zhang (Ferrari and Zhang 2012) proposed a model to calculate marginal costs of five components, i.e. Safety, Travel Time, Vehicle Operations, Agency, and Emissions. Then, the true marginal cost to society of each vehicle is obtained for each roadway segment by combining these component costs. The marginal cost research in this category usually tries to translate those external costs to be stated and understood in terms of monetary value in order to assist in real world decision making.

Marginal Cost research in the second category usually looks into the marginal travel time contribution of an additional unit of flow into the system. Ghali and Smith (1992) defined four levels of Marginal Cost from the perspective of queuing delays. The first level is the travel time on that link that an additional driver experiences; the second level includes the additional delay experienced by all vehicles that traverse that link after that vehicle. The third level extends the definition of the second level to include the additional delay to all vehicles (on all links) whose path travel times are increased due to that vehicle, and the fourth level defines the global link, marginal travel time brought about by that additional vehicle. Peeta (1994) also has a similar definition for marginal travel time in this category from several perspectives: 1) static or time-dependent, 2) global or local, and 3) link-level or path-level.

Various studies can be found on Marginal Cost with respect to the second category, especially in the area of congestion pricing and System Optimal Dynamic Traffic Assignment. To name a few: Sheffi (Sheffi 1985) calculated the marginal travel time cost caused by the additional traveler on Ghali and Smith's first and second level, but the link interactions were ignored in his static System Optimal Assignment problem. Shen et al. (2007) demonstrated by analyzing a series of examples that path marginal costs are not simply the summation of the corresponding link marginal costs, unless the flow perturbation travels with the vehicle unit that initiated the perturbation. They proposed an evaluation method that decomposed path marginal costs to the differing marginal costs for each link of such networks. Qian and Zhang (2011)
proposed a method to compute the total path marginal cost in a network with a cell-transmission-model-based (CTM-based) kinematic wave model, which they used to model traffic dynamics by tracing the changes in the cumulative flow curves of the bottleneck links on which queues form. Alibabai and Mahmassani (2012) built a model to evaluate marginal cost that can guide the search for system optimal traffic strategies. In their model, the anatomy of flow movement is analyzed on the micro level, a perturbation analysis method is designed to quantify the sensitivity of the path travel times to the path flows, and application of the perturbation analysis in solving the SO-DTA problem is described, in conjunction with a new solution method, called SO ordered method. Carey and Watling (2012) developed an SO model that more closely reflects traffic flow theory and derived the marginal costs and externalities from that. They also extended the existing SO formulation using the CTM to allow more general nonlinear flow density functions and derive and interpret system marginal costs and externalities.

One important, noteworthy aspect of real-world application of the SO concept is the partial compliance with the SO principle. Not all users have access to SO information and also not all are able to, or willing to, comply with the SO principle. While assuming all users are SO compliant is unrealistic, one can still exploit the opportunity to have only a fraction of total drivers be SO compliant, if they are motivated to do so. This research aims to exploit such opportunity and investigate the degree of overall system improvement that can be achieved due to such behaviorally induced drivers following routes that are computed based on the SO principle. Most preceding research also assumed fixed demand, limiting the choice dimension only to route but not departure time. This research also aims to relax this limitation and explore both route and departure as dimensions of the SO management decision.

3 METHODOLOGY

In exploiting the overall concept, this research chooses to apply simple marginal calculation combined with the time-dependent shortest path and path finding method as an integrated procedure. The marginal cost of concern includes both route and departure time dimensions. From the departure time standpoint, this research considers that different trip purposes may have different departure or arrival time
flexibility constraints. For example, work trips may usually involve a much higher late-arrival penalty than early arrival. A similar penalty structure may also apply to pick-up or drop-off kid trips and meeting appointments that have a specific targeted arrival time. In such cases, the preferred decision may lean toward either leaving earlier or taking a less congested route. Shopping or social trips may permit higher departure and arrival flexibility and lower early and late arrival penalty.

Generally speaking, several scenarios would exist for a specific trip: 1) if departure time is fixed and only the path is flexible, then the minimal marginal cost path for the said departure time needs to be sought. 2) If route is fixed and departure time is flexible, then the optimal departure time for the said path is sought. 3) If both departure time and route are flexible, then the minimal marginal cost path in both spatial and temporal dimensions would be of interest. For different types of aforementioned trip characteristics, the effectiveness of strategies on the entire traffic network would need to be evaluated. In light of understanding the effect of a varying percentage of departure flexibility, various scenarios with different percentages of users in various user classes and/or different flexibility on travels are also analyzed.

Further, schedule delay (SD) measures the difference between the actual arrival time and travelers’ preferred arrival time (PAT), and is also an essential part of the trip planning decision process. Generally speaking, for the departure time and route choice problem, travel time and arrival schedule delay are considered to be the two travel impedances that are considered by travelers. Sometimes travelers may have a hard constraint or high penalty on arrival time (e.g., catching a flight). A lower penalty for schedule delay may be observed in shopping, social, or entertainment trips, and is explicitly considered in this research.

### 3.1 Research framework

Following the aforementioned concepts, this research proposes a behaviorally induced, system optimal model following the framework illustrated in Figure 3-1. The major body of this work flow is the iteration included in the blue box, which has four major components followed by a link flow update:

1) Marginal Cost calculation / update
2) Time-Dependent Shortest Path search
3) Optimal path finding
4) Link flow update

Figure 3-1: Research Framework

The detailed modeling procedure as described below starts from a pre-defined departure and route choice condition, which could be UE or any other non-UE or non-SO condition:
Step 1: Randomly select a certain percentage of vehicles from the dataset to be the experiment vehicles, for example assume 10% of all vehicles in the city are willing to change their travel behavior

Step 2: Read traffic demand for each link. Here the demand is calculated by scanning through the vehicle trajectory file and sum up all the vehicles on the same link at the same time. The reason for not using the link volume to approximate demand is that the existence of link capacity constraint in the peak hours might cause the underestimation of link demand.

Step 3: Set $i = 1$ and compute the link marginal cost. Use the traffic demand calculated in step 2 together with other available data to derive the marginal cost for each link; the equations used will be demonstrated in Section 3.2

Step 4: Set $t = t_0$, and Find Time-Dependent Minimal Marginal Cost Path using the link marginal cost computed in Step 3, Section 3.3 further describes the TDSP search algorithm in detail.

Step 5: Compute the Schedule Delay for that path, the SD calculation method will be presented in Section 3.4

Step 6: If there is another departure time choice for the user, update the departure time $t$ and go back to Step 4; otherwise go to step 7

Step 7: Among all different departure time and path choices, choose the one with the least generalized cost, which is computed through a linear combination of travel time and schedule delay. The detailed logic of finding the optimal path will be explained in Section 3.5 below

Step 8: Assign that optimal path to user $i$, and update the link volumes of interest, i.e., decrease the link’s volume along the previous route by 1, and increase the link’s volume along the new route by 1.

Step 9: Determine if all experiment vehicles have been assigned a new path; if yes, go to Step 10. Otherwise, set $i = i + 1$ and go back to Step 3 to iterate, until all experiment vehicles have found a new optimal route.
**Step 10:** Re-write the simulation input file using the vehicles' new path and/or departure time information, feed back into the simulation system

**Step 11:** Re-run the simulation for one shot using the updated vehicle and path information as input, and analyze system benefit as well as other performance statistics.

### 3.2 Marginal Cost calculation

In this research, the volume-delay function is used to establish the relations between traffic demand and travel time, see Eq. (1):

\[
S_{at}(v_{at}) = t_a \left(1 + \alpha \left(\frac{v_{at}}{C_a}\right)^\beta\right)
\]

(1)

with

- \(t_a\): free flow travel time on link \(a\)
- \(v_{at}\): traffic demand on link \(a\) at time \(t\)
- \(C_a\): capacity of link \(a\)
- \(S_{at}(v_{at})\): The average travel time for vehicles on link \(a\) at time \(t\)
- \(\alpha\): coefficient
- \(\beta\): exponent

The key variables \(\alpha, \beta\) in Eq. (1) can be calibrated offline using the available dataset from real world, including link travel time, traffic counts, link length, link speed limit, etc.. Relevant research can be found in Cetin et al. (Cetin et al. 2012; Foytik et al. 2013). The parameters of the volume-delay function are calibrated to be static, but such volume-delay function can establish the relations between time dependent travel demand and link travel time.

Based on the volume-delay function, the link marginal cost could be derived by taking the derivative of \(v_a\); see the following Eq. (2):
\[ MC_t = \frac{\delta (v_{at} \cdot S_{at}(v_{at}))}{\delta v_{at}} = S_{at}(v_{at}) + t_a \cdot \frac{\delta (S_{at}(v_{at}))}{\delta v_{at}} \]

\[ = t_a \left( 1 + \alpha \left( \frac{v_{at}}{c_a} \right)^\beta \right) + t_a \cdot \alpha \cdot v_{at} \cdot \beta \cdot \left( \frac{v_{at}}{c_a} \right)^{\beta - 1} \cdot \frac{1}{c_a} \]  

(2)

The marginal cost calculated from Eq. (2) is local static link marginal cost. It includes the travel time that the additional driver experiences on that link (the first part in Eq. (2)), and corresponds to the first level of marginal cost defined by Ghali and Smith (1992). The second part in Eq. (2) is the additional delay experienced by all vehicles that traverse that link after that vehicle, which corresponds to the second level of marginal cost in the same definition.

The reason of using BPR function is it takes in travel demand as input to compute the travel time, which is more reasonable than volume as volume could be constrained by the upstream link capacity. Computing the exact marginal cost of adding one vehicle into the network could very challenging and is not the focus of this research. The point of using BPR function is to compute the marginal cost for different travel choice at a comparable level, and find out the best departure time and route choice with least marginal cost for each individual driver for traffic assignment purpose.

As illustrated in Figure 3-1, at the next step, TDSP will be sought based upon link marginal cost, and the marginal cost on that local link will expand to affect the whole path of that vehicle. Then, assigning the optimal TDSP to each individual vehicle will lead to the changes of links volume along the path, and the marginal cost will need to be recalculated as well. In the end, every vehicle to be assigned in the network at a time later than that previous vehicle will be affected. With this procedure, the marginal cost brought about by that particular user will be expanded to affect all other vehicles of interest at the global network level, which corresponds to the third and fourth levels defined by Ghali and Smith.

### 3.3 TDSP algorithm
A revised Time-Dependent A* algorithm is developed to search for a time dependent shortest marginal cost path for each experiment vehicle. In this revised algorithm, its uniqueness lies in that the time-dependent marginal cost is used as the link travel cost when searching for the shortest path, which makes the algorithm different from the standard A* search algorithm. Three major modifications are listed below:

**Modification 1:** Each link will keep two cost values at the same time; the first one is the travel time cost denoted as $tt_{ij}$, another one is marginal cost value denoted as $mc_{ij}$ which is computed from section 3.2. The reason for keeping two different labels for each link is, TDSP needs to know the actual time stamp that indicates when vehicles arrive at a particular link in order to retrieve the time-dependent link generalized cost, during which process both $tt_{ij}$ and $mc_{ij}$ will be used.

**Modification 2:** The link marginal cost derived by Eq. (2) is used as link cost in the TDSP algorithm instead of the link travel time cost; i.e., in the pseudo-code below, it is the marginal cost of link $(i, j)$ traversing at time $T_i$ (denoted as $mc_{ij}[\tau(T_i)]$) that is used as link cost, instead of the link travel time $tt_{ij}[\tau(T_i)]$.

**Modification 3:** Each node will keep two different time stamps $L_i$ and $T_i$. While $L_i$ is still used to record the upper bound of the minimum travel cost (which is marginal cost now) from origin to node $i$, $T_i$ is introduced to record the actual travel time cost from origin to node $i$. The reason to keep $T_i$ is that, when marginal cost is used to replace travel time as the link cost, $L_i$ is not able to record the arrival time at node $i$ any more, but this information is still needed to retrieve the time-dependent traffic network information.

The pseudo-code of the revised A* TDSP is given as below:
Step 1) Initialization

$L_0 = t_0, F_0 = e_{od}, T_i = t_0$;

$L_j = \infty, F_j = \infty, i \neq o, T_i = \infty$;

$S = \{o\}; \overline{S} = \emptyset$.

