June 2017

Space-Time Transportation System Modelling: from Traveler's Characteristics to the Network Design Problem

Mohsen Parsafard
University of South Florida, m.parsafard@gmail.com

Space-Time Transportation System Modelling: from Traveler's Characteristics to the Network

Design Problem

by

Mohsen Parsafard

A dissertation submitted in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy in Civil Engineering
Department of Civil and Environmental Engineering
College of Engineering
University of South Florida

Major Professor: Xiaopeng Li, Ph.D.
Fred Mannering, Ph.D.
Yu Zhang, Ph.D.
Changhyun Kwon, Ph.D.
Guangqing Chi, Ph.D.

Date of Approval:
May 25, 2017

Keywords: Network Design, Mobility Pattern, Time Geography, Traffic Surveillance, Space-Time Network

DEDICATION

To my lovely parents and my beautiful wife for their endless support, inspiration, and encouragement.

ACKNOWLEDGMENTS

First and foremost, I would like to express my deepest gratitude towards my major advisor Dr. Xiaopeng Li, for all his guidance and tremendous support during my Ph.D. studies. I would also like to thank Dr. Mannering, Dr. Zhang, Dr. Kwon and Dr. Chi for graciously serving on my doctoral committee and their invaluable advice and constructive comments that helped me to improve my dissertation.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

ABSTRACT

Traditional network design problems only consider the long-term stationary travel patterns (e.g., fixed OD demand) and short-term variations of human mobility are ignored. This study aims to integrate human mobility characteristics and travel patterns into network design problems using a space-time network structure. Emerging technologies such as location-based social network platforms provide a unique opportunity for understanding human mobility patterns that can lead to advanced modeling techniques. To reach our goal, at first multimodal network design problems are investigated by considering safety and flow interactions between different modes of transport. We develop a network reconstruction method to expand a single-modal transportation network to a multi-modal network where flow interactions between different modes can be quantified. Then, in our second task, we investigate the trajectory of moving objects to see how they can reveal detailed information about human travel characteristics and presence probability with high-resolution detail. A time geography-based methodology is proposed to not only estimate an individual's space-time trajectory based on his/her limited space-time sample points but also to quantify the accuracy of this estimation in a robust manner. A series of measures including activity bandwidth and normalized activity bandwidth are proposed to quantify the accuracy of trajectory estimation, and cutoff points are suggested for screening data records for mobility analysis.

Finally, a space-time network-based modeling framework is proposed to integrate human mobility into network design problems. We construct a probabilistic network structure to quantify human's presence probability at different locations and time. Then, a Mixed Integer Nonlinear

Programming (MINLP) model is proposed to maximize the spatial and temporal coverage of individual targets. To achieve near optimal solutions for large-scale problems, greedy heuristic, Lagrangian relaxation and simulated annealing algorithms are implemented to solve the problem. The proposed algorithms are implemented on hypothetical and real world numerical examples to demonstrate the performance and effectiveness of the methodology on different network sizes and promising results have been obtained.

CHAPTER 1: INTRODUCTION

Transportation engineers and city planners need to thoroughly understand human travel characteristics and mobility patterns, to economically locate network facilities and better serve travelers. Considering different types of facilities and infrastructures in transportation systems such as highways, intersections, bridges, sensors and digital traffic signs, etc., advanced modeling techniques are required for making the best decisions. For instance, smartly deploying sensors and digital traffic signs that can interact with travelers enables us to best observe travel characteristics and transfer route guidance and other information to the network users. Figure 1 illustrates different components of transportation systems and how closely they are tied to each other, which can be addressed with related modeling techniques.
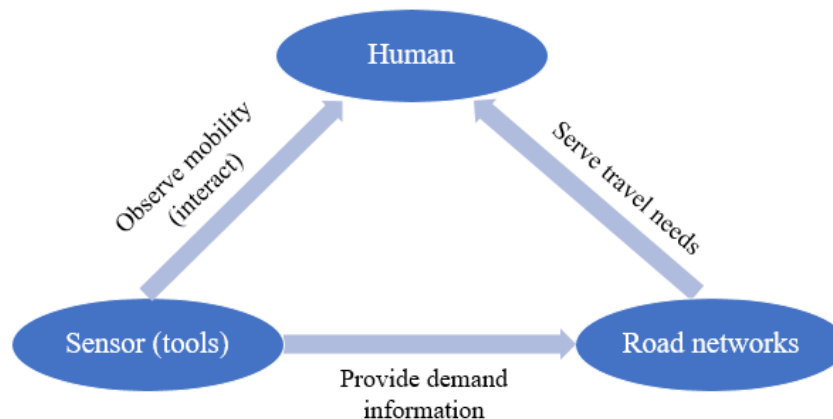
Figure 1 Transportation system components.

Network design problems have been intensively studied for several decades, and researchers have conducted numerous studies investigating theoretical models. The early studies have focused on a deterministic location design problems and have developed models on discrete

[1], [2] and continuous setting [3], [4] formulations. Later, researchers started to investigate the stochastic facility network design to enhance system availability by providing redundancy [5]–[7]. There are many studies on location design problem that address only a single mode, and the literature of multimodal network design problem is relatively limited [8]. Further, in most relevant studies, the design objective is primarily related to travel time, while other important factors e.g. safety and spatial conflicts between different modes are yet to be quantitatively addressed.

In many small communities in the United States that transportation is dominated by autos (single mode network) the lack of walkway infrastructures like sidewalks and crosswalks poses a safety hazard to pedestrians since they must walk along busy streets and highways. Since 1920, there has been a growing concern about pedestrians' safety [9]. Pedestrians' safety is defined as the condition of being protected from danger, risk, or injury caused by accidents with motor vehicles. The Federal Highway Administration (FHWA) estimates that 4,500 pedestrians are killed annually because of traffic accidents with motor vehicles, and as many as 88% of those accidents could have been avoided if walkways separate from travel lanes (sidewalks or crosswalks) had been available to pedestrians [10]. This implies the importance of pedestrians' safety in transportation network design for city planners and governments. Due to the limited budget, identifying the optimal locations for separate walkways is yet a challenging issue.

Motivated by these gaps, the first aim of this research is to investigate a new framework for multimodal network design problems, by considering safety and flow interactions between different modes of transport. A mixed-integer nonlinear programming model is proposed for locating sidewalks and crosswalks in a transportation network to reduce the overall cost, improve pedestrians' safety and enhance walkability.

Traditional modeling frameworks in network design problems only consider the long-term stationary travel patterns (e.g., fixed OD demand) and short-term variations of human mobility are ignored. Recently, the rapid developments of geo-tagged data provide an opportunity to improves our knowledge about people's mobility and travel characteristics, and it can lead to a better approach towards future modeling efforts.

In the second task of this research, we investigate space-time travel patterns of humans at the individual trajectory level with high-resolution detail. Such geo-tagged mobility data are based on objective measurements or samples of human travel paths and contain massive temporal, spatial, and semantic information about individuals. They have become increasingly available and have provided substantial opportunities for understanding human mobility patterns [11]–[15], travel behaviors, and lifestyles [16], [17] and safety analysis [18]. Such data are critical to planning and operations of an efficient and reliable transportation system and eventually to the economic prosperity and long-term sustainability of an urban system [19]. These data can be used to characterize traffic flow patterns for better management of road networks [20]. Further, they have also been used to predict real-time travel demand for more responsive and convenient mobility operations [21].

While geo-tagged mobility data provide abundant information for human travel characteristics, they are in general limited in two aspects, i.e., representativeness and granularity. First, representativeness refers to whether the individuals captured in the data well represent the overall traveler population in the corresponding region. Geo-tagged mobility data often have certain biases on particular traveler groups. For example, Twitter data may over-represent young age groups while under-representing senior age groups [22]. Granularity refers to whether the sample data are dense or frequent enough for an accurate estimation of the corresponding space-

time trajectories. Geo-tagged mobility data essentially contain discrete samples of continuous space-time human (or vehicle) travel trajectories. Many mobility data sets have been recorded with very low sampling rates, which can be even further screened due to privacy concerns [23]. Other than the available sample locations, the remaining portions of trajectories are subject to interpolation-based estimation. It is easy to imagine that sparse samples likely yield higher estimation inaccuracies.

Uncertainties due to low granularities of geo-tagged mobility data have been mainly investigated with two types of approaches. The first type focuses on deterministic geometric or geographical bounds to an individual's activities based on known space-time sample points. Trajcevski et al. (2004) model trajectory bounds as cylinders to facilitate trajectory database queries. The time geography theory [24], [25] uses a space-time prism to bound an object's activity range between two consecutive space-time samples as a prism, i.e., the intersection of two cones oriented in opposite directions. This concept is further generalized to incorporate transportation network structures [26]. The second type assumes stochastic underlying patterns of individual movements. With such stochastic settings, developments on probability distributions [27] and stochastic processes [28] can be applied to describe uncertainties of trajectories estimated with geo-tagged sample points.

Despite these developments in modeling trajectory uncertainties, there is a lack of simple and efficient measures on the quality of using spatiotemporally distributed discrete samples in estimating an individual's continuous trajectory and the suitability of such sample data for studying this individual's travel patterns. Without such measures, transportation planner and operators may have difficulty in identifying whether a particular geo-tagged data set can help them accurately

quantify human travel patterns or not. They may also not be able to identify useful data sets from a vast amount of geo-tagged mobility data while such data become increasingly available.

To bridge these gaps, we propose a set of quantitative measures for the inaccuracies of trajectory estimations with discrete geo-tagged mobility data. We apply a spatiotemporal data model based on time geography for representation and computation of geo-tagged mobility data. Based on this model, we propose two inaccuracy measures to quantify the accuracy of trajectory estimation in a robust manner. We also suggest cutoff points for screening data records for mobility analysis. These measures only have one parameter as an individual's maximum speed and are applicable to different types of geo-tagged mobility data. To alleviate computational load for large-scale data sets, we develop an efficient interpolation method with a lookup table. To demonstrate the applications of these measures, we test multiple sets of real-world geo-tagged trajectory data, including cell phone records and geo-tagged Twitter data. We find that the proposed measures can efficiently estimate associated mobility estimation inaccuracies for a large amount of individual mobility sample data. Further, these results also reveal managerial results into the quality of these data for human mobility studies, including their distribution patterns. Overall, the outcomes of the second task of this study advance our knowledge in understanding the relationship between the spatiotemporal distributions of geo-tagged mobility data and the quality of associated trajectory estimations. They provide a parsimonious and robust tool for evaluating the quality of massive amounts of geo-tagged mobility data and screening useful information from such data for mobility studies.

The final aim of this research is to integrate human mobility characteristics into network design problems. Nowadays, governments, industries and marketing companies want to track their target population, collect their essential data and better interact with them to increase their total

welfare. For instance, in advertisement industry, billboards are the most common outdoor advertising strategy in cities and they have dominant advantages compared to the traditional advertising media such as prominent visibility, low cost per mile, and superior accumulation of local influence. To launch a successful billboard, the most important factors that should be cautiously considered is the geographical locations of billboards to increase the target exposure [29]. In transportation, traffic management centers deploy traffic sensors at various places across a network to estimate and predict network-wide traffic conditions. Electronic digital signs are also located in different spots to interact with travelers and transfer route guidance and information about the network.

Due to limited resources and investment constraints, there is a need for smartly finding the proper locations of facilities over the network to maximize the network coverage. Regardless of the type of facility (e.g. sensors, billboards, etc.), in deterministic location design problems, the short-term variations and dynamics of targets' mobility are ignored. In particular, dynamic space-time relationships between facilities and sensing targets are seldom investigated in existing studies.

To bridge these gaps, we try to integrate human mobility into network design problems by constructing a stochastic framework, where the presence probability of targets could be quantified in a space-time network-based structure. In general, we can expand a physical network by copying it in different time stamps and construct a space-time network. The advantages of the space-time network are that we can work with space-time nodes and links and specify different attributes (e.g., presence probability) of individual targets to them. Using the proposed space-time framework, a Mixed Integer Nonlinear Programming (MINLP) model is developed to maximize the utility from the inspected targets by optimally locating coverage facilities. To achieve near optimal solutions, customized algorithms including greedy heuristic, Lagrangian Relaxation (LR) and Simulated

Annealing (SA) are derived. Numerical examples on hypothetical and real world networks show that the proposed algorithms can efficiently solve realistic problem instances

The organization of this dissertation is as follows: CHAPTER 2 reviews relevant literature and explains why this study contributes to the existing literature. CHAPTER 3 presents the facility location design problem and proposes the network reconstruction process where a multimodal network is reconstructed from a single mode transportation network. The application of this network reconstruction method to a traffic calming network design problem will also be investigated. CHAPTER 4 is about mobility pattern and trajectory analysis of moving objects. Further, we propose a new methodology for quantifying the activity range of moving targets using spatiotemporal data, and more importantly, to quantify the accuracy of trajectory estimation in a robust manner. CHAPTER 5 proposes a probabilistic space-time network structure that enables us to integrate human mobility into network design problems and maximize the network coverage. Finally, CHAPTER 6 concludes this dissertation and discusses possible directions for future research.

CHAPTER 2: LITERATURE REVIEW

2.1. Time Geography, Mobility Pattern and Trajectory Analysis

In transportation studies, it is important to understand human travel characteristics from both microscopic and macroscopic perspectives. At the microscopic scale, we need to understand how vehicles proceed and interact (e.g., via car-following laws [30]) on a road link. Understanding microscopic vehicle trajectory characteristics help us explain root causes to a notorious highway traffic phenomenon, traffic oscillation, which incurs a number of adverse impacts on highway traffic efficiency and sustainability [31]–[33]. At the macroscopic scale, people's space-time travel patterns across a wide space region are of a concern. Studies on human mobility and activity patterns have been applied to several fields, including epidemic modeling [34], traffic prediction [35], urban planning [36], [37], and social networks [38], [39]. With the emergence of rapidly growing geo-tagged Big data from various sources [11], [12], tremendous efforts have been made in the attempt to understand human mobility and activity patterns over time and space.

To name just a few seminal studies, Brockmann et al. (2006) discovered a power-law distribution of human travel distances from anonymous one-dollar bill transactions. González et al. (2008) tracked traces from over ten thousand mobile phone users for a six-month period to quantify the scaling laws of individual humans. With similar cell phone data, Song, Qu, Blumm, & Barabási (2010) further showed that individual human trajectories have a high degree of predictability, although some of their collective measures demonstrate distribution patterns akin to those of scale-free random walks [40], [42].

8

Of the relevant data sources, social media (e.g., Twitter, Facebook) is arguably the newest and the most rapidly growing data source and has drawn enormous interest in many research fields, such as computer science [35], sociology [43], and urban and transportation planning [36], [37], [44], [45]. Social media data are possibly a low-cost, high-information supplement to conventional travel survey methods and contain detailed individual information from semantic messages. In particular, emerging location-based social networks (LBSNs) are a popular form of social media that provide accurate individual location information in addition to semantics [46]. Check-in records on LBSNs contain rich social and geographical information and provide a unique opportunity for researchers to study users' spatial-temporal social behavior [44], [47]–[49]. However, such data have several limitations in determining an individual's activity chain, including concerns about user privacy, lack of detailed descriptions of the activities, missing activities, and a deficiency in individual socioeconomic characteristics [49]–[51].