Step 2) Node selection

$i = \arg\min_{j \in S} F_j, \overline{S} = \overline{S} \cup \{i\}, S = S \setminus \{i\}$.

Step 3) Stopping rule

If $i = d$, then stop; otherwise, continue.

Step 4) Update $F_j$, travel cost label $L_j$ and time cost label $T_i$

For each $j \in A(i)$:

$L_i + mc_{ij}[\tau(T_i)] + e_{jd} < F_j$

Then

$L_j = L_i + mc_{ij}[\tau(T_i)];$

$T_j = T_i + tt_{ij}[\tau(T_i)]$

$F_j = L_i + mc_{ij}[\tau(T_i)] + e_{jd}$

If $j \notin S, S = S \cup \{j\}$, go back to step 2.

where:

- $i, j$: index of nodes
- $S$: The closed set which includes nodes already evaluated
- $\overline{S}$: The open set which includes the tentative nodes to be evaluated
- $t_0$: Departure time at origin node $o$
- $L_i$: The upper bound of the minimum travel cost from origin to node $i$
- $e_{ij}$: The estimated lower bound of the minimum travel cost from node $i$ to node $j$
- $F_i$: The fake label which is the sum of $L_i$ and $e_{id}$, which stands for the estimated minimum travel cost from origin to destination going through node $i$
- $mc_{ij}[\tau(T_i)]$ is the marginal cost of link $(i,j)$ traversing at time $T_i$.
- $T_i$: Travel time cost from origin to node $i$
- $\tau(T_i)$: Time index of timestamp $T_i$

### 3.4 Schedule Delay

Schedule delay (SD) measures the difference between the actual arrival time and travelers’ preferred arrival time (PAT), and is also an essential part of the trip timing decision process (Wang 2011). It is an attribute of a given scheduling choice which measures the deviation inherent in that choice between actual and preferred arrival time, where the deviation may be due either to early or late arrival. Arriving early is likely to involve some time wasted or reduced productivity, and thereby decreases utility, while arriving late usually indicates considerable repercussions and higher penalties (Small 1982).

Figure 3-2 briefly describes the schedule delay function associated with the arrival time, the ideal arrival time window is represented as $[PAT-, PAT+]$, with $PAT-$ as the lower bound for arriving at the destination without incurring schedule delay and $PAT+$ the upper bound. The schedule delay will increase with the increase in arrival deviation.
A variety of schedule delay formulations and parameter setting were used in the past studies (Ran et al. 1992; Bernstein et al. 1993; Wie et al. 1995). A piecewise linear formulation was adopted in this research to represent the schedule delay cost for travelers with an \((o, d, PAT)\) triplet:

\[
SD = \begin{cases} 
\beta_1 * (PAT^- - \tau - TT) & \text{if } \quad PAT^- > \tau + TT \\
0 & \text{if } \quad \tau + TT \leq PAT^- \leq PAT^+ \\
\beta_2 * (\tau + TT - PAT^+) & \text{if } \quad \tau + TT > PAT^+ 
\end{cases}
\]

(3)

Where:

- \(\beta\) is a parameter for measuring the gap between actual arrival time and PAT. \(\beta_1\) and \(\beta_2\) represent early and late cost parameters, respectively.
- \(\tau\): departure time of the trip
- \(TT\): travel time of the trip

The empirical survey shows \(\beta_2 > \beta_1 > 0\) which means that, with the same unit of deviation, the cost generated from being early will be smaller than the penalty due to lateness; this conclusion is true for travelers who usually experience more severe punishment for being late. In this research, we are going to follow this conclusion and use the recommended values for those parameters from the existing literature.

It is noteworthy that different trip purposes may have different schedule delay associated with the same travel plan. For example, a trip to catch a flight is likely to incur a higher schedule delay late penalty, but not as much penalty when arriving early. The case study below includes a scenario to represent this situation with a very high late arrival penalty.

### 3.5 Optimal Path Finding

Generally speaking, for the departure time and route choice problem, travel time and arrival schedule delay are the two travel impedances considered by travelers. For the travelers who are flexible, their actual travel time \(TT\) depends on not only the departure time \(\tau\) but also the route \(r\) to be taken, and the time of arrival will be affected by both departure time \(\tau\) and travel time \(TT\). In other words, the problem of finding the optimal travel plan will in the end come down to the decision of finding out the best departure
time $\tau'$ and route choice $r'$ within the allowed departure time window, so that the total travel cost is minimized.

If we follow the general trend and name this type of travel cost as generalized cost, containing cost from travel time $TT$ as well as from arrival schedule delay $SD$, the generalized cost expression can be depicted as follow:

$$GC_t = \alpha \ast TT_{\tau'} + SD_{\tau'}$$ \hspace{1cm} (4)

Usually the parameter of travel time cost $\alpha$ is intentionally set to be greater than $\beta_1$ but less than $\beta_2$, so as to prevent vehicles from needlessly spending more time in the network to avoid early penalties.

As illustrated in Figure 3-1 and Section 3.1, the purpose of Optimal Path Finding is to find, for each vehicle, the best departure time $\tau'$ and route choice $r'$ within the allowed departure time window, so that the generalized cost can be minimized.

$$GC_{t,r'} = Min \ GC_{t,r'}$$

It is worth noting that the schedule delay cannot be evaluated until the trip has been completed, by which time the trip arrival time will be available. In this research, departure time choices for each individual traveler are simulated with small incremental steps within the allowed departure time window, and TDSP is sought for. The travel time and schedule delay will then be computed, and the one with the minimum generalized cost will be selected as the optimal departure time and route choice for the traveler.

4 CASE STUDY

The case study described in this section was conducted to illustrate the system performance and behavior changes of the proposed System Optimal model. We chose the Tucson I-10 network to be our case study region. The network is constructed in a dynamic traffic assignment (DTA) simulation software DynusT (Dynamic Urban Simulation for Transportation) (Chiu et al. 2011); the resulting network consists of 395 nodes, 830 links and 80 traffic zones. In this network, freeway Interstate-10 runs across the map
diagonally from the northwestern corner to southeastern direction. Figure 4-1 shows the Tucson network in DynusT GUI.

Figure 4-1: Tucson I-10 Network GUI

In order to test fully the performance of the proposed model under various circumstances and understand better the effect of different behavior rule changes, 15 scenarios are designed in this section. For the sake of simplicity, the schedule delay is set to be zero for most case study scenarios except the ones specially noted. A complete list of the scenarios and their brief descriptions are as follows:

- **Baseline Scenario:** The scenario that represents the baseline traffic condition, in this research we use User Equilibrium flow as the baseline scenario which is generated by running DynusT DTA simulation software to UE with O-D demand as the simulation input.

- **Scenario 1 (SC1):** Simulate the scenario that 10% of travelers are willing to change their travel behavior by as much as 15 minutes.
- Scenario 2 (SC2): Simulate the scenario that 10% of travelers are willing to change their travel behavior by as much as 30 minutes.
- Scenario 3 (SC3): Simulate the scenario that 10% of travelers are willing to change their travel behavior by as much as 60 minutes.
- Scenario 4 (SC4): Simulate the scenario that 10% of travelers are willing to change their travel behavior by as much as 120 minutes.
- Scenario 5 (SC5): Simulate the scenario that 5% of travelers are willing to change their travel behavior by as much as 120 minutes.
- Scenario 6 (SC6): Simulate the scenario that 20% of travelers are willing to change their travel behavior by as much as 120 minutes.
- Scenario 7 (SC7): Simulate the scenario that 30% of travelers are willing to change their travel behavior by as much as 120 minutes.
- Scenario 8 (SC8): Simulate the scenario that 50% of travelers are willing to change their travel behavior by as much as 120 minutes.
- Scenario 9 (SC9): Simulate the scenario that 10% of travelers are willing to change their travel behavior, and that the degree of their travel behavior change follows a more realistic normal distribution. The parameters of the normal distribution is described in Figure 4-1.
- Scenario 10 (SC10): Simulate the scenario that 10% of travelers are willing to change their travel behavior, and that the degree of their travel behavior change follows a more realistic normal distribution. The difference from Scenario 9 is that it starts from a non-UE solution which is more similar to the real world traffic flow. The parameters of the normal distribution is described in Figure 4-1, and it is assumed 80% travelers are traveling as UE pattern.
- Scenario 11 (SC11): Simulate the scenario that 10% of travelers are willing to change their travel behavior, and the degree of their travel behavior change follows a more realistic normal distribution. It starts from a non-UE solution which is further from UE flow than Scenario 10. It is assumed only 60% travelers are traveling as UE pattern.
- Scenario 12 (SC12): Simulate the scenario that 10% of travelers are willing to change their travel behavior, and the degree of their travel behavior change follows a normal distribution with larger departure time shift flexibility than Scenario 9. The parameters of the normal distribution is described in Figure 4-1.
- Scenario 13 (SC13): Simulate the scenario that 10% of travelers are willing to change their travel behavior by as much as 60 minutes. But travelers’ travel decisions are subject to the schedule delay. The parameters of SD is $\alpha=0.4$, $\beta_1=0.7$ and $\beta_2=0.2$.

- Scenario 14 (SC14): Simulate the scenario that 10% of travelers are willing to change their travel behavior by as much as 60 minutes. But travelers’ travel decisions are subject to the higher schedule delay penalty than Scenario 13. The parameters of SD is $\alpha=0.4$, $\beta_1=0.7$ and $\beta_2=10$.

The purpose of designing different scenarios and the way they will be used in the case study is shown in Table 4-1. The following effects of different behavior rule changes are tested and analyzed in this research.

- Effect of maximum shift of time window. To test the scenarios in which the number of travelers who are willing to change the travel behavior is fixed, the only variable is the travelers’ maximum shift in their departure time window.

- Effect of flexible traveler percentage. To test the scenarios in which the maximum departure time shift window is fixed, the only variable is the number of travelers who are willing to change travel behavior.

- Combined flexible traveler percentage and willingness to shift departure time. To test the scenarios in which both the maximum shift in time window and the percentage of travelers willing to make changes can vary, a couple of scenarios with different combination will be tested.

- Effect of multiple user classes: To test the scenarios in which the initial solution is not User Equilibrium due to the existence of different user classes in reality.

- Effect of travel behavior distribution: To test the scenarios in which traveler’s departure time shift flexibility is heterogeneous and subject to normal distribution which is more realistic instead of assuming all travelers are homogeneous.

- Consideration of schedule delay: To test the scenarios to see if schedule delay is also taken into consideration during the travel decision making process.

The following sections in this chapter will go through the test list one by one and analyze the data in detail.
<table>
<thead>
<tr>
<th>Simulation Period (4hr)</th>
<th>Baseline</th>
<th>SC1</th>
<th>SC2</th>
<th>SC3</th>
<th>SC4</th>
<th>SC5</th>
<th>SC6</th>
<th>SC7</th>
<th>SC8</th>
<th>SC9</th>
<th>SC10</th>
<th>SC11</th>
<th>SC12</th>
<th>SC13</th>
<th>SC14</th>
</tr>
</thead>
<tbody>
<tr>
<td>Effect of Max Shift of Time Window</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Effect of Flexible Traveler percentage</td>
<td>*</td>
<td></td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Combined Flexible traveler percentage and Willingness to Shift</td>
<td></td>
<td>*</td>
<td>*</td>
<td>*</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Effect of Multiple User Classes</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>*</td>
<td>*</td>
<td>*</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Effect of Travel Behavior Distribution</td>
<td>*</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>*</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Consideration of Schedule Delay</td>
<td>*</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>*</td>
<td>*</td>
<td></td>
</tr>
<tr>
<td>Max Shift Period (+-min)</td>
<td>/</td>
<td>15</td>
<td>30</td>
<td>60</td>
<td>120</td>
<td>120</td>
<td>120</td>
<td>120</td>
<td>120</td>
<td>120</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Flexible Traveler Percentage</td>
<td>/</td>
<td>10.0%</td>
<td>10.0%</td>
<td>10.0%</td>
<td>5.00%</td>
<td>20.0%</td>
<td>30.0%</td>
<td>50.0%</td>
<td>10.0%</td>
<td>10.0%</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
4.1 Baseline Scenario

In baseline scenario, the simulation time length is set to be 4 hours in total, and a demand profile is constructed with the second hour as the peak demand hour. The traffic network is run to User Equilibrium after 20 simulation iterations using O-D demand as the simulation input. All other scenarios listed in Table 4-1 are derived from the baseline scenario dataset.

Heat mapping is used in this research to show the temporal-spatial speed profile of the I-10 freeway corridor in the case study network. The X axis stands for the time and the Y axis represents the physical road segment along the corridor; the color in the heat map describes the congestion levels in the traffic network, with blue as free flow condition and red as severely congested.

From Figure 4-2 and Figure 4-3, we can find in the baseline scenario result, that the congestion starts from the 2nd hour and lasts 1-2 hours depending upon the locations, then the network cools down and the traffic becomes free flowing again at the end of simulation.

Some other detailed numbers of interest in the baseline scenario result are:
• The total system travel time in the whole simulation period is 131,639 hours
• Average travel time for all travelers in the simulation period is 47.4 min
• Average travel time for peak hour travelers is 49.7 min

4.2 Typical Scenario Case study

The proposed algorithm in Section 3 is coded in Java programming language and integrated with DynusT DTA simulation software. In this section, scenario 1 will be selected as the typical case study scenario; its system performance and behavior changes will be analyzed in detail. In this scenario, it is assumed 10% of travelers are willing to change their travel behavior and shift their departure time as much as 15 minutes. Those travelers are selected completely randomly.

4.2.1 Traffic congestion alleviation

To show the traffic congestion alleviation after the behavior changes, the heat maps of freeway I-10 Northbound and Southbound before and after are drawn and shown below.

![Figure 4-4 Northbound Scenario Heat Map – Baseline (L) VS Scenario 1 (R)](image-url)
The left side of Figure 4-4 and Figure 4-5 shows the heat map on the freeway for the baseline scenario, and the right side shows the heat map for scenario 1 after the behavior change. It can be observed from the heat maps on the right side that the color of the congested area becomes lighter and the size of the colored area also shrinks, which represents the result that traffic congestion along the freeway corridor has been alleviated noticeably.

4.2.2 Peak hour demand reduction

Further analysis of the dataset reveals more insight into the result. The first observation is on the peak hour demand reduction. Analysis shows that, compared with the baseline scenario, the traffic demand in the peak hour is reduced by 8.4% in the case study result.

Those travelers who changed their travel behavior are divided further into several categories depending on their behavior change. Figure 4-6 shows the number of travelers from different categories.
Among all travelers in the network, 4.3% travelers changed their departure time but stayed on the same route. For those travelers, 3.7% of them chose to leave earlier than they used to, and the other 0.6% postponed their trip for a later time.

Only a small portion of travelers chose to take a different route but leave at the same time, most likely due to the reason that peak hour traffic was so congested and when they had flexibility at departure time, leaving at a different time will yield greater benefits.

The number of travelers who not only changed their departure time but also traveled on a route that is different from their previous path accounts for 3.9% of the whole population.

Travelers who changed their departure time not only removed themselves from the congestion, but also contributed to reduce the peak hour traffic demand. Thus, in the end, the whole system benefits from their behavior changes. The following section will analyze the benefit each traveler category gets quantitatively.

4.2.3 Travel time saving
Due to the fact that a certain percentage of travelers changed their travel behavior by either shifting the departure time or changing the route they take, we know from the Section 4.2.1 that traffic congestion is alleviated and the system is experiencing lower total travel time, which means drivers in the network are benefiting from their behavioral change by saving travel time.

The system travel time comparison between the baseline scenario case and what occurs after the scenario 1 behavior change is shown in Figure 4-7. The total system travel time decreases by 13.9%, dropping from 131,639 to 113,384 hours. The total travel time saving after the behavior change is 18,255 hours.

![Total System Travel Time Comparison](image)

**Figure 4-7 Total system travel time comparison**

If we divide the travelers into several categories according to their behavior changes, the travel time saving stats will be computed for the different categories respectively.

- **Group 1** - travelers who changed their departure time but stayed on the same route can reduce their own travel time by 20.9 min on average – 21 min for leaving earlier and 2.8 min for leaving later.
- **Group 2** - travelers who changed the route they took but not the departure time can reduce their own travel time by 11.9 min on average.
- **Group 3** - travelers who changed both departure time and their route choice can reduce their own travel time by 18.3 min on average.
- Group 4 - other travelers who did not make any changes to either departure time or route choice can benefit with a 4.0 min travel time reduction, although they did not do anything to contribute to the traffic reduction.

Among all 13688 travelers who changed their travel behavior by either leaving at a different time or taking a different route or both, 200 of them are randomly selected to show their travel time comparison before and after the behavioral changes, and the comparison result can be found in Figure 4-8. We can see that

![Travel time comparison before and after the behavior change](image)

**Figure 4-8 Travel time comparison before and after behavior change**

For the travelers who did not change anything, it turns out, due to other travelers’ behavior changes, they are also experiencing less traffic and can arrive at the destination sooner. As a matter of fact, the average benefit a single traveler can generate by changing either departure time or route choice (with that benefit defined as total travel time reduction of the whole network divided by the number of travelers who changed their travel behavior) is 80.0 minutes. Of these 80.0 minutes, they can save 20.2 minutes for themselves on average, and the other 59.8 minutes travel time saving will go to other travelers. This result
further demonstrates that travelers can not only do good for themselves by removing themselves from the traffic, but that the other drivers in the network can also greatly benefit from their behavior change.

4.2.4 Monetary saving and cost

If we translate the travel time saving into monetary value, the saving brought about by the small group of users’ behavior change can be huge. Here, the value of time from the calculations of the Texas Transportation Institute (TTI) will be used to calculate the monetary saving.

The 2011 Annual Urban Mobility report produced by TTI estimated the congestion cost to be 16.79$ per hour of person travel (TTI 2011). If we use that recommended value, the total annual monetary saving will be:

\[
(131,639 - 113,384) \times 2 \text{ peak hours/day} \times 16.79\$/\text{hour} \times 260 \text{ workdays/year} \\
= 159 \text{ million/year}
\]

The cost involved in this system varies and depends on the specific strategy to be used to induce the change in travel behavior, but generally speaking it will be relatively cheap. For example, in the Spitsmijden experiment in Europe (Knockaert et al. 2007), the cost of persuading travelers to make behavior changes is between 3.99$ per hour to leave earlier and 4.24$ per hour to leave later, which translates to a cost of $7.4 million/year.

\[
(\text{Number of travelers shifting earlier} \times 4.24 \$/\text{hour} + \text{Number of travelers shifting later} \times 3.99 \$/\text{hour}) \times 15\text{ min/(60 min/hour)} \times 2 \times \text{peak hours/day} \times 260 \text{ weekdays/year} \\
= 7.4 \text{ million/year}
\]

A similar experiment was also carried out in the US and the cost involved to persuade travelers to make travel behavior changes is 13.33 $/hour for leaving earlier and 12.62$ for leaving later (Leblanc and Walker 2013), which gives us $23.4 million in total cost:

\[
\text{Number of drivers shifting earlier} \times 13.33 \$/\text{hour} + \text{Number of drivers shifting later} \times 12.62 \$/\text{hour} \times 15\text{ min/(60 min/hour)} \times 2 \times \text{peak hours/day} \times 260 \text{ weekdays/year}
\]
= 23.4 million/year

Compared with the total monetary saving, relatively lower cost can be computed from both research and yields a good Return on Investment (ROI), which indicates the proposed framework is able to reduce the traffic congestion greatly and save a huge amount of money with relative low cost.

4.3 Effect of Max Shift of Time Window

In this section, we fix the percentage of drivers willing to make changes to be 10%, so the only parameter to be changed is the maximum length of time they are willing to shift. Four scenarios are designed and implemented to make the comparison, with the maximum length of the change in departure time to be 15min, 30min, 60min, and 120min respectively, which correspond to SC1, SC2, SC3 and SC4 in Table 4-1.

The following figures, Figure 4-9 and Figure 4-10, show the heat map along the I-10 corridor. The scenarios from left to right are: Baseline scenario, SC1, SC2, SC3 and SC4 respectively. We can tell from the diagrams that, with increases in the allowed maximum departure time shift, the color of the diagram becomes lighter and that network is less congested.
Figure 4-9 Changes of heat map with different max shift of time window – I-10 NB (Baseline, SC1, SC2, SC3, SC4)

Figure 4-10 Changes of heat map with different max shift of time window – I-10 SB (Baseline, SC1, SC2, SC3, SC4)
Those drivers who changed their travel behavior are divided further into several categories depending on the way in which their behavior changed. Figure 4-11 shows the changes in the number of drivers from different categories. Since the total number of drivers willing to make change is fixed to be 10%, the travel behavior change in Figure 4-11 isn’t significant except when the max departure time shift allowed increased to 120 minutes.

![Travel Behavior Changes](image)

**Figure 4-11 Travel behavior changes with different max shift of time window**

The travel time for each group before and after the behavior change is shown in Figure 4-12. We can see that for drivers with different behavior changes, the travel time saving also varies. Also it can be observed that the more flexible the departure time choice is, the more travel time reduction that can be expected. Drivers who change departure time or route can expect to reduce their travel time for about 30 minutes, while the drivers who did not do anything can get a benefit of only 5 minutes of travel time decrease on average.
The system total travel time and travel time saving under different circumstances are also analyzed. We can observe from Figure 4-13 that there is significant travel time saving when drivers are able to make changes of even as little as 15 minutes. With drivers having more flexibility in their departure time choice, the benefit we can get is also increasing, although the marginal benefit becomes smaller.

Figure 4-12 Changes of travel time with different max shift of time window

Figure 4-13 Total travel time changes with different max shift of time window
The monetary saving is also increasing when drivers have more flexibility in their departure time choice; as observed from the diagram below, the benefit we can get is overwhelmingly larger than the potential cost needed.

![Figure 4-14 Total monetary saving with different max shift of time window](image)

4.4 Effect of flexible traveler percentage

In this section, we fix the maximum length of time that travelers are willing to shift to be 120 minutes, so the only parameter to be changed is the percentage of drivers willing to make changes. Five scenarios are designed and implemented to make the comparison, with the percentage of drivers flexible in their travel decision percentage being 5%, 10%, 20%, 30%, and 50% respectively, which correspond to SC5, SC4, SC6, SC7, and SC8 in Table 4-1.

The following figures, Figure 4-13 and Figure 4-14, show the heat map along the I-10 corridor. The scenarios from left to right are: Baseline scenario, SC5, SC4, SC6, SC7, and SC8 respectively. We can tell from the diagrams that, with more drivers willing to change their travel behavior, the area of color becomes smaller and the color of the diagram is lighter as well, results which indicate the congestion in the network is significantly alleviated.
Figure 4-15 Changes of heat map with different flexible traveler percentage – I-10 NB (Baseline, SC5, SC4, SC6, SC7, SC8)

Figure 4-16 Changes of heat map with different flexible traveler percentage – I-10 SB (Baseline, SC5, SC4, SC6, SC7, SC8)
Those travelers who changed their travel behavior are divided further into several categories depending on the way in which their behavior changed. Figure 4-17 shows the changes in the number of drivers from different categories. Since the percentage of drivers willing to make a change increases from 5% all the way to 50%, correspondingly the number of drivers who changed their travel behavior also increases significantly.

![Travel Behavior Changes](image)

**Figure 4-17** Travel behavior changes analysis with different percentages of flexible travelers

The travel time for each group before and after the behavior change is shown in Figure 4-18. We can see that for drivers with different behavior changes, the travel time saving also varies. Also it can be observed that the more drivers willing to change their travel behavior, the more travel time reduction that can be expected.
Figure 4-18 Changes of travel time with different percentages of flexible travelers

On the one hand, for those drivers who made changes, the average travel time saving for them was very high at first and starts to decrease when more drivers are willing to make changes. This is rather understandable because, at the very beginning, the network is so congested that when drivers remove themselves from traffic, the travel time saving can be very high. But when more of them are trying to avoid the congestion and travel wisely, the non-peak hour traffic becomes somewhat congested and, with its difference from with peak traffic travel time being therefore smaller, so is the benefit to those drivers of changing their travel behavior.