While studies on theoretical path modeling are mature (e.g., [52], [53]), tracing and predicting an individual's activity patterns using geo-tagged mobility data are still in the exploratory stage. This study attempts to address the issue of missing activities, that is, an individual's whereabouts are known only at sample time points, but activities at other times are missing from the data.

The time geography theory can be used to estimate the range of an individual's missing whereabouts based on his/her known locations posted on social media. Time geography reveals how participating in an activity at a given place and time is directly related to abilities to participate in activities at other places and times [54]. This concept has been recently applied to transportation network design in innovative ways [55], [56]. Recent developments in time geography can provide intuitive concepts and quantitative measures to describe how discrete sample points can confine

an object's path in a space-time coordinate system. Inspired by the general relativity, time geography quantifies an object's activity range given its mobility capability and sometimes geographical barriers as well [24], [25], [57].

According to the time geography literature, three major factors can constrain an individual's ability to conduct activities in space and time: capability constraints (physiological necessities, such as sleeping), authority constraints (limited access, such as a military area), and coupling constraints (spatial and temporal requirements, such as a meeting at 3 p.m.) [58], [59]. Considering these factors, researchers applied time geography to investigate human activity patterns by integrating time geography concepts with geographic information systems (GIS) [60], [61], three-dimensional geo-visualization of activity-travel patterns [62], and some analytical measurement, such as the space-time path, space-time prism, or station [25].

The space-time path and space-time prism are two fundamental concepts in time geography literature [24], [25], [63]–[65], and they serve as the basis for the proposed measures in this study. The space-time path has been applied to mobile phone logs to study human movement behavior [65] and individual access to urban opportunities [66]. The space-time prism is a more powerful measure for assessing the ability of an individual to travel and participate in activities and is used for measuring accessibility [25], [67], [68].

Recently, the probabilistic models in time geography and spatial databases have been investigated from different perspectives. For instance, studies on measurement error analysis in measurement-based GIS argue the spatial data quality [69] and error propagation [70]. Researchers have also developed mathematical foundations for modeling the distribution of visit probabilities within the space-time prisms using the Random Walk theory [71], [72], the truncated Brownian Bridges method [73] and the moment-design method [74].

These studies develop various means of describing human activity patterns under various probabilistic modeling assumptions. Different from these modeling assumptions, this study only employs the very simple concept of time geography to evaluate to what extent such data can reflect an individual's activity trajectory. Note that because of irregular geometries, finding space-time prisms can be challenging for Big Data with available commercial software. Our proposed measures can smartly circumvent this computational challenge with a look-up table method.

2.2. Location Design Problems

Facility network design problems are one of the most difficult and challenging problems in transportation. The majority of studies in transportation network design problems (TNDP) focus on developing mathematical formulations and solution techniques for improving the utilization of the transportation network through either link improvements (i.e. expanding the capacity) or link additions (e.g. building new streets) [8]. The common objective in TNDP is to make an optimal investment decision in order to minimize the total travel cost in the network. The TNDP is usually formulated as a bilevel, leader–follower problem [75]–[78]. The upper level is the designers' problem in which decision makers (e.g. city planners) design the transportation network. The lower level problem is the travelers' problem in which users decide on their travel route and mode of transport. The bilevel structure allows the decision maker to improve the transportation network while accounting for travelers' route and mode decisions [8].

Although there have been many studies of the TNDP, there are still numerous gaps and limitations in the literature, especially for multimodal network design problems. For instance, many transportation network design problems address only a single mode, and the literature of multimodal network design problem is very limited [8]. We found only one study in multimodal TNDP that considers a non-motorized mode (bicycle) [79]; all other studies focus on either bus

and car or on bus, car, and metro [80]–[83]. In pedestrian transportation literature, most studies are descriptive [84]–[86]. To the best of our knowledge, no study has considered walking as a mode of transportation within a multimodal network design problem. Also, most of the existing studies in multimodal network design problems do not consider flow interaction between different transportation modes [87]–[89]. However, when transportation modes share lanes in reality, the flow of different modes do interact.

Most studies in TNDP have ignored combined mode trips where travelers can use multiple modes of transportation during their trips such as park-and-ride, especially in the strategic level decisions[8]. An important aspect of multimodal transportation systems with combined mode trips is to provide convenient mode transfer possibilities for travelers. Nowadays, with the advent of technologies like Uber, Lyft, SideCar and Curb, combined mode trips seem more viable than ever before. In addition, the objective in most studies in TNDP is primarily related to travel time [78], [90]–[92], or travel cost such as operator cost, and user cost [80], [89]. These gaps are addressed in the second task of this study, where we develop a multimodal TNDP for deployment of pedestrian infrastructures to reduce the overall cost, improve pedestrians' safety and enhance walkability.

Recently, a number of reliable location models have been proposed to address possible probabilistic facility failure risks [93]–[95]. Snyder and Daskin (2005) proposed two reliable facility location model formulations (based on p-median and UFL models) to investigate the effect of probabilistic facility failures on the optimal facility deployment. Li and Ouyang (2010) addressed the correlations between facility failures in the context of reliable facility location design, but the developed continuum approximation model is only suitable for macroscopic problems with a smooth and continuous setting. These discrete and continuous reliable facility

12

location modeling techniques have been adapted to solve traffic surveillance sensor location design problems [98], [99].

Traditionally, the operation of transportation systems has been monitored by fixed sensors that are not movable after installation [100]–[103] to gather valuable information e.g. link flow data. Most of the studies in the literature typically assume the availability of measured link traffic information on all network links. However, in practical applications, the assumption of installed sensors on all links is unrealistic due to budgetary constraints. The existing models in sensor location design problem can be identified into two categories: models solving the sensor location flow-observability problem and models solving the sensor location flow estimation problem [104]. Studies on flow-observability problem determine the optimum deployment of sensors such that flows involved in the system are known [100], [105]–[107]. On the other hand, flow-estimation problems (e.g. origin-destination flow estimation) try to optimally deploy sensors such that the derived flow estimates are the best possible [108]–[111].

Overall, the optimal design of sensor locations across a transportation network yields the best data collection. Different sensor location models exist on different criteria such as sensor types (e.g., counting sensors, image sensors, Automatic Vehicle Identification (AVI) readers), available a priori information and flows of interest (e.g., OD flows, route flows, link flows) [112]–[118]. In addition to the link flow estimation, there is a particular emphasis on the end-to-end travel time prediction problem. Xing et al. (2013), proposed measurement and uncertainty quantification models to explicitly take into account several important sources of errors in the travel time estimation/prediction process, such as the uncertainty associated with prior travel time estimates, measurement errors and sampling errors. In their study, they selected a path travel time uncertainty

criterion to construct a joint sensor location and travel time estimation/prediction framework with a unified modeling of both recurring and non-recurring traffic conditions.

During the last decade, the optimal allocation of plate scanning devices has received very limited attention, even though several scholars have demonstrated that plate scanning devices are much more informative than those based on traditional link count information. Mínguez and Sánchez-Cambronero (2010), provided some techniques for obtaining the optimal number and location of plate scanning devices for a given prior OD distribution pattern under different situations, i.e. maximum route identifiability or budget constraints. Cerrone et al. (2015) proposed four mathematical formulations for vehicle-ID sensors location problem to derive route flow volumes. They have considered both the full observability and estimation version of the problem.

Despite the wide use of fixed sensors in transportation networks, there are evidences that mobile traffic sensors can improve the transportation network surveillance in different ways [121]. New technologies such as Unmanned Aerial Vehicles (UAVs) are more flexible to monitor the real time traffic conditions and special events with regards to both space and time and they can be considered as dynamic sensors [122], [123]. These technologies help the decision makers to change the sensor network design in unexpected events such as network disruptions. Zhang et al. (2015) proposed a mathematical model for traffic state detection of spatially and temporally distributed incidents using UAVs. This study shows that using a number of illustrative and real-world networks, their proposed model offers a unified fixed and dynamic sensor network framework and efficient routing/scheduling algorithms for improving road network observability. Zhu et al. (2014) investigated the motion ability of traffic sensors and proposed a mobile traffic sensor routing problem. The results showed that the mobile traffic sensor had a better network surveillance performance than the fixed sensor.

Despite the considerable number of studies on formulating and solving the sensor location problem, little research in this context has considered short-term variations of human mobility. Nowadays, location-based social data that contain a massive temporal-spatial information about population targets can be integrated into location design problems to improve the network coverage. Considering this as a unique opportunity that is not investigated in the literature, we to develop a new space-time framework and formulate a mathematical programming model to maximize the network coverage. This could be extended to other application e.g. advertising industry. For instance, launching successful billboards large cities to increase the target exposure [29].

CHAPTER 3: NETWORK DESIGN PROBLEM - A DETERMINISTIC APPROACH

3.1. Introduction[1]

Transportation Network Design Problems (TNDP) are concerned with building new transportation infrastructures (e.g. streets or bridges) or expanding the capacity of existing infrastructures. City planners need to consider accessibility and/or mobility related issues such as congestion, air pollution, traffic incidents, etc., while designing a transportation network. TNDP are necessary to improve the systems performance especially in metropolitan areas and city centers. TNDP have been studied during the last five decades with a concentration on the road network and public transit. The primary focus of TNDP is to satisfy traveler's needs, increase relevant social welfare and serve the travelers with the minimum total system cost by best planning and configurations of certain infrastructure facilities.

As discussed in CHAPTER 1, many studies on transportation facility location design address only a single mode, and the literature of multimodal network design problem is relatively limited (see [8] for more details). In this chapter, we investigate a pedestrian network design problem in a multimodal network, where the behavior and characteristics of travelers are fully observed. We develop a network reconstruction method to expand a single-modal transportation network to one suitable for multi-modal traffic analysis considering flow interactions and safety. This reconstruction method further leads to a multimodal transportation network design model to enhance not only network efficiency but also pedestrians' safety by optimally designing traffic

---

[1] Portions of this chapter is published in [125],[126]. Permission is included in Appendix D.

calming facilities. In particular, flexible changes of transportation modes along a traveler's trajectory and interactions of different modes of transportation when they share certain facilities are considered.

We develop a bi-level mathematical programming model for optimally locating sidewalks and crosswalks (S&C) in a transportation network such that the overall transportation cost decreases and pedestrians' safety improves. In the proposed bi-level leader–follower optimization model, the leader (city planners) decides where to locate the traffic calming facilities in the upper lever. In the lower level, the follower (travelers) decides on the travel route and modes of transportation. Travelers can use three modes of transport including auto, public transit and walking and they are allowed to switch between these modes along their trips. The problem is given a limited budget, to determine where to install traffic calming facilities (S&C) in a transportation network, to minimize the total transportation cost, and to improve pedestrians' safety.

3.2. Network Reconstruction

A traditional transportation network, even though hosting multi-modal traffic, often allows the same link to be shared by multi-modal traffic instead of tagging each link with a specific mode. In this section, a network reconstruction process is used to convert the existing transportation networks dataset with only one mode (the auto mode) to a multimodal network structure, which is called mode-link structure. Given the road transportation network dataset, this process adds walking and public transit modes as a set of mode-specific links and nodes to the transportation network. In this framework, pedestrians use walking links, auto passengers use auto links, and public transit passengers use public transit links.

17

The reconstruction process presents a transportation network that consists of a network $\mathcal{G}(\mathcal{N}, \mathcal{L})$ made up of a set of links, $\mathcal{L}$, representing road segments, public transit lines, sidewalks and crosswalks, and a set of nodes, $\mathcal{N}$, representing intersections between these links. Network $\mathcal{G}(\mathcal{N}, \mathcal{L})$ is reconstructed from a given only-auto network $G(N, L)$. The original network $G(N, L)$ is made up of a set of links, L, representing only road segments, and a set of nodes, N, representing the intersections between the road segments. The original and the reconstructed networks for the Small network instance are illustrated in Figure 2. The original network $G(N, L)$ in this example as illustrated in Figure 2(a) consists of $N = \{1, 2, 3, 4\}$, representing nodes for the auto mode, and $L = \{(1,3), (1,4), (2,1), (2,3), (3,4)\}$, representing road segments [127].

The reconstruction mapping function $\mathfrak{R}$ is a process that takes the original network $G(N, L)$ and transforms it into a new network that, in addition to road segments and their intersections, also includes public transit lines and S&C with separate links and nodes. For each link $l \in L$, representing a road segment in G, links l_{s1} and l_{s2} are added to the left and right-hand sides of l in parallel, to represent left- and right- hand sidewalks, respectively, for road segment l. Also, two crosswalk links, $l_{c_{begin}}$ and $l_{c_{end}}$, are added, one at the beginning and one at the end of each link l to the network. We refer to these crosswalks as the begin- and end- crosswalks. The crosswalk links cross the auto link l and connect the left- and right- hand sidewalks. In addition to sidewalk and crosswalk nodes and links, public transit nodes and links are also added to the network. To connect these walking and public transit nodes and links to the original network, transfer links (representing the switch from one mode to another) are added to the network. For each origin and destination in the set of origin-destination pairs, dummy origin and destination nodes are added to the network. To ensure connectivity of the network, each dummy node is connected to its nearest sidewalk node using a connector link with zero travel time. This also ensures that travelers start

and end their trips by walking (using sidewalk links). The reconstructed network for the Small

network instance is illustrated in Figure 2(b). In this figure, the dummy nodes are shown in green

and are represented by 5, 6, 7 and 8. Walking nodes are represented with three digit numbers, e.g.

231. The public transit nodes are represented with four digit numbers, e.g. 100.
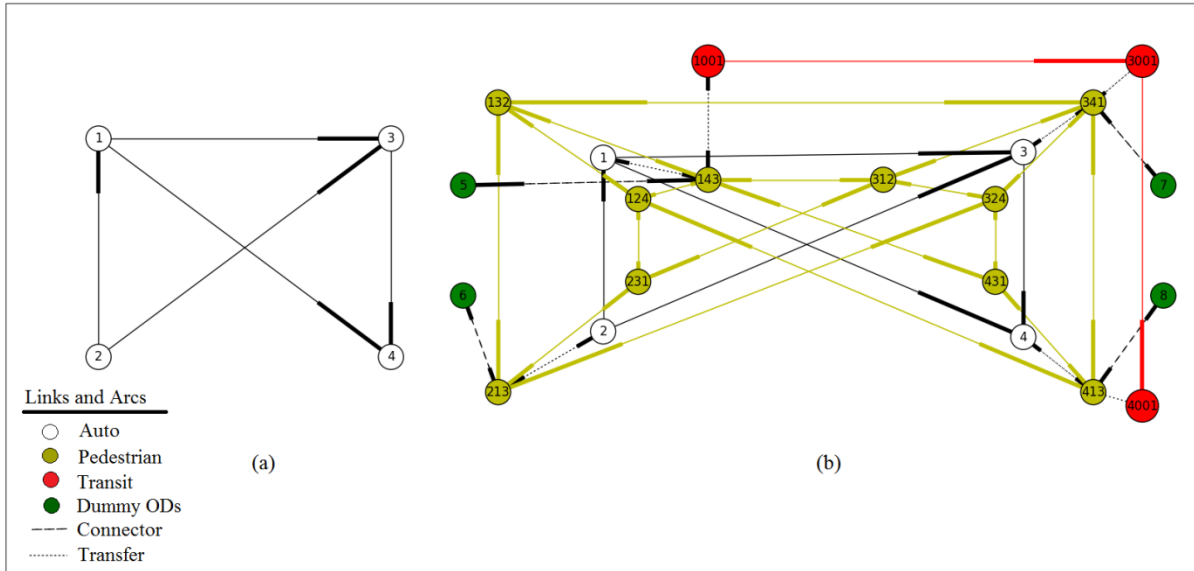
Figure 2 Small network, (a) original network (b) network after reconstruction.