On the other hand, for the drivers who did not make changes but benefited from the other drivers’ behavior change, the benefit is increasing quickly due to the fact that traffic in peak hour is becoming lighter compared with the initial condition, so even if they did not do anything good, the benefit they get is also increasing. In the end, the travel time saving for the two groups of drivers is becoming close.

The total travel time saving under different circumstances are also analyzed. We can observe from Figure 4-19 that there is significant travel time saving when more drivers are willing to make changes. With more drivers having flexibility regarding their travel, the total benefit that the system can get is also increasing.
Figure 4-19 Total travel time changes with varying flexible traveler percentage

The monetary saving is also increasing when drivers have more flexibility in their departure time choice; as observed from the diagram below, the benefit we can get is overwhelmingly larger than the potential cost needed. It is also noteworthy that, with more drivers changing their behavior, the cost of needed incentives is also increasing.

Figure 4-20 Total monetary saving with different flexible traveler percentage

4.5 Combined flexible traveler percentage and willingness to shift
In this section, both the maximum shift of time window and the flexible driver percentage can be changed. Four scenarios are used to make the comparison: SC1 (10% + 15 min), SC3 (10%+60min), SC6 (20%+120min), SC7 (30%+120min) in Table 4-1.

The following figures, Figure 4-21 and Figure 4-22, show the heat map along the I-10 corridor. The scenarios from left to right are: Baseline scenario, SC1, SC3, SC6, SC7 respectively. We can tell from the diagrams that with more drivers willing to change their travel behavior and the more flexibility they possessed, the colored area becomes smaller and the color of the diagram gets lighter as well, which indicates the congestion in the network is significantly alleviated.
Figure 4-21 Changes of heat map with behavior change combinations – I-10 NB (Baseline, SC1, SC3, SC6, SC7)

Figure 4-22 Changes of heat map with behavior change combinations – I-10 SB (Baseline, SC1, SC3, SC6, SC7)
Those drivers who changed their travel behavior are divided further into several categories depending on the ways in which their behavior changed. Figure 4-23 shows the changes in the number of drivers from different categories. Since the percentage of drivers willing to make change increases as well as their max shift of time window, correspondingly the number of drivers who changed their travel behavior also increased significantly.

![Travel Behavior Changes](image)

**Figure 4-23 Travel behavior changes analysis with behavior change combinations**

The travel time for each group before and after the behavior change is shown in Figure 4-23. We can see that for drivers with different behavior changes, the travel time saving also varies. Also it can be observed that the more that drivers are willing to change their travel behavior, or the more that they are willing to change the departure time, the more the reduction in travel time that can be expected.
The total travel time savings under different circumstances are also analyzed. We can observe from Figure 4-25 that there is a significant travel time saving when more drivers are willing to make changes or shift departure time to a larger degree. With more drivers having flexibility regarding their travel, the total benefit the system can get is also increasing.
travel behavior change alone, which indicates the marginal system benefit is the highest when drivers start to change their behavior, and decreases when more behavior changes are made.

The monetary saving is also increasing when drivers have more flexibility in their departure choice; as observed from the diagram below, the benefit we can get is overwhelmingly larger than the potential cost needed.

![Figure 4-26 Total monetary saving with behavior change combinations](image)

**4.6 Effect of travel behavior distribution**

In the previous case study, we assume that all drivers are homogenous in their travel behavior, which means they are all acting exactly the same way and able to shift departure time exactly the same amount as the other drivers do. For example, in Section 4.2, we assume all drivers are able to shift departure time as much as 60 minutes, which may not necessary hold in reality.

Two more realistic distributions are assumed in this section to better approximate the behavior of the drivers in our case study. Generally speaking, drivers usually have an equal tendency to leave earlier or later, which makes a normal distribution with mean value close to zero and variance close to drivers’ max departure time shift window a reasonable approximation to the reality. The first case (upper graph of Figure 4-27) was assumed to have a normal distribution with mean value 0.4 and standard deviation 16 minutes.
The second case is designed to allow more behavior changes, which means the travel behavior curve should be relaxed to be a flatter shape. Normal distribution with mean value 0.7 and standard deviation 21 is assumed to represent the second realistic travel behavior, second graph in Figure 4-27 shows its probability distribution curve.

Figure 4-27 Distribution of travel behavior changes
Corresponding to the two distributions constructed above, two scenarios are designed and implemented to make the comparisons, which correspond to SC9, SC12 in Table 4-1.

The following figures, Figure 4-28 and Figure 4-29, show the heat map along the I-10 corridor. The scenarios from left to right are: Baseline scenario, SC9, and SC12 respectively. Observations from the diagrams tell us that the traffic congestion in SC9 and SC12 has been alleviated noticeably compared with the baseline scenario. It can also be observed that, when more behavior changes are allowed as in SC12, the traffic condition can be much lighter than that in SC9.
Figure 4-28 Changes of heat map with travel behavior distribution – I-10 NB

Figure 4-29 Changes of heat map with travel behavior distribution – I-10 SB
Those drivers who changed their travel behavior are divided further into several categories depending on the ways in which their behavior changed. Figure 4-30 shows the changes of the number of drivers from different categories. Since the total number of drivers willing to make a change is fixed to be 10% for both the SC9 and SC12 scenarios, the travel behavior change in Figure 4-30 is not significant between them.

![Travel Behavior Changes](image)

**Figure 4-30 Travel behavior changes with travel behavior distribution**

The travel time saving for each group before and after the behavior change is shown in Figure 4-31. We can see that for drivers with different behavior changes, the travel time saving also varies. It can be observed that drivers in both scenarios are able save travel time compared with the baseline scenario, but the difference between SC9 and SC12 is small.
The system total travel time and travel time saving under different circumstances are also analyzed. We can observe from Figure 4-32 that both scenarios are able to save travel time compared with the baseline scenario. While the behavior change distribution curve is flatter in SC12 and drivers have more flexibility, its total travel time saving is slightly larger than SC9.
The monetary saving under each of these two scenarios is around US$60 million per year, which is significantly high. Observed from the diagram below, the benefit we can get is overwhelmingly larger than the potential cost needed.

![Monetary saving VS cost](image)

Figure 4-33 Total monetary saving with travel behavior distribution

### 4.7 Effect of multiple user classes

In the previous case study, we assume that the baseline scenario is the User Equilibrium, which is not necessarily the case in reality. Some key preconditions of user equilibrium are: users have complete information about the whole network and traffic condition, and all drivers in the network are willing to follow exactly the shortest path, so that in the end none of the drivers will be able to reduce their travel time by changing the route. It is generally accepted that these assumptions do not hold in reality, as drivers are usually constrained during the travel and will not be able to follow the best route and leave at the best departure time, so the traffic condition in real world is not a User Equilibrium flow.

Drivers in reality are usually constrained either by departure time or the paths to take; for example, the path of housewives who need to do grocery shopping or dry cleaning is usually fixed, so the only aspect in which they have flexibility is the time to leave. However, for some of the daily commuters, it is easier for them to change the route but not the time to leave, which means they are flexible only in term of the
route choice. So, the travel behavior of different driver classes will be significantly different from each other.

In this section, two scenarios are designed to simulate the conditions that are close to the reality.

Scenario 10 has a higher ratio of flexible drivers, 80% of them will be travelling at the User Equilibrium path, 10% of drivers are flexible with respect to departure time but not route, and the remaining 10% of drivers are flexible in route choice but not departure time.

Scenario 11 has a lower ratio of flexible drivers, 60% of them will be travelling at the User Equilibrium path, 20% of drivers are flexible with departure time but not route, and the remaining 20% of drivers are flexible in route choice but not departure time.

In this section’s case study, three scenarios are used to illustrate further the system performance under different circumstances. Besides the above-mentioned SC10 and SC11, SC9 – which assumes all drivers are in a User Equilibrium user class – will also be included in this section.

The following figures, Figure 4-34 and Figure 4-35, show the heat map along the I-10 corridor. The scenarios from left to right are: SC9, SC10, and SC11 respectively. We can see from the diagrams that, from the left to right, on I-10 northbound the congested area shrinks but the color becomes darker, while southbound the congested area is slightly expanded and the color darkness is about the same.
Figure 4-34 Changes of heat map with multiple user classes – I-10 NB

Figure 4-35 Changes of heat map with multiple user classes – I-10 SB
Those drivers who changed their travel behavior are divided further into several categories depending on the ways in which their behavior changed. Figure 4-36 shows the changes in the number of drivers from different categories. Since the total number of drivers willing to make a change is fixed to be 10% in all three scenarios, the number of drivers who changed departure time and drivers who changed both departure time and route remain about the same for all three scenarios. However, the number of drivers who changed route increased dramatically, which is understandable because SC10 and SC11 did not start from User Equilibrium flow, so drivers can reduce their travel time by changing the route.

![Travel Behavior Changes](image)

**Figure 4-36 Travel behavior changes with multiple user classes**

The travel time saving for each group before and after the behavior change is shown in Figure 4-37. We can see that for drivers with different behavior changes, the travel time saving also varies. It can be observed that for drivers who only changed route, their travel time saving increases from the left to the right side significantly. The travel time saving for drivers who changed both departure time and route also increases from left to right. But the saving for drivers who only changed departure time decreases, and drivers who did not change travel behavior can get 2.5 minutes of travel time saving in all three cases.
The system total travel time and travel time saving under different circumstances are also analyzed. We can observe from Figure 4-38 that travel time savings are all positive in all three scenarios and increase slightly from left side to right side. This finding indicates that, with the same amount of drivers changing their travel behavior, the travel time saving achieved can be different, and the worse the traffic condition is, the more saving that can be expected.

The monetary saving is also increasing slowly when drivers have more flexibility in their departure time choice, indicating that the worse the initial solution is, the more improvement and economic benefit
that can be gained. Observed from the diagram below, the benefit we can get is overwhelmingly larger than the potential cost needed.

![Monetary saving VS cost](image)

Figure 4-39 Total monetary saving with multiple user classes

We can observe from Figure 4-39 that replacing the key assumptions in the previous case study scenarios with more realistic input still yields satisfactory system performance improvement; the travel time saving is still significant for all scenarios, and the monetary saving is also overwhelmingly larger than the potential system deployment cost.

It can also be found that the average travel time value can be as low as 40.8 minutes in section 4.2 when only 10% of drivers shift departure time for 15 minutes, while the average travel time for the three scenarios in this section is 46.6, 47.3 and 49.5 respectively, which indicates the great potential of the system: the real world traffic flow is not User Equilibrium yet, but our proposed model is able to improve the traffic a lot better than User Equilibrium flow.

### 4.8 Effect of schedule delay

In this section’s case study, two scenarios are designed to illustrate further the system performance under different circumstances, which correspond to SC13 and SC14 in
Table 4-1; both take account of schedule delay penalty in the travel decision making process, but the schedule delay penalty in SC14 is higher than in SC13. The baseline scenario as well SC3 are also listed and analyzed in this section. SC3 is exactly the same scenario as in SC13 and SC14 except no schedule delay is considered.

It is assumed the number of drivers willing to change travel behavior accounts for 10% in SC3, SC13 and SC14, and the maximum time shift window is 60 min. With respect to the schedule delay, the parameters in SC13 are set to be $\alpha=0.4$, $\beta_1=0.7$ and $\beta_2=0.2$ as recommended by Wang (Wang 2011). For SC14 with higher schedule delay penalty, $\beta_1$ is set to be 10, which is very large number compared with $\alpha$ and $\beta_2$.

The following figures, Figure 4-40 and Figure 4-41, show the heat map along the I-10 corridor. The scenarios from left to right are: Baseline scenario, SC3, SC13, and SC14 respectively. We can tell from the diagrams that, on the one hand, SC3 has the clearest traffic condition; on the other hand, both SC13 and SC14 are able to reduce traffic compared with the baseline scenario.

The heat map also shows that the traffic condition in SC14 is worse than in SC13, which is rather intuitive, as when drivers are faced with higher schedule delay penalty, their travel choice will be further limited and likewise limited will be the space to improve traffic.
Figure 4-40 Changes of heat map with schedule delay – I-10 NB

Figure 4-41 Changes of heat map with schedule delay – I-10 SB
Those drivers who changed their travel behavior are divided further into several categories depending on their behavior change. Figure 4-42 shows the changes in the number of drivers from different categories.

Making a comparison going from SC3 to SC13 and given that schedule delay is considered in SC13, the number of drivers in SC13 who left earlier decreases and those who left later increases. This is due to the fact that in SC3, drivers are allowed to shift departure time as much as 60 minutes and the majority of drivers left earlier, so that when they shift departure time too much, the schedule delay penalty will be imposed, which in the end causes a certain number of drivers to leave later.

From SC13 to SC14, when a higher schedule delay penalty is imposed for arriving late, drivers have a tendency to leave earlier to avoid the high late penalty, which also explains that the number of drivers who left later decreases.

The travel time for each group before and after the behavior change is shown in Figure 4-43. We can see that for drivers with different behavior changes, the travel time saving also varies.
Also it can be observed that the drivers who left earlier can save more travel time, while the rest of the drivers have a higher travel time than in SC3, but about the same as in the baseline scenario.

![Changes of travel time](image)

**Figure 4-43 Changes of travel time with schedule delay**

The system total travel time and travel time saving under different circumstances are also analyzed. We can observe from Figure 4-44 when schedule delay is considered, the travel time saving can still be observed with drivers’ behavior changes, but much less than without the schedule delay. This is because schedule delay can be considered as a constraint to the user trip activity, if the constraint becomes harder, users will have less flexibility in their travel behavior which leads to a less ideal solution, and explains the travel time saving will be less.
The monetary saving is also increasing when drivers have more flexibility in their departure time choice, observed from the diagram below. The benefit we can obtain is overwhelmingly larger than the potential cost needed, regardless of the schedule delay being imposed or not.

We can observe from Figure 4-45 that replacing the key assumptions in the previous case study scenarios with more realistic input still yields excellent system performance improvement,
the travel time saving is still huge for all four scenarios, and the monetary saving is also overwhelmingly larger than the potential system deployment cost.

5 CONCLUDING REMARKS

This paper documents the research effort in developing a Behaviorally Induced, System Optimal model to improve the system performance towards the System Optimal condition. Both analytical derivation and numerical analysis on a hypothetical network resembling the traffic network structure in Tucson, Arizona, have been conducted. The outcome of this study shows that our proposed method offers a promising modeling approach for improving the existing network traffic condition towards System Optimal, and that the amount of economic value that can be saved by the proposed system can be huge.

The proposed algorithm incorporates the marginal cost calculation, TDSP search algorithm, and schedule delay as well as optimal path finding modules. A numerical analysis was conducted on the Tucson I-10. The benefit for the system as a whole as well as for each user class was evaluated, and the effect of different behavior change rules was analyzed in the case study. The outcome of the case study shows that the benefit of our proposed algorithm includes significant alleviation of traffic congestion, with both travel time and monetary saving, at a relatively low cost.
6 REFERENCES


APPENDIX C: A CONSTRAINTED TIME-DEPENDENT K SHORTEST PATHS ALGORITHM ADDRESSING OVERLAP AND TRAVEL TIME DEVIATION

Paper was prepared to be submitted to a journal
Theoretical research assumes all users are homogeneous in their route choice decision and will always pick the route with the shortest travel cost, which is not necessarily the case in reality. Researchers in Minnesota found that only 34% of drivers strictly traveled on the shortest path. This paper documents the research effort in developing a Constrained Time-Dependent K Shortest Paths Algorithm in order to find K Shortest Paths between two given nodes. The goal of this research is to provide sound route options to drivers in order to assist their route choice decision process, during which process the overlap and travel time deviation issues between the K paths need to be considered. The proposed algorithm balancing overlap and travel time deviation is developed in this research. A numerical analysis is conducted on the Tucson I-10 network, the outcome of the case study shows that our proposed algorithm is able to find different shortest paths with a reasonable degree of similarity and close travel time, which indicates that the result of the proposed algorithm is satisfactory.

Keywords: K shortest Paths Algorithm (KSP), Time Dependent Shortest Path, Constrained Shortest Path, Route Overlap, Travel time Deviation, Active Traffic and Demand Management (ATDM), Route Choice Model
1. BACKGROUND

A Shortest Path Algorithm (SPA) aims at solving the Shortest Path Problem (SPP), i.e., finding a path from a given origin node O to a given destination node D along the existing links such that the cost of the path is minimized. It is a problem that receives a vast amount of research and application attention in different areas, for example, Transportation, Computer Science, Geographic Information Systems, Speech Recognition and so on. Some commonly seen applications of SPA include finding, in a traffic network, the best route with least travel time from location O to location D, sending the user requested data from server A to client computer B with fastest response within the internet network and, in the speech recognition field, interpreting a sentence from audio with the least likelihood of transition between words.

In the field of transportation engineering, there is an interesting problem derived from SPP that has been widely researched theoretically and employed in real world application, which is how to find K Shortest Paths (KSP) between a given origin address O and destination address D with certain constraints; in other words, how to find several different Shortest Paths (SP) and rank them within a certain network. This problem is of particular interest due to the deficiency of adopting a single SP option in the real world; one of the most obvious examples is that use of a single such option assumes all users are homogeneous in their route choice decision and will always pick the route with the shortest travel cost. This approach, thus, failing to consider other factors involved in the decision making process, such as user preference with respect to different roads. Based on the three-week real GPS data, researchers in Minnesota found only 34% drivers strictly traveled on the shortest path; and even if the standard was relaxed to allow 10% deviation spatially, the number was merely increased to 40% (Zhu and Levinson 2012).
Various reasons might exist that cause drivers to be not necessarily travel on the shortest path in reality. Generally speaking, the driver’s trip choice decision is made based on considering several dimensions, to name a few: traffic congestion, fuel consumption, route distance, activity location constraints such as dropping off or picking up kids, preference for certain roads, and so on. So even though in the theoretical research on transportation, the SP is always sought for and assigned to the user in the traffic assignment model, it is not necessarily the path traveled in the real world.

One of the many merits of KSP research lies in its capability of providing multiple path options, which can find not only the shortest path between origin and destination addresses, but also the K-1 other paths in order of increasing travel cost. Providing the multiple path options generated by the KSP algorithm to drivers would be helpful and give them the flexibility to choose among these offered options in their trip decision making process. Another advantage of a KSP algorithm is its ability to incorporate additional constraints that cannot be solved by ordinary shortest path algorithms. Such constraints can come from drivers’ different travel behavior preferences: towards a freeway or arterial; about the recurring traffic congestion they experienced before and want to avoid; or simply more familiarity with certain roads and greater willingness to stay on those routes.

In this paper, a Constrained K Shortest Paths Algorithm to find the Time-Dependent K Shortest Paths between two given nodes will be proposed. The objective of finding KSP in this research is to provide multiple sound route options to the drivers in order to assist their route choice decision process. With this in mind, we defined the KSP satisfying the following constraints:

1. The K paths found should not be highly overlapped; otherwise they do not make much difference to the users and there is a high chance they will be perceived as the same path. However, it should also be noted that it does no harm to allow the existence of a
certain degree of overlap, as long as significant spatial deviation is observed between
the routes. This feature distinguishes the proposed algorithm from the disjoint KSP
research in previous literature, either node disjoint or link disjoint method.

2. Travel times are comparable between those routes found. If the travel time of route K
is significantly higher than route K-1, although it might still be a very different route,
it can be overwhelmingly worse in terms of travel cost than others available, a situation
which will make it unattractive to drivers, so that the chance of drivers selecting this
route becomes very low.

This research proposes an algorithm that takes into consideration the aforementioned
constraints. In addition to the algorithmic development, a numerical analysis is conducted on the
Tucson I-10 network, and the degree of similarity is calculated to show the degree of overlap
between paths found. The outcome of the case study shows that our proposed algorithm is able to
find the different shortest paths with reasonable similarity degree and close travel time, which
indicates the result of the proposed algorithm is satisfactory.

The rest of this paper is organized as follows: Section 2 reviews literature relating to
K Shortest Path, including the studies that are critical to this research. Section 3 discusses the
proposed algorithm framework and procedures in depth; the pseudocode is presented in this section
as well. Section 4 presents the case study results and analysis. Section 5 concludes this research.

2. LITERATURE REVIEW

A rich body of literature exists with respect to research that has been conducted on the
Shortest Path area since the 1950s. These Shortest Path-related research efforts can be generally
classified into four categories: (1) Static Shortest Path Algorithms (SSP), 2) Time-Dependent
Shortest Path Algorithms (TDSP), (3) Shortest Path Re-optimization Algorithms (SPR), and (4) K-
Shortest Paths Algorithms.
Research on the Static Shortest Path Algorithms dates back to the 1950s. Dijkstra proposed what is called Dijkstra's algorithm to find the Shortest Path from one node to all other nodes, with the pre-requisite of a non-negative arc cost. Bellman and Lester computed the single-source Shortest Path in a weight diagraph; their algorithm is able to work on the graph with negative cost. Hart, Nilsson and Raphael (Hart et al. 1968; Hart et al. 1972) first described the A star algorithm, which led to its widespread use because of its performance and accuracy. A star solves the 1-to-1 Shortest Path problem, and it achieves better performance by using heuristics to determine distance between two nodes. Floyd published the Floyd-Warshall algorithm to solve the problem of all-to-all Shortest Paths, which is essentially the same as the previously published algorithms by Bernard Roy in 1959 and also by Stephen Warshall in 1962 (Weisstein 2009). Johnson's algorithm (Johnson 1977) is also able to solve the all-to-all Shortest Path and, on the sparse network, it is able to work faster than the Floyd-Warshall algorithm. Dial (Dial 1969) proposed an implementation method to sort the nodes and store them in the buckets, an added feature which is able to make the Shortest Path Algorithm more efficient.

The time Dependent Shortest Path algorithm studies the dynamic Shortest Path problems, where the arc travel time changes dynamically. Back in 1966, Cooke and Halsey (Cooke and Halsey 1966) were the first to take into consideration in their algorithm the situation that, in the real world, the time required for travel between any two vertices may not be constant, so they developed a modified form of Bellman's iteration scheme to find the shortest route between any two vertices in a network for applications in the generalized case. After that, the idea of Time Dependent Shortest Path began to be popular. Ziliaskopoulos (Ziliaskopoulos 1993) proposed an algorithm to solve the all-to-one time-dependent Shortest Path problem; his algorithm calculated the path by operating backwards starting from the destination node. Chabini (Chabini 1997) proposed the most efficient
algorithm, known as DOT algorithm, to compute all-to-one shortest paths in discrete dynamic networks. He proved the new and simple solution algorithm had an optimal run time complexity.

For the Shortest Path Re-optimization topic, as early as the 1970s, research started to appear aimed at calculating the shortest path tree by updating the previous shortest path tree when the network data changes. Gallo (Gallo 1979) proposed an algorithm to update the Shortest Path Tree (SPT) when 1) the node from which the shortest paths are to be determined is changed and 2) the cost of one arc is modified (either increased or decreased). One issue that needs to be addressed in this model is that it can deal only with the scenario in which one arc cost changes; if multiple arc costs have changed, the algorithm has to be run multiple times to calculate the updated SPT. Pallottino (Pallottino and Scutella 2003) presented an algorithm to re-optimize a shortest path when a subset of arc costs have changed, which is a more general situation, and the method was proved to be more efficient than applying the standard SP algorithm again. Regarding the scenario in which the origin node changes, Florian (Florian et al. 1981) proposed a dual simplex algorithm to find all shortest paths. Given a shortest path arborescence rooted at node r, the change of root to a new origin s, renders the arborescence rooted at r dual feasible and primal infeasible for the new problem. Desrochers and Soumis (Desrochers and Soumis 1988) proposed a primal-dual re-optimization algorithm for the Shortest Path Problem with time windows, i.e., each node should be visited within time interval \([a_i, b_i]\); it can solve a problem with up to 2500 nodes and can deal with both positive and negative cost arcs.