With the reconstruction process, the inter-modal interactions and transfers between

different modes can be modeled in a robust manner. The proposed reconstruction process, which

converts the single mode transportation network to a multimodal network structure, can be adapted

to convert the physical transportation networks to a space-time network structure in CHAPTER 5

(for a more detailed discussion on the reconstruction process and algorithm, please see [127]).

3.3. Modeling Framework

The integrated network design problem is formulated by a bi-level model structure. In the

proposed bi-level model, the upper level problem locates and installs S&C on the reconstructed

network, and the lower level problem solves the corresponding traffic assignment problem and

user equilibrium. The reconstructed network is represented with $\mathcal{G}(\mathcal{N}, \mathcal{L})$ in which $\mathcal{N}$ and $\mathcal{L}$ are

the set of nodes and links. Links in $\mathcal{L}$ can be divided into six categories: A for auto links (which is equal to set L in the original network $G(N, L)$), T for public transit links, S for sidewalk links, C for crosswalk links, F for transfer links that connect walking links to auto and public transit links, and R for connector links which connect the whole network together. The sets of incoming and outgoing links to and from node $i \epsilon \mathcal{N}$, respectively, are represented with $I(i)$ and $O(i)$. Let M be the set of transportation modes that contains 'a', 't' and 'w' representing auto, public transit and walking modes, respectively. Let K represents the set of trips in the transportation network. For trip $k \in K$, d_k is the transportation demand, starting from the origin $\mathcal{O}_k$ and ending at the destination $\mathcal{D}_k$. The set of traffic calming facilities (S&C) is represented with J. Although we assume that the installation of sidewalks includes implementing both left and right-hand sidewalks, for the begin- and end- crosswalks ("c_{begin}" and "c_{end}"), they can be installed independently. If the set of traffic calming facilities available for link $l_a \in A$ (the subscript a indicates that link l is an auto link) is represented by J_{l_a}, then the decision for city planners to make is whether to install the traffic calming facility $j \in J_{l_a}$ that costs $c_{J_{l_a}}$ on the link l_a and imposes the transportation cost $\varphi_l(.)$. The criterion for this decision making is whether the overall transportation cost decreases while pedestrians' safety increases. This decision is limited to budget constraint on implementing traffic calming facilities (not exceeding b) on the transportation network. Table 1 shows all the notations used in the proposed model.

As mentioned earlier, two traffic calming facilities are considered in this research, sidewalks and crosswalks. For crosswalks, we assume that there can be two crosswalk lines on each auto link: one at the beginning (represented by $l_{c_{begin}}$) and one at the end (represented by $l_{c_{end}}$). We assume that the begin- and end-crosswalks can exist and work independently. However, for sidewalks, we assume that the left- and the right-hand sidewalks are dependent and cannot exist

20

Table 1 Notations.

Sets and Indices	N:	Set of nodes indexed by $i = 1, \dots, N$.
	L:	Set of links, indexed by $l = 1, \dots, L$ or (i, i').
	$\mathcal{N}$	Set of nodes in the reconstructed graph indexed by $i = 1, \dots, \mathcal{N}$.
	$\mathcal{L}$	Set of links in the reconstructed network, indexed by $l = 1, \dots, \mathcal{L}$ or (i, i').
	$I(i)$	Set of incoming links to node i.
	$O(i)$	Set of outgoing links from node i.
	M:	Set of transportation modes, denoted by $m = a, t, w$ where "a", "t", and "w" represent auto, public transit, and walking modes respectively.
	J:	Set of traffic calming facilities indexed by $j = c_{begin}, c_{end}, s$, where "$c_{begin}$", "$c_{end}$", and "s" are the begin-crosswalk, end-crosswalk and sidewalk, respectively.
	J_{la}:	Set of traffic calming facilities available on auto link la.
	K:	Set of trips indexed by k.
Parameters and Functions	b:	The budget (\$)
	ϑ:	Value of time (\$/h)
	ω:	Transit-auto equivalent factor
	θ:	Automobile-transit equivalent factor
	τ_{l_f}:	Transfer cost for switching between walking mode to auto or public transit modes (the transfer cost is different for auto and public transit).
	σ	Average cost of a pedestrian crash in dollars (\$)
	t_{l_m}:	Free-flow travel time for link l_m
	$d_k, \mathcal{O}_k, \mathcal{D}_k$	Demand, origin and destination of trip k
	γ_{l_m}:	Capacity of link l_m
	δ:	Safety weight factor that quantifies the travelers' preference between time delay and safety.
	$\varphi_l(.)$:	Travel cost function for link l in the upper level problem
	$\acute{\phi}_l(.)$:	Travel cost function for link l in the lower level problem
	$\psi(X_{l_s})$	The probability of a pedestrian getting into a crash when walking along the auto link adjacent to the sidewalk link l_s
	$P_{la}(X_{l_s})$	Pedestrians' crash probability function on auto link la
	$\alpha_1, \alpha_2, \beta_1, \dots, \beta_6$	Multipliers and powers used in the objective function formulation
Variables	$\pi_{i,k}$:	Auxiliary variable (the dual variables of the corresponding shortest path problem)
	$X_{k,l}$:	Number of trip k flows on link l
	X_l:	The vector of flow variables $(X_{k,l})$ for all trips on link l: $X_l = \{X_{k,l} \vert \forall k \in K\}$
	X:	The vector of all flow variables (X_l) on all links for all trips: $X = \{X_l \vert \forall l \in L\}$
	$Y_{l,j}$:	1 if traffic calming facility j is implemented on auto link l, 0 otherwise
	Y_l	The vector of all $Y_{l,j}$ on link l: $Y_l = \{Y_{l,j} \vert \forall j \in J\}$
	Y	The vector of all Y_l on all links: $Y = \{Y_l \vert \forall l \in L\}$

independently, meaning that if we build a sidewalk for a road, we build it for both sides of the road. We formulate OTCIMTP as a link-based model. Each link in the transportation network has a transportation cost that includes the travel time converted to a dollar value.

3.3.1. Travel Cost Function

For auto links, $l_a \in A$, the travel cost function is as follows:

$$\varphi_{l_a}(X,Y) = \left(t_{l_a} \left(1 + \alpha_1 \left(\frac{\sum_{k \in K} X_{k,l_a} + \sum_{k \in K} \omega X_{k,l_t}}{\gamma_{l_a}} \right)^{\beta_1} \right) \right. \tag{i}$$

$$+ \left(\left(\frac{\sum_{k \in K} X_{k,l_{s1}}}{\gamma_{l_{s1}}} \right)^{\beta_2} + \left(\frac{\sum_{k \in K} X_{k,l_{s2}}}{\gamma_{l_{s2}}} \right)^{\beta_2} \right) \times \left(1 - y_{l_a,s1} \right) \tag{ii}$$

$$+ \left(\frac{\sum_{k \in K} X_{k,l_{c_{begin}}}}{\gamma_{l_{c_{begin}}}} \right)^{\beta_3} \times \left(y_{l_a,c_{begin}} \right) + \left(\frac{\sum_{k \in K} X_{k,l_{c_{end}}}}{\gamma_{l_{c_{end}}}} \right)^{\beta_3} \times \left(y_{l_a,c_{end}} \right) \left. \right) \times \vartheta + \mu_{l_a} \quad (iii) \tag{1}$$

The travel cost of using the auto mode on link $l_a \in A$ is affected by (a) the amount of flow on auto link l_a, (b) the amount of flow on the other links associated with link l_a (public transit and walking modes), and (c) whether traffic calming facilities, S&C, are installed on link l_a. To incorporate these three factors into the travel cost of auto links, we divide the auto travel cost into three parts. We use the coefficient ϑ (the value of time) to convert all these travel costs to a dollar value.

Part (i) reflects the effect of traffic flows of the auto mode in conjunction with the flows of public transit on link l_a. In this formula, t_{l_a} is the free flow travel time on link l_a. The quantities α_1 and β_1 are model parameters. X_{k,l_a} and X_{k,l_t} are the amount of flows for auto and public transit respectively. Here, γ_{l_a} is the capacity of link l_a. Since there is a difference between the size of auto and public transit options (say, the capacity of the public transit is ω times bigger than that of

22

auto), we add ω as a multiplier to X_{k,l_t} to make up for this difference. We refer to ω as the transit-auto equivalent factor.

Part (ii) reflects the effect of traffic flows on the left- and right-hand sidewalks on auto link l_a. $X_{k,l_{s1}}$ and $X_{k,l_{s2}}$ are the number of flows in the left- and right-hand sidewalks, respectively, and $\gamma_{l_{s1}}$ and $\gamma_{l_{s2}}$ are the capacity of the left- and right-hand sidewalks respectively. The quantity β_2 is a model parameter. When a sidewalk is installed, we assume that both the left- and the right-hand sidewalks are built on the sides of a street. The $y_{l_{a,s1}}$ indicates whether sidewalks are installed on the street represented by link l_a or not. When $y_{l_{a,s1}} = 1$, it indicates that two sidewalks are installed on both sides of l_a. Installing sidewalks on a link creates separate walkways for pedestrians which can decrease the travel cost of the auto mode, as the flow of pedestrians no longer interferes with the flow of autos.

Part (iii) reflects the effect of traffic flows on crosswalks (both the begin- and end-crosswalks) on auto link l_a. In this formula $X_{k,l_{c_{begin}}}$ and $X_{k,l_{c_{end}}}$ are the amount of flows in the begin- and end crosswalks respectively, and $\gamma_{l_{c_{begin}}}$ and $\gamma_{l_{c_{end}}}$ are the capacity of these crosswalks. The quantity β_3 is a model parameter. $y_{l_{a,c_{begin}}}$ and $y_{l_{a,c_{end}}}$ are decision variables indicating whether sidewalks are installed or not (at the beginning or at the end of a street). We assume the begin- and end-crosswalks can be implemented separately and independently (as opposed to sidewalks). Installing begin- or end-crosswalks can encourage more pedestrians to cross the auto link l_a and increases its travel cost. This is reflected in part (ii) of equation (1). When a crosswalk is installed ($y_{l_{a,c_{begin}}} = 1$ or $y_{l_{a,c_{end}}} = 1$), the flow of pedestrians affects the travel cost of auto on link l. In this formula, μ_{la} represents the out-of-pocket cost (the indirect costs such as gas,

insurance etc.) that auto drivers have to pay for using auto (there is no such cost when walking or using public transit).

The travel cost function for public transit link $\varphi_{l_t}(.)$ is as follows:

$$\varphi_{l_t}(X,Y) = \vartheta \times t_{l_t}\left(1 + \alpha_2\left(\frac{\sum_{k\in K}(X_{k,l_t} + \theta X_{kl_a})}{\gamma_{l_t}}\right)^{\beta_4}\right) \tag{2}$$

The travel cost of public transit reflects the effect of traffic flows of public transit in conjunction with the flows of auto on link l_t. In this formula, t_{l_t} is the free flow travel time on link l_t. The quantities α_2 and β_4 are model parameters, and γ_{l_t} is the capacity of link l_t. Similar to equation (1), we use a multiplier, θ, this time for X_{k,l_a}, to make up for the difference between the capacity of an auto and that of a public transit. We refer to this multiplier as the auto-transit equivalent factor. We use the coefficient ϑ (the value of time) to convert the travel cost to a dollar value.

The travel cost function for crosswalk link $\varphi_{l_c}(.)$ is as follows:

$$\varphi_{l_c}(X,Y) = \vartheta \times t_{l_c}\left(1 + \alpha_3\left(\frac{X_{l_c}}{\gamma_{l_c}}\right)^{\beta_5}\right) \tag{3}$$

It reflects the effect of traffic flow of the crosswalk on link l_c. In this formula, t_{l_c} is the free flow travel time for crossing a street on link l_c. The free flow travel time can be found based on the length of l_c and walking speed. The X_{l_c} is the flow of pedestrians crossing a street using the crossover link l_c, which has the capacity of γ_{l_c}. The quantities α_3 and β_5 are model parameters. We use the coefficient ϑ (the value of time) to convert the travel cost to a dollar value.

The travel cost function for sidewalk links $\varphi_{l_s}(.)$ is as follows:

$$\varphi_{l_s}(X,Y) = \vartheta \times t_{l_s}\left(1 + \alpha_4\left(\frac{X_{l_s}}{\gamma_{l_s}}\right)^{\beta_6}\right) \tag{i}$$

$$+(1 - y_{l_s}) \times \sigma P_{l_a}(X_{l_a})X_{l_s} \tag{ii} \tag{4}$$

This travel cost function consists of two parts. part (i) reflects sidewalks' travel time and it is a function of the effect of the flow of pedestrians (X_{l_s}) on sidewalk link l_s. In this formula, t_{l_s} is the free-flow travel time, and γ_{l_s} is the capacity of link l_s. The quantities α_4 and β_6 are model parameters. Part (ii) on the other hand, reflects pedestrians' safety and computes the expected cost of pedestrians' crashes (as a penalty cost in the absence of installed sidewalks) on the auto link l_a adjacent to sidewalk l_s. When no sidewalk is installed, pedestrians have to walk along the streets, which is unsafe. To compute the expected cost of pedestrians' crashes, we multiply the probability that a given pedestrian will get in a crash ($P_{l_a}(X_{l_a})X_{l_s}$) by the average cost of a pedestrian crash (σ) [128]. The term $P_{l_a}(X_{l_a})$ is the pedestrian crash probability function when pedestrians walk along the auto link l_a. This function is obtained by dividing the total number of crashes on a given road segment by the total traffic flow on that segment. Then, by using a simple linear regression among all the streets, we calculate the crash probability function. For this regression, we used the historical crash data from Starkville, Mississippi, U.S.A. (see [33] for more details). In this formula, y_{l_s} indicates whether or not a sidewalk is installed, which can enforce the cost associated with pedestrians' crashes in the absence of an installed sidewalk.

For sidewalk links, the travel cost function for the upper level problem (equation (4)) is slightly different from the travel cost function in the lower level. The travel cost function for the lower level is as follows:

$$\acute{\varphi}_{l_s}(X,Y) = \vartheta(1-\delta) \times t_{l_s}\left(1 + \propto_4 \left(\frac{X_{l_s}}{\gamma_{l_s}}\right)^{\beta_6}\right) \qquad (i)$$
$$+ \delta(1 - y_{l_s}) \times \sigma\psi(X_{l_s}) \qquad (ii) \qquad (5)$$

The travel cost function for the lower level also consists of two parts: part (i) reflects sidewalks' travel time, and part (ii) reflects pedestrians' safety on sidewalks. Part (i) is similar to

what is used in equation (4). However, here we also use an adjustment safety weight factor, δ, to combine the travel time part and the safety part. The adjustment weight factor can be used by city planners for find a balance between travel safety and travel time when designing a transportation network. Part (ii) computes the expected cost for an individual when walking along the auto link l_a (adjacent to the sidewalk l_s) for which the sidewalk has not been built. In this formula, $\psi(X_{l_s})$ is the probability that a pedestrian gets into a crash when walking along the auto link l_a (due to the lack of an installed sidewalk) and is computed as follows:

$$\psi(X_{l_s}) = \left(\frac{P_{l_a}(X_{l_a})X_{l_s}}{0.01 \times \gamma_{l_s}} \right) \tag{6}$$

The numerator in (6) computes the expected number of pedestrians' crashes on auto link l_a when pedestrians walk along the street. The denominator in (6) is an estimate of the average number of pedestrians walking along auto link l_a. Due to a lack of data, we assume the average number of pedestrians walking along a street is 1% of the capacity of the adjacent sidewalk if it was built. The links that connect the pedestrian links to auto and public transit links are called "transfer links." The travel cost function for transfer links $\varphi_{l_f}(.)$ is a constant value which indicates the total walking time to reach the transfer station and the total waiting time in the transfer station:

$$\varphi_{l_f}(X,Y) = \vartheta \times \tau_{l_f} \tag{7}$$

There are two types of transfer links: auto-walking and transit-walking transfer links. In order to transfer from public transit to auto (or auto to public transit), both of these transfer links must be used. Finally, the travel cost function for connector links $\varphi_{l_r}(.)$ is assumed to be 0 because these links are hypothetical:

$$\varphi_{l_r}(X,Y) = 0 \tag{8}$$

The objective functions in the upper level and the lower level are different. This is due to incorporating pedestrians' safety on sidewalks into the objective function of the transportation problem in the upper level problem. However, pedestrians' safety is also incorporated into the objective function of the lower level problem. We assume that the travel cost for all links, except sidewalk links, is the same for the upper level problem and the lower level problem: $\acute{\varphi}_l(X, Y) = \varphi_l(X, Y)$ $\forall\, l \epsilon \mathcal{L} \backslash \{l_s | s \in S\}$.

3.3.2. Mathematical Formulation

The proposed mathematical model is formulated as follows:

$$Min \sum_{k \in K} \sum_{l \in L} \varphi_l(X, Y) X_{k,l} \tag{9}$$

$$s.t. \;\; X_{k,l}\left(\acute{\varphi}_l(X, Y) - \left(\pi_{i',k} - \pi_{i,k}\right)\right) = 0 \;\;\;\; , \forall\, l = (i, i') \epsilon \mathcal{L}, k \epsilon K \tag{10}$$

$$\sum_{l \in O(\mathcal{O}_k)} X_{k,l} = d_k \;\;\;\; , \forall\, k \in K \tag{11}$$

$$\sum_{l \in I(i)} X_{k,l} \;\; - \sum_{l' \in O(i)} X_{k,l'} = 0 \;\;\;\; , \forall\, i \epsilon \mathcal{N} \backslash \{\mathcal{O}_k, \mathcal{D}_k\}, k \epsilon K \tag{12}$$

$$\sum_{j \in J} \sum_{l \in L} c_{j,l} y_{l,j} \leq b \tag{13}$$

$$X_{k,l} \geq 0 \;\;\;\; , \forall\, k \epsilon K, l \epsilon \mathcal{L} \tag{14}$$

$$\pi_{ik} \geq 0 \;\;\;\; , \forall\, k \epsilon K, i \epsilon I \tag{15}$$

$$y_{l,j} \epsilon \{0,1\} \;\;\;\; , \forall\, j \epsilon J, l \epsilon L \tag{16}$$

The objective function (9) minimizes the total transportation cost, including the cost of travel time due to a lack of pedestrians' safety, in the network (in the objective function, a lack of safety is penalized with a dollar-value equivalent; therefore, by minimizing total cost, we are also minimizing lack of safety). Constraint (10) enforces the optimal flow solution to be at travel cost equilibrium. Constraint (11) and (12) represent flow conservation for all the nodes and trips. Constraint (13) is the budget constraint, and constraints (14) - (16) serve to restrict the range of

variables. Though this problem is a non-convex mixed-integer nonlinear programming problem, the following proposition shows that the lower level problem is convex.

Proposition 1: The objective function of the lower level problem, $\varphi_l(.)$, is convex.

Proof. See Appendix A. □

For any solution to the upper level problem which locates and installs S&C, we use the nonlinear complementary algorithm [129] to solve the traffic assignment and user equilibrium in the lower level. The nonlinear complementary algorithm does not require the lower level problem to be convex. However, since the problem is convex, any other algorithm that requires convexity can also be used.

3.4. Solution Algorithm

Solving a bi-level network design problem is difficult as the problem is NP-hard. In fact, as Ben-Ayed and Blair [130] showed, even a linear bi-level problem (or bi-level linear problem; BLP) is NP-hard. Therefore, solving the problem for large scale instances using exact solution methods requires extensive computational resources. To solve the proposed bi-level transportation network design problem in this chapter, we use an exact approach and two heuristic algorithms. For the exact approach we implement the model in YALMIP (version 20141030) [131] and solve it using the BARON solver (version v1.69) [132], a computational system that can solve mixed integer nonlinear programming problems. We also develop a greedy heuristic (GH) and a simulated annealing (SA) algorithm. These algorithms are used to solve the upper level problem that is where to install S&C in the network considering the limited budget. Then, a nonlinear complementary algorithm [129] is used to solve the lower level problem (the user equilibrium traffic assignment problem on the reconstructed network). We use a link-list dynamic data structure proposed by Toobaei et al. [133], which was reported to outperform the Frank-Wolfe

algorithm [134]. The advantages of the nonlinear complementary algorithm are its speed and the fact that it allows for a general cost function (i.e. the travel cost is a function of all the flows in the network). In the following sections, the greedy heuristic and the simulated annealing algorithm are described in more detail.

3.4.1. Greedy Heuristic

The greedy heuristic (GH) is a simple heuristic algorithm that makes the locally optimal choice at each stage with the hope of finding a global optimum. The GH in this study starts with a null solution X (no sidewalks or crosswalks installed) and iteratively finds a new solution suggesting where to install a new sidewalk or crosswalk from the set of all possible options ($\acute{X}$) until the budget is exhausted. For a given solution X, the nonlinear complementary algorithm is used to compute its corresponding objective value, $f(X)$. The benefit-cost ratio $\frac{f(X)}{Cost(X)}$ is computed to evaluate the suitability of that solution. The solution that has the largest benefit-cost ratio in that iteration is accepted (i.e. the corresponding sidewalk or crosswalk is installed). Additional S&C are installed at successive iterations in a similar fashion until the budget is exhausted.

Greedy Heuristic

1	Initialization:
	$X \leftarrow \{\}$
	$\acute{X} \leftarrow \{All\ sidewalks\ and\ crosswalks\}$
2	**while** ($Budget\ is\ enough$) **do**
3	Find the sidewalk or crosswalk with the largest cost-benefit gain:
	$x \leftarrow BestTrafficCalming(\acute{X})$
4	$\acute{X} \leftarrow \acute{X} - \{x\}$
5	$Budget \leftarrow Budget - InstallationCost(x)$
6	**end while**

Figure 3 The pseudo code for the proposed GH.

In this study, we assume that the cost of installing S&C at any location is the same. Therefore, instead of $\frac{f(X)}{Cost(X)}$ we can use $f(X)$ in evaluating solution X. The pseudo code of the GH is shown in Figure 3.

3.4.2. Simulated Annealing

An SA algorithm repeats an iterative neighbor generation procedure and follows search directions that improve the objective function value. To escape from local optima, the SA algorithm offers the possibility to accept worse solutions with a probability that decreases as the algorithm moves toward completion. In each iteration, the difference between the objective value of the current solution $f(X)$ and the new solution $f(\acute{X})$ is evaluated as $\Delta = f(X) - f(\acute{X})$. If $\Delta \geq 0$ (for a minimization problem), the new solution $\acute{x}$ is accepted; otherwise, it will be accepted with a probability of $p = \exp\left(\frac{\Delta}{T}\right)$, in which T is a parameter called the temperature of the current state. The factors that influence acceptance probability are the degree of objective function value degradation Δ and the temperature T. Smaller degradation and higher temperature induce higher acceptance probability. The temperature can be controlled by a process called the cooling schedule, which specifies how it should be progressively reduced to make the procedure more selective as the search progresses to neighborhoods of good solutions [135].

The cooling schedule starts with a high temperature T_{max} so that it allows acceptance of new neighbor solutions with higher probability. An attenuation factor $\propto$ ($0 <\propto< 1$) is used to decrease the temperature in each iteration, so the acceptance probability decreases. The algorithm is terminated when the current temperature reaches the minimum temperature (T_{min}).

The neighborhood search used in this study consists of two steps: (1) removing an installed traffic calming facility and (2) installing a traffic calming facility. For a given solution, we first

select an auto link arbitrarily from the pool of auto links with installed traffic calming and uninstall

its traffic calming (if more than one traffic calming is installed on that auto link, one is arbitrarily

selected and uninstalled). Then, we update the budget (the cost of the traffic calming that was

uninstalled is reimbursed).

Simulated Annealing
1   Initialization: generate a random solution $X \leftarrow RandomSolution()$

2   Solve the traffic assignment problem at equilibrium for
    each link:
    $(flow, cost) \leftarrow NonLinearComplementary(X)$

3   Evaluate the objective value for the current solution: $f(X) \leftarrow EvaluateObjVal(flow, cost)$
4    $T \leftarrow T_{max}$
5   **while** $(T > T_{min})$ **do**
6        $i \leftarrow 1$
7   **while** $(i < iter_{max})$ **do**
8            $X' \leftarrow FindNeighborSolution(X)$
9            $(flow, cost)$ $\leftarrow NonLinearComplementary(X')$
10          $f(X') \leftarrow EvaluateObjVal(flow, cost)$
11          $\Delta \leftarrow (f(X) - f(X'))$
12   **if** $(\Delta < 0)$ **then**
13              $BoltzmannValue \leftarrow \exp\left(\frac{-\Delta}{T}\right)$
14   **if** $(BoltzmannValue \leq Random(0,1))$ **then**
15   Accept the new solution: $X \leftarrow X'$
16   **end if**
17   **Else**
18   Accept the new solution: $X \leftarrow X'$
19   **end if**
20          $i \leftarrow i + 1$
21   **end while**
22        $T \leftarrow T * \propto$
23   **end while**

Figure 4 The pseudo code for the proposed SA.

Next, we arbitrarily select another auto link (from the pool of all auto links) and install a

traffic calming (either sidewalk or crosswalk, arbitrarily chosen) on that auto link. If the selected

auto link already has that traffic calming installed, we discard that link and select another one.

Figure 4 shows the pseudo code for the proposed SA, and Table 2 shows the tuned parameters used in the cooling schedule.

Table 2 The SA parameters and their values.

Parameter	Value
T_{max}	42000
T_{min}	0.00001
$Iter_{max}$	20
$\propto$	0.80

3.5. Case Studies

Three sample networks are used for experimentation: a small hypothetical network (called "Small network"), the Hearn network [136] and the Sioux Falls network. The characteristics of these networks are given in Table 3. For the Sioux Falls network, the demand for each origin-destination pair is available. However, for the Small and Hearn networks, which are hypothetical transportation network instances, no such data are available. Therefore, we generate these data arbitrarily for experimentation. Table 4 presents the arbitrarily-generated demand data and Table 5 shows the parameters we use in the travel cost functions that were presented in section 2.

Table 3 Sample transportation networks.

Network	Num. of OD pairs	Original network		Reconstructed Network	
		Num. of nodes	Num. of links	Num. of nodes	Num. of links
Small	4	4	5	21	65
Hearn	4	9	18	55	192
Sioux Falls	552	24	76	143	523

Table 4 Demand for different OD pairs for the Small and Hearn Networks.

Small Network			Hearn Network		
OD Pairs		Demand	OD Pairs		Demand
from	to		from	to	
5	7	10	10	12	20
5	8	40	10	13	40
6	7	20	11	12	60
6	8	60	11	13	80

The parameters used as coefficients in the travel cost functions in Table 5 are commonly

used values. It is assumed that a public transit carries, on average, 20 passengers; 5 times more

than that of an auto [129]. This makes $\omega = \frac{4}{20} = 0.2$, and $\theta = \frac{20}{4} = 5$ as shown in Table 5. The

average cost of a pedestrian crash ($\sigma = \$78300$) is computed using data from Gårder [128]. The

Highway Economic Recruitments System (HERS) [137] considers $19.86 as the value of time (per

hour), which is equal to $ 0.33 per minute. Therefore we assume that $\vartheta = 0.33$ (the value of time).

The coefficients of the intercept and the slope of the linear regression model ($P_{la}(X_{la})$) that is used

in the pedestrian crash probability function in presented equations (4) and (6) are 1.7^{-7} and

3.59^{-10} respectively. Therefore, $P_{la}(X_{la}) = 1.7^{-7} X_{ls} + 3.59^{-10}$ (for more detail we refer the

reader to Parsafard et al. [138])).

Table 5 Parameters' values in travel cost functions used in this study.

Parameter	Value	Description
α_1	0.15	Coefficient of the BPR-like function in (1)
α_2	0.15	Coefficient of the BPR-like function in (2)
α_3	2	Coefficient of the BPR-like function in (3)
α_4	2	Coefficient of the BPR like function in (4) and (5)
β_1	4	Power in equation (1)
β_2, β_3	2	Power in equation (1)
β_4	4	Power in equation (2)
β_5	2	Power in equation (3)
β_6	2	Power in equation (4) and (5)
ω	0.2	Public transit to auto equivalent factor in equation (1)
θ	5	Automobile to public transit equivalent factor in equation (2)
τ	3, 5	The cost for transferring from the walking mode to auto and public transit modes respectively and vice versa (6)
σ	78300	Average cost of a pedestrian crash in dollars (4), (5)
ϑ	0.33	Value of time in (1)-(6)

We used the BARON solver to solve the mathematical model. After running for 24 hours,

BARON did not provide any feasible solution, even for the Small network. On the other hand, the

GH and the SA produced competitive solutions in a reasonable time (as shown in Table 6). Table

6 shows the computational results and the solutions produced by the SA and the GH algorithms

after installing S&C in the transportation networks.

Table 6 A comparison of the computational performance of the SA and the GH
(% reduction in cost $= \frac{f(x^0) - f(x^*)}{f(x^0)}$, x^0: The null solution i.e. when no S&C are installed,
x^*: Solution found by our algorithms).

Networks	% Reduction in cost		Time (sec)	
	SA	GH	SA	GH
Small	% 12	% 12	20.7	0.92
Hearn	% 5	% 5	31	2.6
Sioux Falls	% 59	% 58	199	504

The average demand per OD pair for the Small, Hearn and the Sioux Falls networks is 33,

50, and 653, respectively, and the average demand per link is 2, 1 and 689, respectively. We

speculate that the higher demand per link and per OD pair in the Sioux Falls network might be the

reason for the higher reduction in cost for this network. We will investigate the impact of demand

for these networks later in section 4.3. Regarding the computation time of the two algorithms for

the hypothetical Small and Hearn networks, as shown in Table 6, the GH is faster than the SA;

however, for the Sioux Falls network, the SA is faster than the GH. To investigate the cause of

these differences, we counted the number of times that the nonlinear complementary algorithm is

called by these two algorithms (for solving the user equilibrium in the lower level problem). We

learned that for the Small and Hearn networks, the nonlinear complementary algorithm is called

more often in the SA than the GH. For the Sioux Falls network, however, as the budget increases,

the nonlinear complementary algorithm is called more often in the GH than the SA. Therefore, the

GH becomes computationally more expensive and less efficient than the SA for the larger Sioux

Falls network (Figure 5). In summary, for the larger transportation network (the Sioux Falls), the

SA outperforms the GH both in solution quality and computation time.