Considerable literature can be found treating various aspects of the K-Shortest Paths research area. Martins published his Shortest Paths Ranking algorithm as early as 1983 (Martins 1983) on how to find the K-Shortest Paths between an initial and a terminal node in a network and how then to rank them according to their cost. This algorithm belongs to the class of deletion path algorithm based on the optimality principle. Martins continued to devote himself to research in the
ranking algorithm field (Azevedo et al. 1990; Azevedo et al. 1993; Deazevedo et al. 1994) in order to improve the computation efficiency. Researches about the Node-Disjoint Shortest Paths can be found in several studies (Frish 1967; Steiglite 1971; Suurballe 1974), where nodes are not allowed to repeat in the different paths found. Researches about the Edge-Disjoint Shortest Paths can be found in a number of studies (Y. Perl 1978; Suurballe 1982; Suurballe and Tarjan 1984), with the constraint of no repeating links. Yen (Yen 1971) also proposed a K-Shortest Path algorithm to find K loopless paths; the significance of his algorithm was in having computation of the upper bound increase only linearly with the value of K. Other work (Rosen et al. 1991; Chen 1993; Chen and Hung 1994; Chen 1994) studied how to find the K-Shortest paths when a capacity constraint exists on the links between two nodes. Eppstein (Eppstein 1994) developed an algorithm to find K shortest path through use of a shortest path tree and a graph to represent all possible deviations from the shortest path, and later Jimenez (Jimenez and Marzal 2003) modified this algorithm to improve its practical performance. Many other researchers also studied K-Shortest Paths algorithms from different perspectives, see (Dreyfus 1969; Wongseelastote 1976; Horne 1980; Katoh et al. 1982; Perko 1986).

Route choice modeling is an area that receives increased attention and involves KSP research and application, and these models can usually be found in Stochastic User Equilibrium, traffic assignment, OD calibration and other research fields (Prashker and Bekhor 2004). A straightforward and conventional way of finding these paths is to compute a sufficiently large number of overall shortest paths, i.e., using a path enumeration method, and select ones that satisfy the constraints or delete the ones that do not satisfy the constraints. With reference to route choice, the most commonly seen options include Multinomial Logit (MNL) and probit models (Ben Akiva and Lerman 1985; Sheffi 1985; Cascetta 1990; Ortuzar and Williansen 1990). Cascetta proposed a modified specification of the logit model (C-Logit) which overcomes the main shortcoming of
MNL, i.e., the overlap issue, by introducing a "commonality factor" in the logit utility function (Cascetta et al. 1996). However, as noted by the author, the application of the C-Logit model requires the explicit enumeration of alternative paths and the algorithm might work better if a limited number of reasonable paths, i.e., paths of comparable cost are generated.

Path enumeration approaches are, however, time consuming and CPU expensive; and researchers have been trying to deal with constructing the route choice model, while avoiding an excessive computational burden. The Probit Stochastic Method (SAM) proposed by Maher and Hughes (Maher and Hughes 1997) does not suffer from the burden of path enumeration by its assumption of a 'Markovian' routing strategy, and the capacity restraint was also considered in the SAM model. Castillo and other researchers (Castillo 2008; Castillo et al. 2008) used Bayesian network in his OD calibration model and traffic flow prediction model also without path enumeration. Zijpp and Catalano (Zijpp and Catalano 2003) proposed a method to find the feasible shortest paths directly and in combination with predefined constraints, by using the ordinary shortest path computation as its elementary operation. Other non-enumeration approach can be referred to (Bell 1995; Akamatsu 1996; Damberg et al. 2002; Qin et al. 2005).

In the last decades, KSP algorithms were also being researched and applied in other areas, such as the public transit area, which is similar to finding the K shortest paths in a network with multiple costs constrained and time-window constraints. Xu et al. (Xu et al. 2010) proposed an algorithm to enumerate the paths and identify the first K shortest paths in a schedule-based transit network, in which each node has a list of scheduled departure times for transit vehicles. Zhao et al. (Zhao et al. 2008) designed a least transfer times algorithm and the K shortest transit path algorithm in the stochastic transit network, in which the transfer times, travel time, travel time reliability and cost of each transit path plan are taken into account. Malihe Niksirat (Niksirat et al. 2012) solved the K-Shortest viable path problem in a transportation network including multiple modes by using
Greedy Algorithm Simulated Annealing and a Bi-directional searching Ant Colony System, and applied the algorithm in the Tehran public transit network. Jerald Jarlyasunant et al. (Jarlyasunant et al. 2011) developed an algorithm to calculate travel time of KSP in a public transportation network with real time traffic conditions. The trick is they had previously used their computer to create a lookup table of feasible paths between the origin stop of every bus route to the terminus of every other bus route using the major transfer points, which significantly reduced the algorithm running time. Yang, Wang et al. (Yang et al. 2012) solved the K-earliest arrival problem on timetable-information-based public transportation systems, and Sun et al. (Sun et al. 2007) used a node projecting method and two dummy links to solve the KSP problem based on this topological structure in the transit network. They also introduced a partial overlap indicator to evaluate a park and ride system and scatter the park and ride sites to alleviate the congestion. Besides transit, KSP algorithms are also being studied and applied in other transportation fields. Nikola Besinovic et al. (Besinovic et al. 2013) proposed a KSP model to allocate weigh-in-motion checkpoints while considering trucks trying to bypass checkpoints along the shortest paths. The problem is formulated as a binary program and applied to minimize the damage due to overweight trucks. Jerome Berclaz (Berclaz et al. 2011) developed an KSP algorithm to do the multi-object tracking by detecting objects in individual frames and then linking the detected items across frames.

Our problem of interest in this research is to find K Shortest Path without path enumeration, and to allow reasonable overlap while still being perceived as different routes. Also, the KSP problem to be solved is constrained by the total travel time. The goal of this proposed K shortest path is to provide meaningful and useful travel options to the drivers in order to assist them in their travel decision making process. Section 3 will present the algorithm framework and procedures in detail.

3. METHODOLOGY
In this chapter, a Constrained K Shortest Paths algorithm will be presented to find K-Shortest Paths between two given nodes satisfying the pre-defined constraints. The major input for this model is a traffic dataset of the network. This dataset includes a node and links topology, links connectivity and travel time information. Procedures used to search for the constrained K Shortest Path will be explained in detail.

3.1 Concept

The goal of this proposed K shortest path is to provide sound route options to the drivers in order to assist their route choice decision process and to do so without going through the path enumeration process. The following circumstances should be avoided in the K paths search:

1. Same paths: if the returned KSPs turn out to be the same paths, then the algorithm fails to meet the objective of finding different paths.
2. Loops in the path: having a loop in the path can make the Kth path different from K-1th path, but this indicates erroneous implementation unless negative costs are present in the link costs.
3. Inappropriately detoured path: finding a path that is inappropriately detoured in the opposite direction or through a congested area just to avoid some overlap.
4. Paths with large travel time deviation: if the travel time of the Kth path is significantly higher than K-1th path, this K-1th path is less likely to be preferred by drivers.

A successful KSP solution of interest should be loopless, with some but not excessive overlap, and with low travel time deviation. Here, we define the path choices that are meaningful and useful to the drivers as:

1. The K shortest paths can share a certain a degree of similarity, but not with excessive overlap in order not to fail to permit a noticeable difference for drivers.
2. The K shortest paths should avoid any unreasonable detours stemming from the effort to avoid overlap.
3. The travel times deviation between K paths should be a controllable parameter in order to increase the likelihood for the K paths to be desirable.
To find the KSP result that accomplishes the above goals, a Constrained K Shortest Paths Algorithm is presented in this section. The K Shortest Paths Algorithm proposed in this research includes a network update model which ensures that the Kth path will always be different from the K-1th path, as well as including a penalty function to penalize the links already in the SP solution to address the overlap issue. In the end, the travel time constraints are built into the TDSP search logic to make sure the travel time of the K paths found will be within a reasonable range of the first shortest path. This research derives from the ranking algorithm proposed by Martins (Martins 1983) with substantial improvements to avoid the issues related to overlap and travel time deviation. The following sections in this chapter will explain the research framework and procedure in detail.

Assume we have a traffic network $G = (N, A)$ where $N$ is the set of nodes $\{1, 2, \ldots, n\}$ and $A$ is the set of links / edges. Each link $(i, j)$ is uniquely defined by the source node $i$ and sink node $j$, and has time-dependent travel time $t_{ij}$ as its cost.

### 3.2 Research framework

The research framework for searching for K-Shortest Paths is illustrated in Figure 3-1. The KSP search starts from the existing dataset and finds the 1st SP. After that, in order to find the next SP that satisfies the overlap and travel time deviation constraints, three main models will be introduced, i.e., network update, link cost penalty, and Kth SP search, respectively.

- **Network update**: the main purpose of network update is, after the 1st SP is found, to update the network in such a way that the next SP to be sought will not be identical to the previously one. The method is to expand the network with hyperlinks as well as hyper nodes that are added to the existing network.

- **Link cost update**: the purpose of the link cost update procedure is to update the cost of links of interest in order to lower the chance of these links being selected in the next SP search, but not to entirely prohibit such occurrence. The idea is to introduce a link penalty and a link generalized cost parameter.
• K\textsuperscript{th} TDSP search: this procedure searches the next SP based on the updated network and link generalized cost. The link generalized cost is used as travel cost of the links during search, which can be different from the actual travel time a vehicle needs to across the links; so the TDSP algorithm needs to store both types of information for different purposes. Also, the route travel time constraints will be built into the TDSP search logic.

In this framework, the proposed KSP algorithm is structured to encompass several major procedures:

a) **Initialization**

Read the network structure and travel time information into the memory; construct the adjacency list for the nodes and links within the network.

b) **Find first SP**
Find the first Shortest Path for the given origin node o and destination node d. It does not matter which SP algorithm is used to find this route, although A star algorithm is recommended as this is a one-to-one SP problem and it runs quickly.

c) **Determine whether all routes have been found**

If all K paths have been found, stop. Otherwise continue to search for the next SP.

d) **Network update**

In this step, the network will be updated based on the shortest paths that have been found so far. As described above, the main purpose of network update is to update the network in such a way that the next SP to be sought for will not be exactly the same as the previous one, which is accomplished through the network expansion and adding hyperlinks as well as hypernodes to the existing network. The detailed logic of the network update procedure will be presented in Section 3.3.

e) **Link cost update**

In this step, the link cost in the expanded network will be updated with a penalty intended to lessen, but not prohibit, the chance of particular links being selected in the next SP, the purpose being to address the overlap issue. The particular links incurring a penalty will be those that were in the previous SP. The level of penalty can affect the overlap degree in the next SP to be found, so it needs to be carefully calibrated. The detailed logic of link cost update will be described in Section 3.4.

f) **Find K^{th} Shortest Path**

After updating the network structure and link cost, find the next Shortest Path with travel time deviation less than the pre-defined threshold. The logic and pseudocode of TDSP incorporating the travel time deviation constraint will be presented in Section 3.5.
If all routes have been found, stop the process. Otherwise go back to step c and continue until all K routes have been found.

3.3 Network update

In this section, given the previously found SP information, the network structure will be updated by expanding the network and adding hyperlinks as well as hyper nodes to the existing network. The purpose of doing this is to update the network in such a way that the next SP found will not be identical to the previous one.

Figure 3-2: Network update procedure
Figure 3-2 describes the procedure for updating the network. The major work on the network update was presented in Martins (Martins 1983). Generally speaking, the network update effort includes the following steps:

I. set i=1;

II. find the first node \( s_1 \) in the \((k-1)^{st}\) Shortest Path with multiple incoming arcs; create a node copy \( s_1' \) for this node;

III. If node copy \( s_1' \) is already in the network, find the first node following \( s_1 \) in the \((k-1)^{st}\) SP that satisfies the following two requirements: 1) it has multiple incoming arcs in the network, and 2) its node copy does not exist in the network. Denote this node as \( s_1 \), then create a node copy \( s_1' \) for this node;

IV. if \( s_i \) equal the destination node \( d \), stop and no more additional paths can be found. Otherwise, add a node copy \( s_i' \) to the network;

V. In the existing network, find all predecessor nodes of \( s_i \). For each predecessor node \( p_j \) \((\forall p_j \in \text{Pred}(s_i))\), add a hyper link \((p_j, s_i')\) into the network, except for the link \((p_j, s_i')\) that had its corresponding physical link \((p_j, s_i)\) included in the \((k-1)^{st}\) Shortest Path.