As shown in Table 6, installing S&C reduces the total transportation cost. To test the robustness of the results, sensitivity analysis is presented for different parameters such as budget, demand and the safety weight factor.

In the previous experiment (Table 6), we assumed that the budget for installing S&C was unlimited. To test the performance of the GH and the SA with a limited budget, we ran another set of experiments imposing a limit on the budget for installing S&C. Our results show that, as the budget increases and more S&C are installed, the overall transportation cost decreases in all three networks; however, for the Sioux Falls network, this change is more significant. For all three networks, as budget increases, the degree by which the overall cost decreases diminishes, and there exists a point where adding more S&C no longer affects the overall cost (see Figure 6).

(a)

(b)

(c)
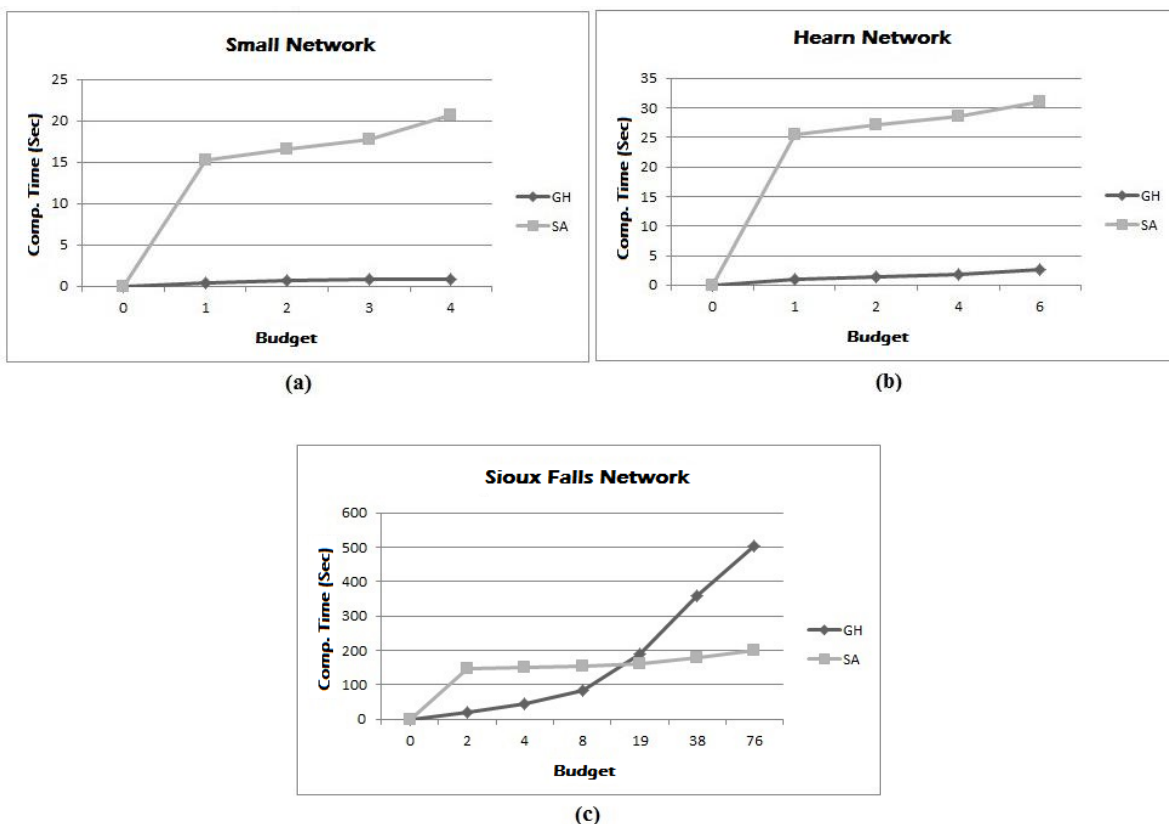
Figure 5 Computation time, the GH vs. the SA for the (a) Small network, (b) Hearn network and (c) Sioux Falls network
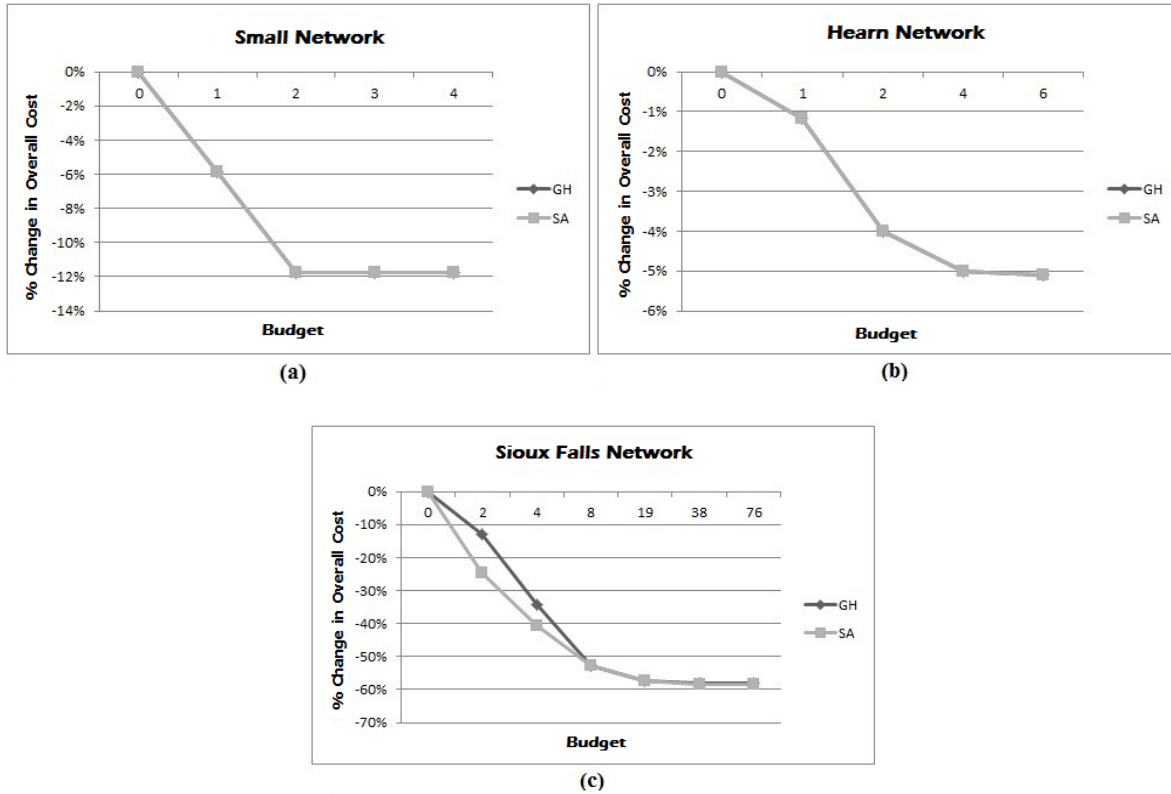
Figure 6 Percent change in the overall cost for different budget values. (a) Small network, (b) Hearn network and (c) Sioux Falls network

For the hypothetical Small and the Hearn networks, there is no significant difference between the quality of the solutions produced by the GH and the SA as shown in Figure 6(a) and (b). However, for the Sioux Falls network, the SA produces better solutions and outperforms the GH with low budget (Figure 6(c)). As the budget increases, the difference between the two algorithms diminishes, however, the SA is still faster.

Although the optimal installation of S&C decreases the overall cost in transportation networks, it has different impacts on the three modes of transportation. For auto and public transit modes, installing S&C means building separate walkways for pedestrians, and it can decrease the travel time of auto and public transit modes. However, for pedestrians, installing separate walkways can decrease crashes and therefore increase safety. For example, in the Sioux Falls network, as shown in Figure 7, installing S&C causes a larger reduction in walking travel costs

than public transit and auto costs. This can be explained by the effect that S&C have on pedestrians' safety (equation (4)).
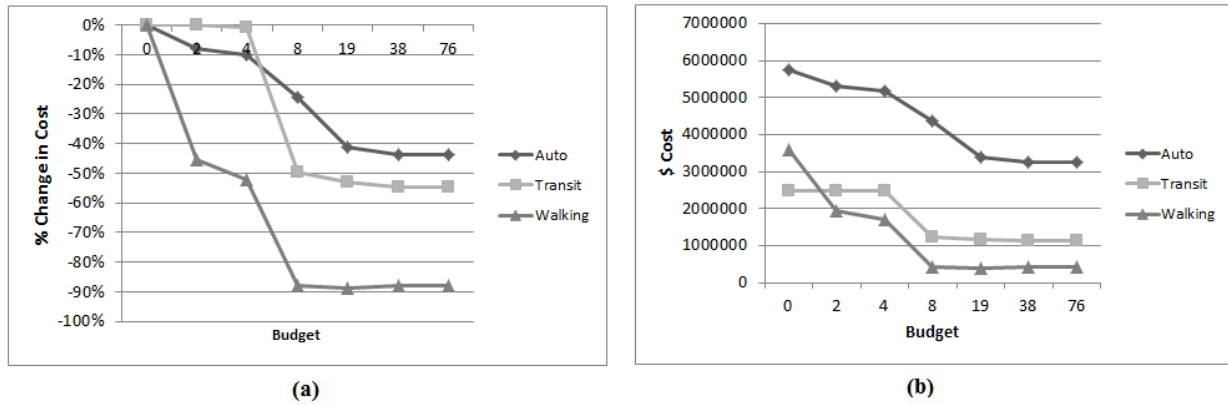
Figure 7 Changes in the overall transportation cost over different transportation modes for the Sioux Falls network. (a) Relative changes (percent of changes) in the overall cost, (b) Absolute changes in the overall cost

To see the impact of installing S&C on the traffic flow in transportation networks considering changes in demand and the safety weight factor, a series of experiments are presented for the Small network. Three scenarios are considered: (1) before and after installing S&C (with a safety weight factor of $\delta = 0.5$, and a demand factor of 100%), (2) for low and high traffic congestion after installing S&C (a low and high demand factor), and (3) for low and high values of the safety weight factor (after installing S&C).

Figure 8(1a) and Figure 8(1b) show the flows in the network before and after installing S&C, respectively. In general, we observed that after installing S&C, fewer autos and more walking are used. More specifically, the auto link (1, 3) is no longer used after installing S&C, and the unused public transit link (1001, 3001) is used after installing S&C. Some walking links are also used more often after installing S&C, such as (2, 324), (324, 431), and (431, 413). The public transit link (3001, 4001) is used less often after installing S&C.

Figure 8(2a) and Figure 8(2b) show the differences under low and high traffic congestion (low and high demand factor), respectively. The model suggests that when the transportation

37

network becomes more crowded, fewer autos and more walking and public transit are used. This result is as expected: when there is high auto traffic congestion, people are more likely to walk than to use their own vehicles because auto traffic congestion causes slow travel speeds.

Figure 8(3a) and Figure 8(3b) show the flow differences for low safety weight ($\delta = 0.1$) and high safety weight ($\delta = 0.9$), respectively. As shown in Figure 8(3a) and Figure 8(3b), the auto links (1, 3) and (2, 1) and the walking links (124, 413), (143, 124), (132, 124), (324, 431), and (431, 413) are used more often for a higher value of the safety weight factor. On the other hand, public transit links (1001, 3001) and (3001, 4001) are used less often. We speculate that increasing the safety weight factor (and therefore putting more emphasis on pedestrians' safety) leads to more use of walking and auto links and less use of the public transit mode. The presence of safe walkways that are separate from roads can encourage more people to walk. As a result, the flow of pedestrians would cause less disturbance to the flow of auto, which can decrease the travel time of auto, increase the use of auto and decrease that of public transit.

3.6. Results and Discussion

The proposed model in this chapter provides insights into how implementing traffic calming facilities such as sidewalks and crosswalks (S&C) affects pedestrians' safety and the transportation cost in a multimodal transportation network. Pedestrians' safety is considered as an important factor in designing a transportation network. A mixed-integer nonlinear programming model is developed for optimally locating S&C in a transportation network. The model is implemented in YALMIP (version 20141030) and solved using the BARON solver (version v1.69), one of the most advanced solvers on the market. However, because of the computational difficulty, greedy heuristic and a simulated annealing algorithms are developed for finding a near optimal solution. Experiments with three sample networks show that these algorithms outperform

38

the BARON solver. Specifically, the SA algorithm is more efficient in producing better quality

solutions for the Sioux Falls network.
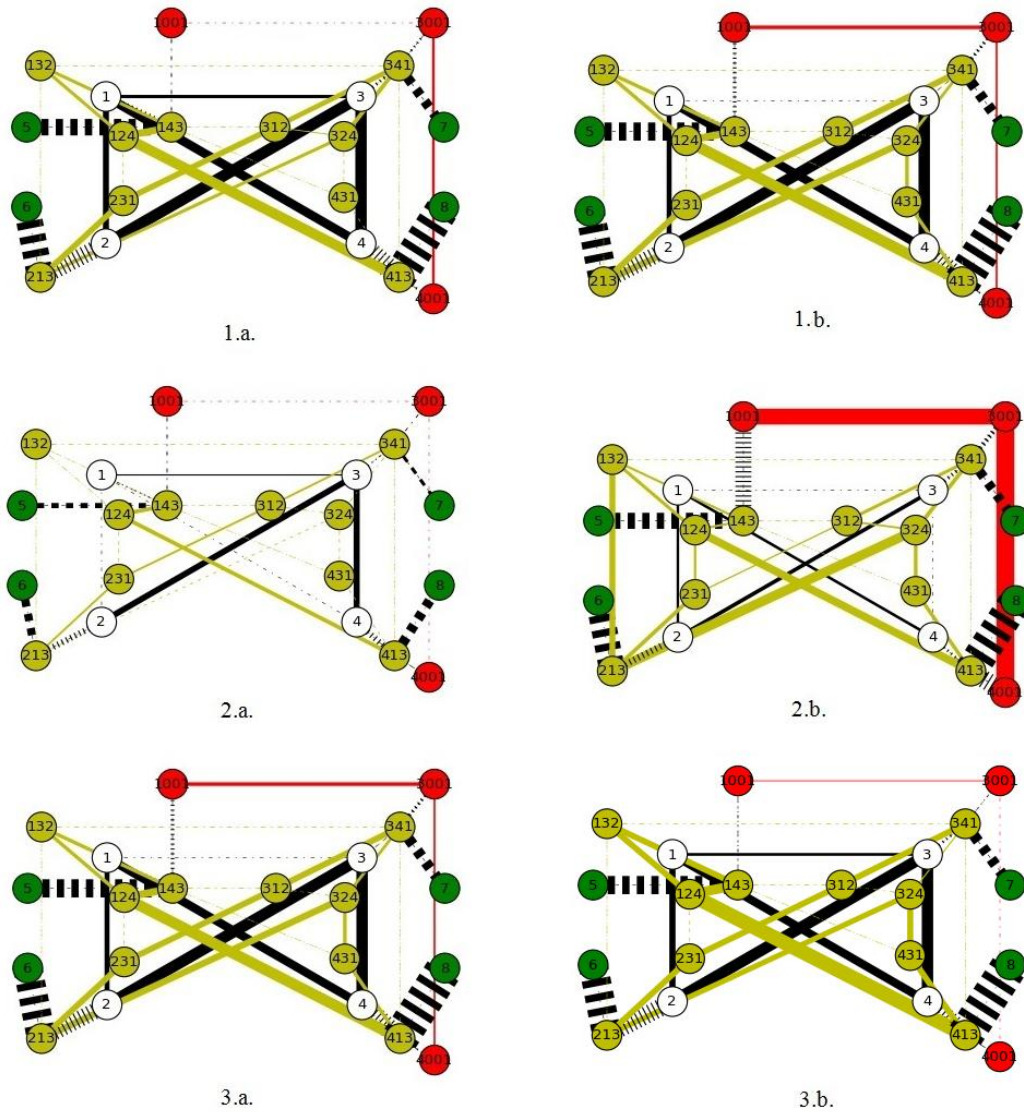
Figure 8 The flow in the Small transportation network under different scenarios. (1.a.) before
and (1.b.) after implementing traffic calming facilities, (2.a.) low and (2.b.) high congestion, and
(3.a.) low and (3.b.) high safety weight factors.

Although these two approximate algorithms do not necessarily find the optimal solutions,

the results are useful in better understanding the impact of traffic calming facilities (in this study,

S&C) in multimodal transportation networks. The results show that installing S&C according to

the solution obtained by the SA algorithm reduces the total transportation cost by 12%, 5% and

39

59% for the Small, Hearn and Sioux Falls networks, respectively. The optimal solutions are expected to yield an even greater reduction in transportation cost. The results suggest that installing S&C not only improves pedestrians' safety, but also reduces the total transportation cost (including the travel cost of auto and public transit). However, S&C provide a smaller reduction of the total transportation cost in a network with low traffic.

This study can be viewed as a foundation for further research on pedestrian transportation network design. Future research can be conducted in several directions. First, the only traffic calming facilities considered in this study are sidewalks and crosswalks. However, there are many more to consider such as speed bumps, stop lights, stop signs and police patrols. One can also study the effect of crosswalks on pedestrians' safety. We assumed that the cost of installing S&C at any location in a transportation network is the same. This assumption can be relaxed by acquiring relevant data. Although safety is an important factor in promoting walking, connectivity is also important in designing a walkable transportation system [139]. Restrictions can be added to ensure connectivity when designing walkways. Further, the problem we studied in this study is deterministic. Therefore, considering uncertainty (in demand, capacity, travel time, etc.) is another way to extend this research. Regarding the solution methodology, since the heuristic methods produce approximate solutions, a useful next step is to develop exact methods for finding the optimal solution. As the problem is bi-level in nature, implementing a decomposition-based method is recommended.

CHAPTER 4: TRAJECTORY ANALYSIS OF MOVING OBJECTS

4.1. Introduction

The trajectory analysis of moving objects, can reveal detailed information about human travel characteristics and provide opportunities to find the users' mobility patterns and presence probability. In a microscopic scale, we can investigate the dynamics of vehicle kinematics and quantify the traffic oscillation based on the space-time trajectories of a set of conductive vehicles following one another along a highway link or segment [31]. In a macroscopic scale, check-in records on location-based social networks (LBSN) contain rich social and geographical information and provide a unique opportunity for researchers to study users' mobility in the network. However, caution must be exercised when studying mobility pattern of moving objects according to their spatiotemporal data. The representativity and the granularity of such data could be relatively limited.

To take the Twitter data as an example, Twitter users might be a special group (e.g., youth) rather than a sound representation of the general population. Despite of a non-representative sample of the whole population, the tremendous available information in such data can be complementary to traditional travel surveys. Furthermore, tweet data sample location and semantic information are available for a user only when he/she tweets, and his/her remaining activities are subject to inaccurate estimation or interpolation. This could be an issue for the data collected in other geo-tagged mobility data (e.g. cellphone data) as well. We address this issue by proposing a new method for trajectory estimation and a set of measures to quantify the accuracy of estimated

trajectories in a robust manner. The first measure is based on absolute inaccuracies between the estimated and ground truth trajectories and the second measure on relative inaccuracies with respect to an individual's overall activity area. Then, we propose cutoff points for screening and selecting the suitable data for mobility analysis. To enable their efficient applications to large-scale data sets (or Big Data), we develop an efficient interpolation method with a lookup table to efficiently solve these measures for geo-tagged data involving a large number of individuals. Multiple sets of real-world geo-tagged data including cell phone records and geo-tagged twitter data are considered as case studies. The results reveal managerial insights on accuracy of trajectory estimations, how it relates to spatiotemporal distributions of geo-tagged data, and overall data qualities for mobility pattern analysis.

4.2. Methodology - Human Trajectory Analysis with Time Geography

This section describes the proposed work in this dissertation on estimating human travel trajectories and their error bounds in a regional scale [140]. We basically extend the techniques in the moving objects literature to develop rigorous definitions of people's activity range, called activity bandwidth. To explain some details, let consider a time period $\mathbf{T}:= [0, T]$ (e.g., a typical day) and a geographical space $\mathbf{C}$ (e.g., a city). We call a traveler's trajectory in space $\mathbf{C}$ over time period $\mathbf{T}$ a *space-time path.* A space-time path typically comprises a number of static stays (or activities) at discrete locations (e.g., home, workplaces, shopping centers, restaurants) according to certain schedules and trips connecting these activities, as illustrated in Figure 9(a). In that figure, the bottom plane denotes space $\mathbf{C}$, and the vertical axis marks time period $\mathbf{T}$. We define a space-time point as a pair of location and time measurements, denoted by (c, t), which mark the traveler's presence at time $t \in \mathbf{T}$ and location $c \in \mathbf{C}$. A space-time path can be specified by the number of critical space-time points $(\bar{c}_1, \bar{t}_1), (\bar{c}_2, \bar{t}_2), \ldots, (\bar{c}_N, \bar{t}_N)$ that mark either the beginning or the ending

of an activity, where N is the total number of activities. With these critical points, the coordinates of this path at any time point $t \in \mathbf{T}$ can be denoted as an interpolation of the neighboring critical points, as follows:

$$\bar{P}(t) := \begin{cases} \dfrac{1}{\bar{t}_{i+1} - \bar{t}_i}\left((\bar{t}_{i+1} - t)\bar{c}_i + (t - \bar{t}_i)\bar{c}_{i+1}\right), & if\ \bar{t}_i \leq t \leq \bar{t}_{i+1}, 1 \leq i \leq N; \\ \bar{c}_1, & if\ 0 < t < \bar{t}_1; \\ \bar{c}_N, & if\ \bar{t}_N < t < T; \end{cases} \tag{17}$$

For ease of notation, we denote this space-time path by $\bar{\mathbf{P}} := \{\bar{P}(t), \forall t \in \mathbf{T}\}$. Note that when a traveler stays at the same location to perform a certain activity over a period of time, the corresponding path segment would be vertical, and its projection to the space plane is a single location. Otherwise, when the individual travels between two activities, the path segment is slanted, and its slope marks the individual's travel speed. In this study, we assume that the maximum speed a traveler could reach is $\bar{v}$. This implies that the inverse of the slope of each path segment should be no greater than $\bar{v}$.

Although it is difficult to track a complete space-time path, discrete sample points along the path may be available in massive social media data such as geo-tagged tweets (Figure 9(b)). With these sample points, we can estimate the individual's trajectory by simply connecting the points with linear segments, as illustrated by the solid curve in Figure 9(b). However, the estimated path is likely different from the ground truth path, particularly when the sample points are sparse. Fortunately, we can use the concept of time geography to quantify the inaccuracy range between the estimated trajectory and any possible ground truth trajectory. We first consider the case when $\mathbf{C}$ is a one-dimensional space.

A space-time cone, as illustrated by the shaded area in Figure 10(a), represents the movement boundary that an individual with a speed limit of $\bar{v}$ can possibly reach if only one space-

time sample point (c_i, t_i) on his/her space-time path $\bar{\mathbf{P}}$ is known. Because of the speed limit, at a

time $t \in [t_i, T]$, this individual has to be at a location, $c_i - \bar{v}(t - t_i)$, if traveling backward at the

maximum speed and $c_i + \bar{v}(t - t_i)$ if traveling forward at the maximum speed. Therefore, his/her

possible presence at time t has to be no less than $c_i - \bar{v}(t - t_i)$ and no greater than $c_i + \bar{v}(t - t_i)$.

With this, we can formulate the upper cone (i.e., the shaded area above point (c_i, t_i)) as follows:

$$\mathbf{O}^+_{(c_i, t_i)} := \{(c, t) | \ |c - c_i| \leq \bar{v}(t - t_i), t \in [t_i, T]\}. \tag{18}$$
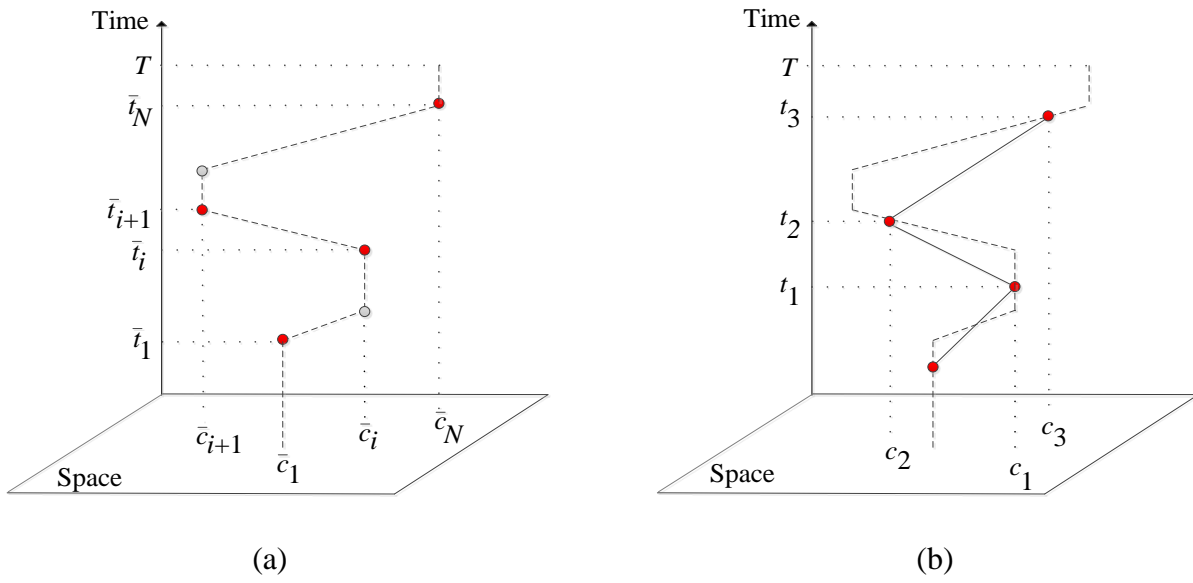
Figure 9 (a) Space-time path and (b) accessible control points on space-time path.
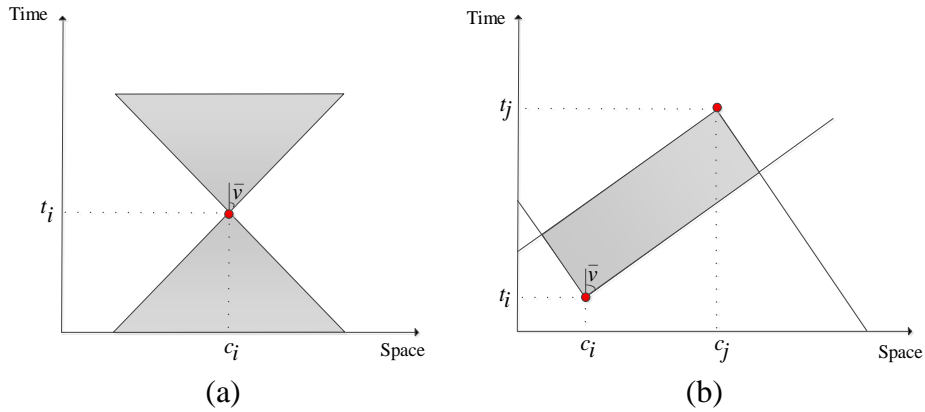
Figure 10 (a) Space-time cone and (b) space-time prism in one-dimensional space

Similarly, at time $t \in [0, t_i]$, this individual has to be between locations $c_i - \bar{v}(t_i - t)$ and $c := c_i + \bar{v}(t_i - t)$. With this information, we can formulate the lower cone (i.e., the shaded area below point (c_i, t_i)) as follows:

$$\boldsymbol{O}^-_{(c_i, t_i)} := \{(c, t) | \ |c - c_i| \le \bar{v}(t_i - t), t \in [0, t_i]\} \,. \tag{19}$$

Then the time-space cone with respect to (c_i, t_i) is simply the union of $\boldsymbol{O}^+_{(c_i, t_i)}$ and $\boldsymbol{O}^-_{(c_i, t_i)}$:

$$\boldsymbol{O}_{(c_i, t_i)} := \boldsymbol{O}^+_{(c_i, t_i)} \cup \boldsymbol{O}^-_{(c_i, t_i)} = \{(c, t) | \ |c - c_i| \le \bar{v}|t_i - t| \quad, t \in [0, T]\} \,. \tag{20}$$

Now suppose that we observe two sample points (c_i, t_i) and (c_j, t_j) of $\overline{\mathbf{P}}$. The space-time range this individual can potentially reach during time period $[t_i, t_j]$ is shown as the shaded area in Figure 10(b). We call this area a *space-time prism*, which is essentially the intersection between $\boldsymbol{O}^+_{(c_i, t_i)}$ and $\boldsymbol{O}^-_{(c_j, t_j)}$, i.e.,

$$\boldsymbol{R}_{(c_i, t_i)(c_j, t_j)} := \{(c, t) | \ |c - c_i| \le \bar{v}(t - t_i), |c - c_j| \le \bar{v}(t_j - t), \quad t \in [t_i, t_j]\} \,. \tag{21}$$

These definitions for a one-dimensional space can easily be extended to a two-dimensional space. Figure 11(a) illustrates the space-time cones, and Figure 11(b) shows the space-time prism in a two-dimensional space.

(a)   (b)

Figure 11   (a) Space-time cones and (b) space-time prism in two-dimensional space.