VI. Along the \((k-1)^{st}\) Shortest Path, find the next node after \( s_i \), and denote it as \( s_{(i+1)} \).

VII. let \( i=i+1 \), go back to step (iv).

3.4 Link cost update

Assume a shortest path has already been found in the previous iteration. When searching for the next shortest path, those links contained in the already found shortest path should be avoided. A penalty function is introduced here to penalize those links, in order to reduce their chance of being selected again in the next SP, so as to address the overlap issue but not prohibit it.

The level of penalty added to the links can affect the overlap degree in the next SP to be found. The greater the penalty added, the less the overlap is likely to occur and vice versa. Section 4.1 presents the changes in the similarity between one path and another in response to imposition of a different penalty.
The link penalty and updated link cost is calculated through the penalty function Eq. (1), (2).

\[
PN(i,j) = t_{ij} \ast \alpha_{ij} \ast w_{ij} \]  

\[
t_{ij}' = t_{ij} + PN(i,j) \]  

In the equation,
- \(PN(i,j)\): Penalty value for link \((i,j)\)
- \(t_{ij}\): original cost for link \((i,j)\), travel time is used here as the cost
- \(\alpha_{ij}\): penalty factor
- \(w_{ij}\): facility type specific weight for each link \((i,j)\).
- \(t_{ij}'\): updated cost for link \((i,j)\)

The penalty factor parameter \(\alpha_{i,j}\) is introduced herein to reflect the degree of the penalty.

In this model, we penalize the links by percentages of the travel time. For example, when \(\alpha = 0.1\) and \(w=1\), 10% of the original travel time will be added as the penalty to the link \((i,j)\).

The parameter weight \(w\) is introduced to represent a different impact of different road types. Generally speaking, a user may feel two routes are different if both routes include different freeways, whereas such a sentiment may be weaker if only arterials are included in the different routes. In other words, freeways links may trigger a stronger distinction among different routes than arterials links. As such, different link categories may be treated differently by applying link specific penalty values.

The procedure to update the link cost is: for each arc \((i,j)\) in the network,

I. Determine whether node \(i\) or node \(j\) is a node copy, if yes, go to step (ii), otherwise go to step (iii).
II. Update the cost of arc \((i,j)\) to be equal to the cost in the true link in the original network. For example, let the cost of arc \((2',5'')\) be equal to the cost of arc \((2,5)\) in the original network;

III. Determine whether arc \((i,j)\) is included in the \((k-1)\)th Shortest Path. If yes, calculate the penalty value by Eq. (1), then update the arc cost according to Eq. (2).

3.5 \(K^{\text{th}}\) TDSP search

A revised Time-Dependent A star algorithm is developed to search the next TDSP based on the updated network and link generalized cost. During the search, the link generalized cost is calculated and used as travel cost of the links, which can be different from the actual time, so the TDSP algorithm needs to store both pieces of information for different purposes. Also, the route travel time constraints will be built into the TDSP search logic to ensure the next route travel time will not deviate too much from the previous ones.

In this revised algorithm, several major changes exist compared with the commonly seen A star TDSP:

Modification 1: Use the destination node copy \(d^{(k)}\) as the destination instead of the real physical node \(d\). By doing this after the network expansion, the TDSP is guaranteed to find a different path.

Modification 2: Each link will keep two cost values at the same time; the first one is the travel time cost denoted as \(tt_{ij}\), the other one is link generalized cost \(gc_{ij}\) which equals the updated link cost \(tt'_{ij}\) from Eq. (2). The reason for keeping two different labels for each link is, TDSP needs to know the actual timestamp of when vehicles arrive at a particular link in order to retrieve the time-dependent link generalized cost, during which process both \(tt_{ij}\) and \(gc_{ij}\) will be used.

Modification 3: The link generalized cost, which equals to the updated link cost \(tt'_{ij}\) from Eq. (2), is used as link cost in the TDSP algorithm, instead of the link travel time cost. I.e., in the
pseudo-code below, it is the generalized cost of traversing link \((i,j)\) at time \(T_igc_{ij}[\tau(T_i)]\), that is used as link cost, instead of the link travel time \(tt_{ij}[\tau(T_i)]\).

**Modification 4:** Each node will keep two different time stamps \(L_i\) and \(T_i\), instead of one as before. While \(L_i\) is still used to record the upper bound of the minimum travel cost (which is generalized cost now) from origin to node \(i\), \(T_i\) is introduced to record the actual travel time cost from origin to node \(i\). The reason to keep \(T_i\) is that, when generalized cost is used to replace travel time as the link cost, \(L_i\) is not able to record the arrival time at node \(i\) any more, but this information is still needed to retrieve the time-dependent traffic network information.

**Modification 5:** We want to guarantee the travel time of \(k^{th}\) Shortest Path \(tt_k\) found will not deviate beyond a certain limit from that of the first Shortest Path \(tt_1\) (assume the threshold is \(x\)), i.e., \(tt_k \leq tt_1 + x\). This constraint is imposed within the stopping criteria: when the node with minimum F label is the destination, check whether this constraint is satisfied, if yes then stop, otherwise pick the node with second minimum F label and start the search again from that node.

The pseudocode of the proposed A star TDSP search algorithm is illustrated below.

<table>
<thead>
<tr>
<th>Step 1) Initialization</th>
</tr>
</thead>
<tbody>
<tr>
<td>(L_0 = t_0, F_0 = e_{od}, T_i = t_0;)</td>
</tr>
<tr>
<td>(L_j = \infty, F_j = \infty, \forall j \neq o, T_i = \infty;)</td>
</tr>
<tr>
<td>(S = {o}; \bar{S} = \emptyset.)</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Step 2) Node selection</th>
</tr>
</thead>
<tbody>
<tr>
<td>(i = \arg\min_{j \in \bar{S}} F_j)</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Step 3) Stopping rule</th>
</tr>
</thead>
</table>

If $i = d$:

- if $(Ti - to) \leq tt + x$, stop;

- else if $(Ti - to) > tt + x$:

  - if $S \neq \emptyset$, $i = \arg\min_{j \in (S - d)} F_j$, continue

    - else stop

- else, continue.

**Step 4) Update $F_j$, travel cost label $L_j$ and time cost label $T_j$**

$S = S \cup \{i\}, S = S \setminus \{i\}$.

For each $j \in A(i)$:

- if $L_i + gc_{ij}[\tau(T_i)] + e_{jd} < F_j$

Then

- $L_j = L_i + gc_{ij}[\tau(T_i)]$

- $T_j = T_i + lc_{ij}[\tau(T_i)]$

- $F_j = L_i + gc_{ij}[\tau(T_i)] + e_{jd}$

If $j \not\in S, S = S \cup \{j\}$, go back to step 2

where:

- $i$ and $j$: index of nodes.
- $S$: The closed set which includes nodes already evaluated.
- $\bar{S}$: The open set which includes the tentative nodes to be evaluated.
- $t_0$: Departure time at origin node o.
- $L_i$: The upper bound of the minimum travel cost from origin to node $i$.
- $e_{ij}$: The estimated lower bound of the minimum travel cost from node $i$ to node $j$. 
- $F_i$: The fake label which is the sum of $L_i$ and $e_{id}$, which means the estimated minimum travel cost from origin to destination going through node $i$.
- $g_{c_{ij}}[\tau(T_i)]$ is the general cost of link $(i,j)$ traversing at time $T_i$.
- $T_i$: Travel time cost from origin to node $i$.
- $\tau(T_i)$: Time index of timestamp $T_i$.

4. CASE STUDY

A case study was conducted to illustrate the performance of the proposed constrained KSP algorithm. We chose the Tucson I-10 network to be our case study region. The network is constructed in DynusT (Dynamic Urban Simulation for Transportation) (Chiu et al. 2011) Dynamic Traffic Assignment Simulation software; the resulting network consists of 395 nodes, 830 links and 80 traffic zones as illustrated in Figure 4-1.

![Figure 4-1: Tucson I-10 network GUI](image)
In this research, the link-based travel time is used as the arc cost in the network. From the DynusT network dataset, information is read into the memory concerning the network topology and the link speed, which represent the network structure and traffic condition respectively. The KSP algorithm is coded in the Java programming language.

4.1 Penalty factor calibration

One feature of the K shortest paths algorithm proposed in this paper is the paths found can share a certain degree of similarity, but not be overlapped too much which might cause the user to fail to notice the difference and to perceive them as the same route, nor unnecessarily to take a detour just to avoid overlap but thus leading to an unreasonable result. This calibration is done by imposing the penalty upon the links in the previous SP. The level of penalty added to the links can affect the overlap degree in the next SP to be found: the more penalty we add, the lower degree of overlap will be expected, and vice versa. According to Eq. (1), one of the parameters influencing the link penalty \( PN(i,j) \) is the penalty factor \( \alpha_{i,j} \), the higher \( \alpha_{i,j} \) is: the higher the penalty that will be imposed upon the links of interest. So the parameter \( \alpha_{i,j} \) itself needs to be carefully calibrated.

In this section, the term similarity is introduced to evaluate the degree of overlap between different paths. When the value of \( \alpha_{i,j} \) changes, the value of similarity which describes the paths’ overlap degree will be updated as well, which can be used for the sensitivity analysis.

4.1.1 Degree of similarity

Assume in the K-Shortest Paths \( \{p_1, p_2...p_k\} \) that have been found, the link sequence for each route \( i \) is: \( p_i=\{l_{i1},l_{i2}...l_{in_i}\} \). For the \( j^{th} \) link in \( p_i \), the repeat times of link \( l_{ij} \) in all other \( k-1 \) paths is calculated and denoted as \( r_{ij} \). The total degree of similarity for route, \( p_i \) is calculated by Eq. (3).
The degree of similarity of those paths can be calculated by Eq. (4).

\[ r = \frac{\sum_{i=1}^{k} r^i}{\sum_{i=1}^{k} n_i} \]  

In these equations:

- \( r^i \): The number of times link \( j \) in path \( i \) show up in all other \( k-1 \) paths.
- \( n_i \): The total number of links in path \( i \).
- \( r^i \): Degree of similarity of path \( i \).
- \( r \): Degree of link similarity of the totality of \( k \) paths.

The range of similarity degree is \( [0,1] \), where 0 meaning the paths are completely non-overlapped and 1 meaning those paths are exactly the same. The higher \( r \) the value is, the higher the degree of overlap observed.

4.1.2 Similarity test

In this similarity test experiment, three KSP algorithms are tested, including the constrained KSP algorithm proposed in this paper and two other classical KSP algorithms, i.e., the Ranking KSP algorithm proposed by Martins (Martins 1983) which laid the groundwork for this research but did not address the overlap and travel time deviation issue, and the Link-Disjoint KSP algorithms which take out every link that was in the previous SP found.

The similarity test is carried out following the procedures below:

1. Set a range for \( \alpha \) value to be tested
2. For each \( \alpha \) value to be tested, calculate the degree of similarity \( \gamma(oda) \) for a pair of nodes (from node \( o \) to node \( d \)) by Eq. (4)

3. Calculate the similarity \( \gamma(oda) \) for every other node pair in the network

4. Compute the degree of similarity of each \( \alpha \) value \( \gamma(\alpha) \), by taking an average of the \( \gamma(oda) \) computed in step 2 & 3

5. Calculate the degree of similarity for every other \( \alpha \) value, and draw the degree of similarity diagram as in Figure 4-2

The changes in degree of similarity due to different values of the penalty factor \( \alpha \) are illustrated in Figure 4-3. The top orange line stands for the Ranking algorithm, the bottom grey line represents the Link Disjoint algorithm, while the blue curve in the middle is the result of the proposed algorithm in this research.

![Figure 4-2: Similarity test](image)

From Figure 4-2, we can see:

I. The similarity degree of our proposed constrained KSP algorithm is able to adjust due to the change of penalty factor \( \alpha \), while the ranking algorithm and disjoint algorithm always keep the same similarity degree. The similarity degree for the ranking algorithm can be as high as 97.6% and disjoint algorithm routes are completely non-overlapped.

II. The similarity degree decreases when \( \alpha \) value increases in the proposed algorithm. This is rather intuitive because the more the links in the previous shortest paths are penalized, the less likely that they are going to be chosen in
the next shortest paths. Thus, the degree of similarity between K-Shortest Paths will decrease.