We essentially need only to revise the distance measure to the Euclidean metric in a two-dimensional space. Then, the cone and prism definitions in equations (18)-(21) can be adapted as follows:

$$\boldsymbol{O}^+_{(c_i,t_i)} := \{(c,t)\mid \|c - c_i\| \leq \bar{v}(t - t_i), t \in [t_i, T]\}, \tag{22}$$

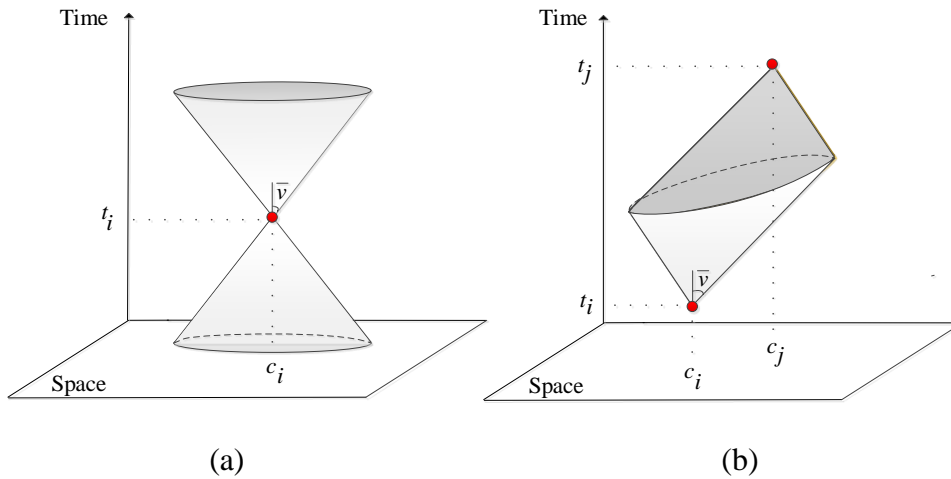$$\boldsymbol{O}^-_{(c_i,t_i)} := \{(c,t)\mid \|c - c_i\| \leq \bar{v}(t_i - t), t \in [0, t_i]\}, \tag{23}$$

$$\boldsymbol{O}_{(c_i,t_i)} := \{(c,t)\mid \|c - c_i\| \leq \bar{v}|t_i - t|, t \in [0, T]\}, \tag{24}$$

$$\boldsymbol{R}_{(c_i,t_i)(c_j,t_j)} := \left\{(c,t)\mid \|c - c_i\| \leq \bar{v}(t - t_i), \|c - c_j\| \leq \bar{v}(t_j - t), t \in [t_i, t_j]\right\}. \tag{25}$$

On the basis of these time geography concepts, some new measures are proposed for characterizing the space-time range of a traveler's trajectory with geo-tagged social media data.

4.2.1. Activity Bandwidth

Given M consecutive sample points and $\boldsymbol{S} := \{(c_i, t_i)\}_{i=1,2,\ldots,M}$ along an unknown underlying space-time path $\overline{\boldsymbol{P}}$ of an individual, we can estimate his/her underlying path with a proper interpolation method (e.g., linear interpolation). As long as the interpolation operation complies with speed limit $\bar{v}$, an estimated path will always be confined within a prism chain, as illustrated in Figure 12 (i.e., the series of space-time cones and prisms determined by $\boldsymbol{S}$:, as follows:

$$\boldsymbol{H}(\boldsymbol{S}) := \left[\boldsymbol{O}^-_{(c_1,t_1)}, \boldsymbol{R}_{(c_1,t_1)(c_2,t_2)}, \boldsymbol{R}_{(c_2,t_2)(c_3,t_3)}, \ldots, \boldsymbol{R}_{(c_{M-1},t_{M-1})(c_M,t_M)}, \boldsymbol{O}^+_{(c_M,t_M)}\right]. \tag{26}$$

Note that the size of $\boldsymbol{H}(\boldsymbol{S})$ bounds the space-time region for any possible ground truth $\overline{\boldsymbol{P}}$. As illustrated in Figure 12, we define $\boldsymbol{P_S}$ as the centerline between sample points $\boldsymbol{S}$ in prism chain $\boldsymbol{H}(\boldsymbol{S})$, or, equivalently, the estimated trajectory obtained with $\boldsymbol{S}$ using linear interpolation:

$$\boldsymbol{P_S}(t) := \begin{cases} \dfrac{1}{t_{i+1} - t_i}\left((t_{i+1} - t)c_i + (t - t_i)c_{i+1}\right), & \text{if } t_i \leq t \leq t_{i+1}, 1 \leq i < M; \\ c_1, & \text{if } 0 < t < t_1; \\ c_M, & \text{if } t_M < t < T; \end{cases} \tag{27}$$

Then, we define an *activity bandwidth* with respect to $\mathbf{S}$, denoted by $B(\mathbf{S})$, which is the average distance between a generic point c in $\mathbf{H}(\mathbf{S})$ and the corresponding $\mathbf{P_S}(t)$ divided by the chain volume, as indicated in Figure 12:

$$B(\mathbf{S}) = \frac{\int_{(c,t)\in H(\mathbf{S})} \|c - \mathbf{P_S}(t)\| \, dcdt}{\int_{(c,t)\in H(\mathbf{S})} dcdt}, \qquad (28)$$

where operator $\|\cdot\|$ in the numerator is the Euclidean distance between two points on a two-dimensional plane, and the denominator is the volume of prism chain $\mathbf{H}(\mathbf{S})$. Note that a large $B(\mathbf{S})$ indicates that the inaccuracy between $\mathbf{P_S}$ and $\overline{\mathbf{P}}$ is likely high, whereas a small $B(\mathbf{S})$ value implies that $\mathbf{P_S}$ is likely close to the $\overline{\mathbf{P}}$.
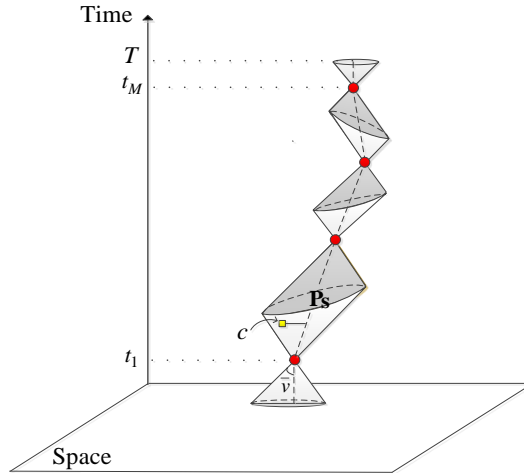
Figure 12 Space-time prism chain.

Despite the compact formulation, the integral equation (28) cannot be resolved into an analytical form and has to be solved numerically. To discuss the problem, we first decompose the prism chain to its cones and prisms. Basically, for M consecutive sample points, there are $M + 1$ space-time cones and prisms in the prism chain. After decomposition, both numerator and denominator in equation (28) are decomposed to $M + 1$ components, where each component corresponds to a single cone or prism. If U_m^c and D_m^c $(m = 1, M)$ specify the cone components in

47

the numerator and denominator, respectively, and U_m^{p} and D_m^{p} ($m = 2, 3, \dots, M$) correspond to the

prism components, then equation (28) could be rewritten as follows:

$$B(S) = \frac{U_1^c + \sum_{m=2}^{M} U_m^{\mathrm{p}} + U_M^c}{D_1^c + \sum_{m=2}^{M} D_m^{\mathrm{p}} + D_M^c}, \tag{29}$$

where

$$D_1^c = \int_{(c,t)\in O_{(c_1,t_1)}^-} dc\,dt, \tag{30}$$

$$D_m^p = \int_{(c,t)\in R_{(c_{m-1},t_{m-1})(c_m,t_m)}} dc\,dt, \tag{31}$$

$$D_M^c = \int_{(c,t)\in O_{(c_M,t_M)}^+} dc\,dt, \tag{32}$$

and

$$U_1^c = \int_{(c,t)\in O_{(c_1,t_1)}^-} \|c - P_S(t)\|\,dc\,dt, \tag{33}$$

$$U_m^p = \int_{(c,t)\in R_{(c_{m-1},t_{m-1})(c_m,t_m)}} \|c - P_S(t)\|\,dc\,dt, \tag{34}$$

$$U_M^c = \int_{(c,t)\in O_{(c_M,t_M)}^+} \|c - P_S(t)\|\,dc\,dt. \tag{35}$$

In equation (29), D_1^c and U_1^c are associated with the lower cone $O_{(c_1,t_1)}^-$, D_M^c and U_M^c are

associated with the upper cone $O_{(c_M,t_M)}^+$, and D_m^{p} and $U_m^{\mathrm{p}}, \forall\, m = 2, \dots, M$ are associated with the

prisms between. Note that the terms in the denominator, D_1^c, D_M^c, and D_m^{p}, are essentially the

volumes of the corresponding cones and prisms, and the terms in numerator, U_1^c, U_M^c and U_m^{p},

correspond to their angular momentums. Actually, these terms can be calculated as certain

functions of the relative difference between corresponding sample points, as described in the

following propositions (see Appendix B for the proofs).

Proposition 2: Given (c_1, t_1) and (c_M, t_M), we have $D_1^c = D^c(t_1)$ and $D_M^c = D^c(T - t_M)$, where function $D^c(t) := \frac{1}{3}\pi\bar{v}^2 t^3, \forall t \in [0, \infty)$ (note that $\bar{v}$ and T are given parameters).

Proposition 3: Given (c_1, t_1) and (c_M, t_M), we have $U_1^c = U^c(t_1)$ and $U_M^c = U^c(T - t_M)$, where function

$$U^c(t) := \int_0^{2\pi} \int_0^{tan^{-1}(\bar{v})} \int_0^{\frac{t}{cos\varphi}} \sqrt{(\rho sin\varphi cos\theta)^2 + (\rho sin\varphi sin\theta)^2}\,\rho^2 sin\varphi\,d\rho d\varphi d\theta, \forall t \in [0, \infty).$$

Proposition 4: Given two consecutive control points (c_{m-1}, t_{m-1}) and (c_m, t_m), $D_m^p = D^p(\|c_m - c_{m-1}\|, |t_m - t_{m-1}|), \forall 2 \leq m \leq M$, where function

$$D^p(c, t) := 2\int_0^{2\pi} \int_0^{tan^{-1}(\bar{v})} \int_0^{\frac{\bar{v}^2 t^2 - c^2}{(2\bar{v}^2 t)cos\theta - (2c)sin\theta sin\varphi}} \rho^2 sin\varphi\,d\rho d\varphi d\theta, \quad \forall c, t \in [0, \infty)$$

Proposition 5: Given (c_{m-1}, t_{m-1}) and (c_m, t_m), $U_m^p = U^p(\|c_m - c_{m-1}\|, |t_m - t_{m-1}|), \forall 2 \leq m \leq M$, where

$$U^p(c, t) := 2\int_0^{2\pi} \int_0^{tan^{-1}(\bar{v})} \int_0^{\frac{\bar{v}^2 t^2 - c^2}{(2\bar{v}^2 t)cos\theta - (2c)sin\theta sin\varphi}} Q\rho^2 sin\varphi\,d\rho d\varphi d\theta, \quad \forall c, t \in [0, \infty),$$

and $:= \sqrt{(\rho sin\varphi cos\theta)^2 + \left(\rho sin\varphi sin\theta - (\frac{c\rho cos\varphi}{t})\right)^2}$.

We see that only function $D^c(t)$ can be solved analytically in a closed form equation defined in Proposition 2, and all other functions defined in Propositions 3-5 do not have closed form formulations and thus have to be solved numerically. Because the numerical solution to a complex integral takes much longer than an analytical computation, calculating these terms for big datasets (e.g., tweet data from millions of travelers) would consume excessive computation resources.

By observing that these functions have at most two variables, we propose a lookup table–based interpolation method that circumvents the need for a time-consuming numerical solution

approach to alleviate the computational load. Basically, for each variable in each function, we identify a finite interval that can cover most practical values of the variable, and then we position a number of ticks along that interval. If we have a set of sample values for the variable, we can place denser ticks in areas in which more sample values more likely fall, instead of evenly distributing them. For a variable with sample values, we first divide their span into K consecutive intervals with equal length l, and f_k denote the number of samples in the kth interval, $\forall k = 1, \cdots, K$, where K is a proper number picked based on the sample distribution. Then we evenly place a number of ticks in each interval k, and this number is calculated as

$$\omega_k = \frac{\sqrt{(A_k f_k)}}{\sum_{j \in K} \sqrt{(A_j f_j)}} \Omega \quad, \quad \forall k \in K \;, \tag{36}$$

where Ω is the total number of ticks selected based on the computational resources. The number K should be selected such that each interval has a sufficient number of samples and there are enough intervals to allow the ticks to be heterogeneously distributed across the entire feasible range of the variable. Once we obtain the ticks for all variables, the combinations of these ticks across the variables form a mesh that covers the feasible region of this function.

We first pre-calculate the function value at each grid point on the mesh and store the function value in a lookup table indexed by the corresponding variable values. This pre-calculation need be executed only once, and then every time when receiving a set of variable values, we can quickly approximate the corresponding function value by linearly interpolating the lookup table values at the nearest grid points. Table 7 is a schematic view of a lookup table for a general function $z = f(c, t)$, where the number of ticks for variable c is Ω and for t is Ψ.

The lookup table method provides significant savings in computational time compared with the numerical approach, particularly when the data set is big. Although the lookup table method is

essentially an interpolation approach and may have approximation inaccuracies, our case study shows that those inaccuracies are well controlled.

4.2.2. Normalized Activity Bandwidth

Activity bandwidth $B(\mathbf{S})$ indicates the inaccuracy bound between an estimated path (e.g., centerline $\mathbf{P_S}$) and ground truth path $\overline{\mathbf{P}}$. A relatively small activity bandwidth means that $\mathbf{P_S}$ is likely close to $\overline{\mathbf{P}}$. However, when control points $\mathbf{S}$ are spatially close to each other, even if the absolute activity bandwidth value is small, it is still difficult to discern an individual's activities.

Table 7 A schematic form of lookup table for estimation of $z = f(c, t)$.

c	t			
	t_1	t_2	$\ldots$	t_Ψ
c_1	$f(c_1, t_1)$	$f(c_1, t_2)$	$\ldots$	$f(c_1, t_\Psi)$
c_2	$f(c_2, t_1)$	$f(c_2, t_2)$	$\ldots$	$f(c_2, t_\Psi)$
.	.	.	.	.
.	.	.	.	.
.	.	.	.	.
c_Ω	$f(c_\Omega, t_1)$	$f(c_\Omega, t_2)$	$\ldots$	$f(c_\Omega, t_\Psi)$

As illustrated in Figure 13, although the activity bandwidth of the chain on the left is smaller than that on the right (which indicates smaller estimation inaccuracies), the control points are clustered around the same location, so it is difficult to use them to estimate the various activities of this individual over time. For example, if this location is mixed land use, the individual could conduct a series of different activities that may not be reflected by the clustered control points. On the other hand, although the activity bandwidth of the chain on the right is larger (which indicates larger estimation inaccuracies), the control points are far apart. The associated activity types may therefore more easily be inferred based on the different characteristics of these locations; thus, this chain may better help us understand this individual's activity pattern.
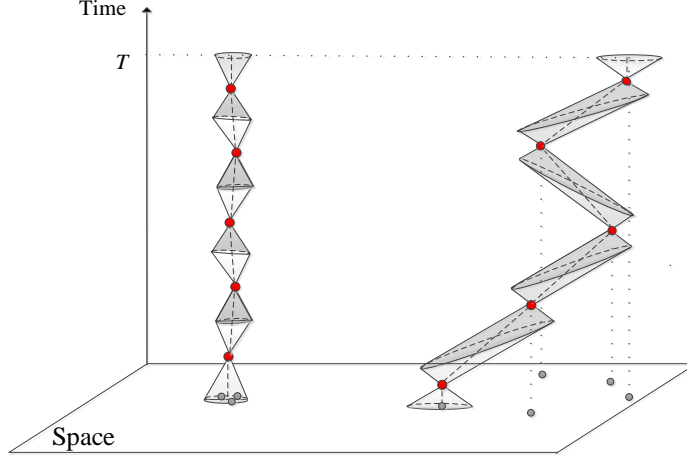
Figure 13 Space-time prism chain with small and large radiuses of gyration.