III. When \( \alpha \) equals 0, the similarity degree of the proposed model is equal to the Ranking algorithm as no links will be penalized, which means the routes found are highly overlapped. When \( \alpha \) is relatively large, the similarity of the proposed model will be close to the result of the Edge-Disjoint algorithm, which means the paths will be non-overlapped.

Observed from Figure 4-2, when the penalty factor \( \alpha_{i,j} \) takes the value between 0.3 and 0.8, the degree of similarity between those paths found is around 25-35%. If we assume 25-35 percent overlap is a reasonable range, then \( \alpha \) value between 0.3-0.8 is recommended. In the case study experiments in the following sections, the penalty factor will take the value of 0.6 unless specially noted.

4.2 Case study

In this section, numeric analysis is presented to show the difference between the algorithm proposed in this research and the ranking algorithm and link-disjoint algorithm as described in section 4.1. The parameter values used are as follows: \( K=3, \alpha_{i,j} = 0.6, w_{i,j}=1 \) for both arterials and freeways. Maximum travel time deviation allowed is 5 minutes or 20% of the travel time of the first route, whichever is larger.

Two examples will be shown to illustrate the performance of the proposed algorithm.

4.2.1 Case study 1

Node 1414 at the top left corner is selected as the origin, and another node 2290 located at the middle right is chosen as the destination in this numeric analysis. Three shortest paths for each algorithm adopted in this research are found and shown in the following figures. Figure 4-3, Figure 4-4 and Figure 4-5 represent the paths for the ranking algorithm, link-disjoint algorithm and
proposed algorithm respectively. The red route represents the 1st shortest path, the orange color stands for the path with 2nd shortest travel time, and the yellow color route is the 3rd shortest path.

Figure 4-3: Shortest paths by ranking algorithm – case study 1

Figure 4-3 shows the shortest paths using the ranking KSP algorithm. We can see the three routes are highly overlapped, with the only minor difference that the orange route makes a U-turn on a road before entering the freeway, and the yellow route makes a U-turn after leaving the freeway. Except for that, the remainder of these routes are identical. The travel times for the three routes are: 9.1, 9.3, and 9.4 minutes respectively. The excessive overlap degree will make the three routes look like one, and would not make much difference to the users.
Figure 4-4: Shortest paths using link disjoint algorithm – case study 1

Figure 4-4 shows the shortest paths result using the link-disjoint algorithm. We can see that none of the links in the three routes is overlapped. It should be noticed the 2nd route (orange color) is going through the top left corner just to avoid the links already taken by the 1st route, and the 3rd route (yellow color) is going through the bottom right corner also to avoid the links already in the 1st route. This does not make much sense because usually drivers will not travel along the opposite direction which requires longer travel distance and higher travel time.

Also, the travel times for the three routes are: 9.1, 17.3, and 17.6 minutes respectively. We can tell that the travel time on the 2nd and 3rd routes are almost double that of the 1st route, and under
most circumstances users will not consider traveling on those routes as good options due to the large travel time difference, which makes the 2\textsuperscript{nd} and 3\textsuperscript{rd} routes of limited utility to the users and, thus, they fail to meet the objective of providing good route options.

Figure 4-5: Shortest paths using proposed algorithm—case study 1

The three routes shown in Figure 4-5 appear to be more promising; certain parts of the routes overlap while some do not. None of the routes found overlap too much to be considered identical, nor do they unnecessarily detour into the opposite direction. Also, the travel times for these three routes are 9.1, 12.1, and 12.6 minutes respectively, which deviate a little bit but are within the user acceptable range. It is very intuitive to know that these are good alternative routes
that make sense to the drivers, and are much better than the routes computed by the other two algorithms.

4.2.2 Case study 2

A new node pair is chosen for the other case study example in this section. The origin node is 2242 located at the top middle of the network, and destination node is 2227 located at the bottom right corner of the network. Three shortest paths for each algorithm adopted in this research are found and shown in the following figures. Figure 4-6, Figure 4-7 and Figure 4-8 represent respectively the paths for the ranking algorithm, link-disjoint algorithm and proposed algorithm. The color codes are the same as in the previous section.

Figure 4-6 shows the shortest paths using the ranking KSP algorithm. We can see the three routes are highly overlapped, the only difference with the orange route is it makes a U-turn on an arterial at the very beginning, and the yellow route contains a loop immediately before entering the freeway. Except for these minor features, the remainder these routes are identical to each other. The travel times for the three routes are: 8.8, 8.9 and 8.9 minutes respectively. The highly overlapped three routes fail to provide different travel options to the users.
Figure 4-6 Shortest paths by ranking algorithm – case study 2

Figure 4-7 shows the shortest paths result using the link-disjoint algorithm. We can see that none of the links in the three routes is overlapped. The red and orange paths look fine, but the 3rd route (yellow color) contains quite a few zigzags just to completely avoid the links already in the first two routes. The travel times for the three routes are: 8.8, 13.5 and 18.2 minutes respectively. We can tell the travel time on the 3rd route is more than double that on the 1st route. Under most circumstances, users will not consider traveling on this route as a good option due to the huge traffic
time difference, which makes it of no help to the users and it fails to meet the objective of providing good route options.

Figure 4-7 Shortest paths using link disjoint algorithm – case study 2

The three routes shown in Figure 4-8 are results from the proposed algorithm and make more sense; certain parts of the routes overlap while some do not. The 1\textsuperscript{st} and 2\textsuperscript{nd} paths go through freeway and the 3\textsuperscript{rd} path is completely on the arterial. None of the routes found overlap so much as
to be identical, nor do they unnecessarily detour into the opposite direction. Also, the travel times for these three routes are: 8.8, 9.7, and 12.0 minutes respectively, which deviate a little bit but are within the user acceptable range.

Figure 4-8 Shortest paths using proposed algorithm– case study 2

Table 4-1 summarizes the comparison among the different shortest paths from all three algorithms. The number of asterisks given stands for the magnitude of value with respect to that
criterion. For example, the Rank Algorithm leads to the largest overlap and least travel time deviation, whereas the Disjoint Algorithm leads to the minimal route overlap and maximal travel time deviation. It is apparent that the proposed algorithm is able to find the routes with both moderate overlap degree and acceptable travel time deviation, which demonstrates its capability to provide drivers with multiple meaningful and useful travel options.

Table 4-1 Comparison among KSP algorithms

<table>
<thead>
<tr>
<th>Algorithm</th>
<th>Overlap</th>
<th>Travel time deviation</th>
</tr>
</thead>
<tbody>
<tr>
<td>Rank Algorithm</td>
<td>★★★★★</td>
<td>★</td>
</tr>
<tr>
<td>Disjoint Algorithm</td>
<td>★</td>
<td>★★★★★</td>
</tr>
<tr>
<td>Proposed Algorithm</td>
<td>★★☆☆</td>
<td>★★☆☆</td>
</tr>
</tbody>
</table>

4.3 Sensitivity analysis

In this section, the sensitivity analysis of the proposed algorithm is carried out from two aspects: (1) different K Shortest Paths results due to user preference with respect to different road facility types, and (2) different K Shortest Paths results due to traffic condition improvement along certain corridors.

The same OD pair from section 4.2.1 is used in this sensitive analysis. The parameters used in that case study example remain unchanged unless specially noted.

4.3.1 User preference test

As described in Section 3.4, a difference in the penalty added to links can affect the degree of overlap in the newly found SP; the higher the penalty we add, the less is the overlap and vice
versa. Besides the general penalty factor $\alpha$, the parameter weight $w$ is introduced to represent the different user perception of different road types. Generally speaking, it is assumed that a user feels two routes are more distinct if both include two different freeways, while he/she will feel less difference if he/she is being guided to take different secondary roads and use different ramps to the same freeway.

In this section, combinations of $w_{i,j}$ values for freeway and arterials will be tested and their results will be analyzed.

Combination 1: $w_{i,j}$ takes an equal value for freeway and arterial. The two link categories are penalized the same amount, and users do not have any preference for either one of them over the other.

Combination 2: $w_{i,j}=2$ for freeway and 1 for arterial. The freeway links are penalized twice as much as arterial links, and users show a preference to avoid using the same freeway.

Combination 3: $w_{i,j}=10$ for freeway and 0.1 for arterial. The freeway links are penalized ten times as much as arterial links, which indicates users have a very strong preference to avoid freeway and stay on the arterial.

Figure 4-5 already describes the proposed algorithm result for the first combination: we can see route 1 and route 3 share the same part of the freeway, but route 3 deviates from the freeway and most links on its path are arterial links.

Figure 4-9 below shows the proposed algorithm’s three SP results with weight combination 2. It can be observed that route 1 and route 2 are identical with route 1 and route 2 in Figure 4-5. The only difference is – since users have stronger preference to avoid using the same freeway and a higher penalty has been imposed upon the I-10 freeway, – the first part of the yellow route is not
traveling on the freeway anymore but goes via an arterial that is parallel to I-10. The rest of the yellow route is still the same as before.

Figure 4-9: Sensitivity analysis – combination 2

When the value of $w_{i,j}$ becomes very high for freeway and low for arterial, users have very strong preference to avoid the same freeway but low aversion for the repeated arterial links. Figure 4-10 shows the KSP result for combination 3. The 3rd route will travel all the way along the arterial parallel to the freeway and the last road segment overlaps with the other two routes.
Figure 4-10: Sensitivity analysis – combination 3

The sensitivity analysis through the 3 combinations of $w_{i,j}$ on freeway and arterial shows that the proposed algorithm is able to find expected different KSP results for the different penalty parameter combinations, whenever users do not have significant preference for freeway or arterial, or significantly prefer freeway over arterials. This sensitivity result demonstrates the capability of the proposed algorithm to adjust the outcomes according to the predefined user preference, and its flexibility to be applied under different circumstances.

4.3.2 Traffic improvement test
In this part, the sensitivity of the shortest paths to traffic condition changes will be tested. It is assumed that, due to the recently finished construction, the road segments along the following two corridors increase from the previous 2 lanes per direction to be 3 lanes per direction, and the traffic condition along the two arterial corridors has been significantly improved.

These two corridors are:

1) Corridor A: West Ina Road from North Silverbell Road to North 1st Avenue
2) Corridor B: North Oracle Road, from West Ina Road to East River Road.

The two corridors are marked in thick green color in Figure 4-11, which also shows the Shortest Paths result after the corridor traffic condition improvement.

Figure 4-11: Updated Shortest Paths result along two improved corridors
It can be observed from Figure 4-11 that, after the traffic condition has been improved along the two corridors, the 2nd shortest path is traveling through both traffic corridors as expected.

Comparing the above result in Figure 4-11 and the previous result from Figure 4-5, we can find the 1st route is identical and still going through the I-10 freeway. However, the previous 2nd path becomes the 3rd route in the current solution, and the previous 3rd SP deviates to the improved arterial corridors and its new total travel cost turns out to be superior than the previous 2nd route.

This experiment shows that, after the traffic condition has been improved along the two corridors, the proposed constrained KSP algorithm is able intelligently to detect the updated traffic condition and include these two light traffic corridors in the updated shortest path solution. It can also be observed that in the new result, the 3 routes are slightly overlapped which is considered reasonable. Further, the travel times of those three new routes are 9.1, 11.6 and 12.1 min respectively, and the travel time deviation between these paths is acceptable. We can tell the 3 new paths found are sound travel options in time as well as in routing.

5. CONCLUDING REMARKS

This paper documents the research effort in developing a unique constrained K Shortest Paths Algorithm to find K Shortest Paths between two given nodes. The K Shortest Paths Algorithm proposed in this research is capable of finding K Shortest Paths without path enumeration, and allows paths to have reasonable overlap while still being perceived as different routes. Also, the KSP problem to be solved is constrained by the total travel time. The goal of this research is to provide sound route choice options to the drivers in order to assist their route choice decision process, during which process the overlap and travel time deviation issues between the K paths need to be considered.
The proposed algorithm incorporating the above-mentioned overlap and travel time deviation constraints is developed and coded in this research, a numerical analysis is conducted on the Tucson I-10 network using DynusT DTA simulation software, and the similarity is calculated to show the degree of overlap between the paths found. The outcome of the case study shows that our proposed algorithm is able to find the different shortest paths with reasonable similarity degree and close travel time, which indicates the result of the proposed algorithm is satisfactory.
6. REFERENCES