To solve this challenge, we normalize the activity bandwidth by the *radius of gyration* [141], which measures the spread of an individual's locations around his/her center of mass (the standard deviation of distances between these locations and the individual's center of mass). For an individual with M consecutive control points (as defined in Figure 13), the center of mass of the control points is formulated as $\bar{c} := \sum_{m=1}^{M} c_m / M$, and the radius of gyration is defined as follows:

$$g := \sqrt{\frac{\sum_{m=1}^{M}(c_m - \bar{c})^2}{M}} .$$

(37)

In fact, g measures the dispersion of an individual's sampled locations and indicates how far he/she moves on average. A small g value means that, overall, the individual travels locally, while a large g implies long-distance travels. With this definition, we can adapt $B(\mathbf{S})$ to *normalized activity bandwidth*, defined as follows:

$$NB(\mathbf{S}) := \frac{B(\mathbf{S})}{g},$$

(38)

With this measure, suitable data for studying individual activity patterns are the ones with small normalized activity bandwidths (e.g., with a small activity bandwidth and a large gyration).

4.3. Case Studies

4.3.1. Twitter Data

This section presents a case study on a geo-tagged tweet data set gathered in New York City from 10:47 p.m. on June 28 through 4:27 a.m. on July 18, 2013. Note that tweet data have a number of intrinsic limitations for mobility studies, such as biased representativeness, sparse sampling rates, and potential location errors. Despite so, tweet data are available for free, and for a study that focuses on the methodology, tweet data are reasonable for illustrating the applications of the proposed methods.

Each tweet consists of a number of fields, including the tweeter's name, the tweet ID, the date and time of the tweet, the geographic coordinates of the tweet, the language, the tweeter's number of followers, and the text of the tweet. The format of the tweet data is illustrated below (where we modified certain fields to anonymize this sample tweet).

"Azama_2_", 350329451143384562, Thu Jun 27 19:07:04 +0000 2013, 40.6823018, -73.3945501, en, 128, "Hello!".

This study uses only the tweet ID, the date and time, and the geographic coordinates. The tweet ID is used to connect tweets from the same individual. The date and time and the geographic coordinates in all collected tweets from the same individual (sorted by time in ascending order) specify the sample space-time points $\{(c_m, t_m)\}$ for the individual. According to our data set, the basic problem settings and assumptions are as follows:

- For more than 98% of individuals traveling in New York City, the travel speed ($\bar{v}$) falls below 30 km/h. Thus, the analysis in this section are presented for six different $\bar{v}$ values including 5, 10, 15, 20, 25 and 30 km/h.

- For two consecutive points (c_{m-1}, t_{m-1}) and (c_m, t_m) for an individual, the activity bandwidth is set to zero if one of the following three events happens: $\|c_m - c_{m-1}\| <$ 0.1km, $|t_m - t_{m-1}| < 0.01$hr, or $\|c_m - c_{m-1}\|/|t_m - t_{m-1}|$ is greater than or equal to the corresponding $\bar{v}$.

- We screen out individuals without any tweets in the first three days (June 28 through July 1) or any tweets in the last three days (July 15 through July 18) because the prism chains of those users have larger lower or upper cones and do not contain much information about activity patterns.

The original data contain information on 93,316 individuals and 1,012,912 tweets during this three-week period. The distribution of the number of tweets per individual has been shown in Figure 14. Note that this distribution is aggregated across the time and may not be comprehensive in evaluating tweet data and understanding the estimation accuracy. For instance, if a Twitter user tweets very frequently during a short period of time but keeps silent afterwards, the sample points may not be as useful as those users that tweet with the same frequency but more evenly distributed over the time. After applying the screening criteria explained above, we keep only 11,734 individuals with 486,114 tweets. Note that we assume that the geo-tagged tweet data is detected from those users that have been enabled tweeting with location and device's precise location is identified. Usually, the GPS devices such as smartphone's GPS sensor or wifi hotspots provide location accuracy to within a few meters. Note that for a few users the tagged locations (e.g. when one tags a neighborhood to a tweet) may be differ from the actual location. However, it is easy to identify user tagged locations versus actual GPS locations and we can only use the actual locations instead. Although a large number of the original data are screened (which is not surprising because social media data often require a considerable amount of cleaning), a relatively large number of

54

tweets remain, and they contain rich information about activity patterns. For the purpose of verification, the following analysis shows that the screened data inherit certain fundamental statistical properties of the original tweet data discovered in the literature. Figure 15 plots scattered data for each individual's displacement (or travel distance) between two consecutive locations and for each individual's radius of gyration (g).
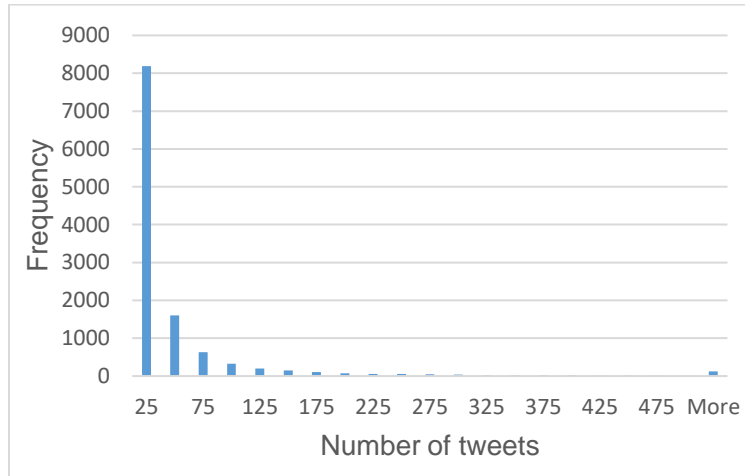
Figure 14 The distribution of number of tweets.

We find that the displacement r fit a power-law distribution $p(r) \sim r^{-(1+\beta)}$, where $\beta = 4.64$ (see Appendix C for goodness of fit test). In addition to the displacement, the gyration also reasonably fits a power-law distribution after the initial portion is truncated for $g > 8$ km and $\beta = 1.40$. These findings are consistent with previous studies on general social media data [11], [40], which implies that the screening does not introduce much bias against important statistical properties. Note that the observed exponents for the gyration is pretty close to González et al., (2008) approximation, where $\beta = 1.65 \pm 0.15$. However, for the displacement, the coefficients are slightly different where in their study $\beta = 1.75 \pm 0.15$. This could be because of differences in geographical areas, data sources and power-law fitting models (the previous study uses $p(r) = (r + r_0)^{-\beta} \exp(^r/_\kappa)$ where $r_0 = 15\ km$ and $\kappa = 400\ km$).
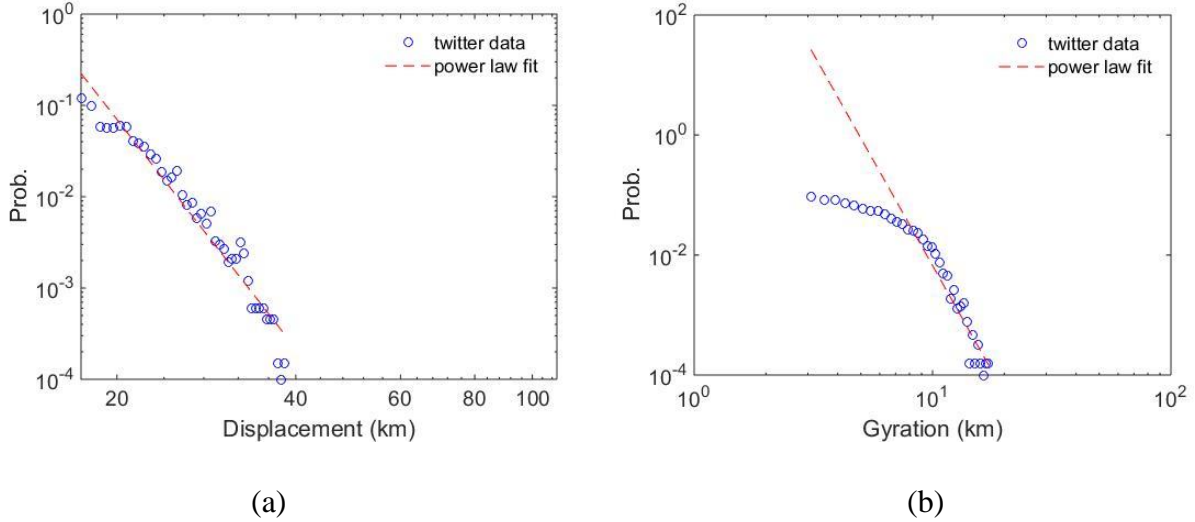
Figure 15 Power-law distribution (a) displacement and (b) radiuses of gyration.

To calculate $B(\mathbf{S})$ for each individual, we first prepare the lookup tables for all the terms

in equation (29) that do not have closed form formulations (functions defined in Propositions 3–

5). The entries in the tables are calculated only once but can be repeatedly used to approximate the

corresponding terms for arbitrary individuals. As explained in Propositions 2–5, for a given $\bar{v}$ each

of the terms has at most two variables, t and c. The number of ticks of each variable is determined

by equation (36), where Ω is set to 100 (which is manageable for our available computational

resource and serves the illustration purpose well). For instance, given $\bar{v} = 30$ km/h, Table 8

illustrates the lookup table for function $U^{\mathrm{c}}(t)$ in Proposition 3, where t is the only variable in this

function. Also, Table 9 and Table 10 show a snapshot of the lookup tables for the $D^{\mathrm{p}}(c, t)$ and

$U^{\mathrm{p}}(c, t)$ functions, respectively. Both functions include c and t as variables. It is theoretically

difficult to calculate the derivatives of $B(\mathbf{S})$ with respect to t or c, and examine how its value is

changing when we increase (decrease) these variables. However, considering Table 8 to Table 10

we can see how different terms in $B(\mathbf{S})$ are changing with respect to c or t. If we integrate these

tables together and calculate $B(\mathbf{S})$, we observe that the value of $B(\mathbf{S})$ increases with the increase of t, and decreases with the increase of c.

Table 8 Lookup table for $U^{\mathrm{c}}(t)$.

t (hour)	0.205	0.409	0.614	…	455.4	451.95	458
$U^{\mathrm{c}}(t)$	1693.77	26836.89	136305.2	…	3.77E+16	4.00E+16	4.24E+16

Table 9 Lookup table for $D^{\mathrm{p}}(c, t)$.

c (km)	t (hour)			
	0.205	0.409	…	451.950	458.5
---	---	---	---	---	---
0.062	2.029	16.119	…	21,751,154,251	22,710,628,653
0.125	2.028	16.118	…	21,751,154,249	22,710,628,651
.	.	.	.		
.	.	.	.		
.	.	.	.		
39.690	0	0	…	21,750,874,668	22,710,345,018
40.500	0	0	…	21,750,863,140	22,710,333,323

To illustrate the efficiency of the proposed lookup tables in calculating $B(\mathbf{S})$, Table 11 compares the solution times from the lookup table approach with those from the traditional numerical approach for different data sets. For these experiments, we randomly select four different geo-tagged Twitter data sets of relatively small sample sizes (so that the experiments are manageable). All the experiments are run on a typical PC with 2.2 GHz CPU and 8 GB RAM. The Scipy module in the Python programming language is used for the numerical approach. We see that for all instances in Table 11, the lookup table approach dramatically reduces the solution time compared to the numerical approach: the ratio of the numerical approach solution time to the lookup table solution time is always greater than 20,000. With this performance, we expect that the absolute computational time savings is even more considerable as the data size further increases. Therefore, for larger data sets in realistic mobility pattern studies, the numerical method

may not be feasible (taking months), while the lookup table approach can yield solutions in a very

short time (a few minutes).

Table 10 Lookup table for $U^{\mathrm{p}}(c,t)$.

c (km)	t (hour)				
	0.205	0.409	...	451.950	458.5
0.062	3.120	49.447	...	73,728,256,227E+3	78,096,174,278E+3
0.125	3.118	49.438	...	73,728,256,216E+3	78,096,174,267E+3
.	.	.	.	.	.
.	.	.	.	.	.
.	.	.	.	.	.
39.690	0	0	...	73,726,834E+6	78,094,711E+6
40.500	0	0	...	73,726,776E+6	78,094,650E+6

Table 11 Comparison of solution time of $B(\mathbf{S})$ with numerical approach and lookup table.

Data set	Sample size (number of individuals)	Number of tweets	Solution time (minutes)		Solution time ratio (numerical approach: lookup table)	Relative error (E_r)	
			Numerical approach	Lookup table		Average	95 percentile
1	7	2000	19.31	0.0009	21,456 : 1	0.0083	—
2	102	4000	86.21	0.0016	53,881 : 1	0.0133	0.01
3	192	8000	206.77	0.0037	55,884 : 1	0.0104	0.02
4	303	16000	325.90	0.0077	42,325 : 1	0.0116	0.02

Despite the superior computational performance, the lookup table approach produces an

approximation error caused by linear interpolation. To quantify this approximation error, a relative

error is formulated to measure the difference between the activity bandwidth obtained from the

numerical approach, denoted by $B_{num}(\mathbf{S})$, and that obtained from the lookup table, denoted by

$B_{lookup}(\mathbf{S})$, for one individual:

$$E_r := \frac{|B_{num}(\mathbf{S}) - B_{lookup}(\mathbf{S})|}{B_{num}(\mathbf{S})}. \tag{39}$$

Table 11 reports the average E_r value and the 95 percentiles E_r value across all individuals in each instance. Overall, the average E_r values are no greater than 0.01. Also, E_r is no greater than 0.02 for more than 95% of the population in all instances. Although the results in Table 11 are based on a given $\bar{v} = 30$, the same E_r value is gained by replication of the experiment for all the other predefined $\bar{v}$ values ($\bar{v} = 5, 10, 15, 20, 25, 30$). Such an error magnitude is acceptable for most engineering applications. Note that this error can be reduced even further as we increase the density of the lookup table ticks.

The distribution of B for all $\bar{v}$ values is shown in Figure 16 and, interestingly, this distribution can be well fit with a power-law distribution $B^{-\beta}$. We used a python package called "power law" [142] for fitting the power-law distributions and compare it with the other alternative heavy tailed distributions. As an observation, by increasing $\bar{v}$ from 5 km/h to 30 km/h, the power-law exponent (β) decreases from 2.93 to 1.68, which suggests the power-law fit has a well-defined mean ($\beta > 2$) for the smaller $\bar{v}$ values. We also use a log likelihood ratio test to compare the exponential distribution with power-law distribution and identify which of these two fits the data better. A large positive loglikelihood ratio and a very small p-value for all cases indicate that the data is more likely in the power-law distribution (To evaluate the power-law distribution individually, the goodness-of-fit results are provided in Appendix C). This finding reveals an interesting pattern about how frequently people travel and tweet. The long tail of this power-law distribution indicates that the majority of the individuals have relatively large activity bandwidths; thus, the ground truth trajectory cannot be accurately estimated. However, around the head of the distribution, we find a large number of people who have small activity bandwidths as a result of high frequencies of their tweets, and their data can be used to construct relatively accurate space-time trajectories.

Figure 16 Power-law and exponential distribution fit for activity bandwidth with different traveling speed limits (km/h). (a) $\bar{v} = 5$ (b) $\bar{v} = 10$ (c) $\bar{v} = 15$ (d) $\bar{v} = 20$ (e) $\bar{v} = 25$ (f) $\bar{v} = 30$.
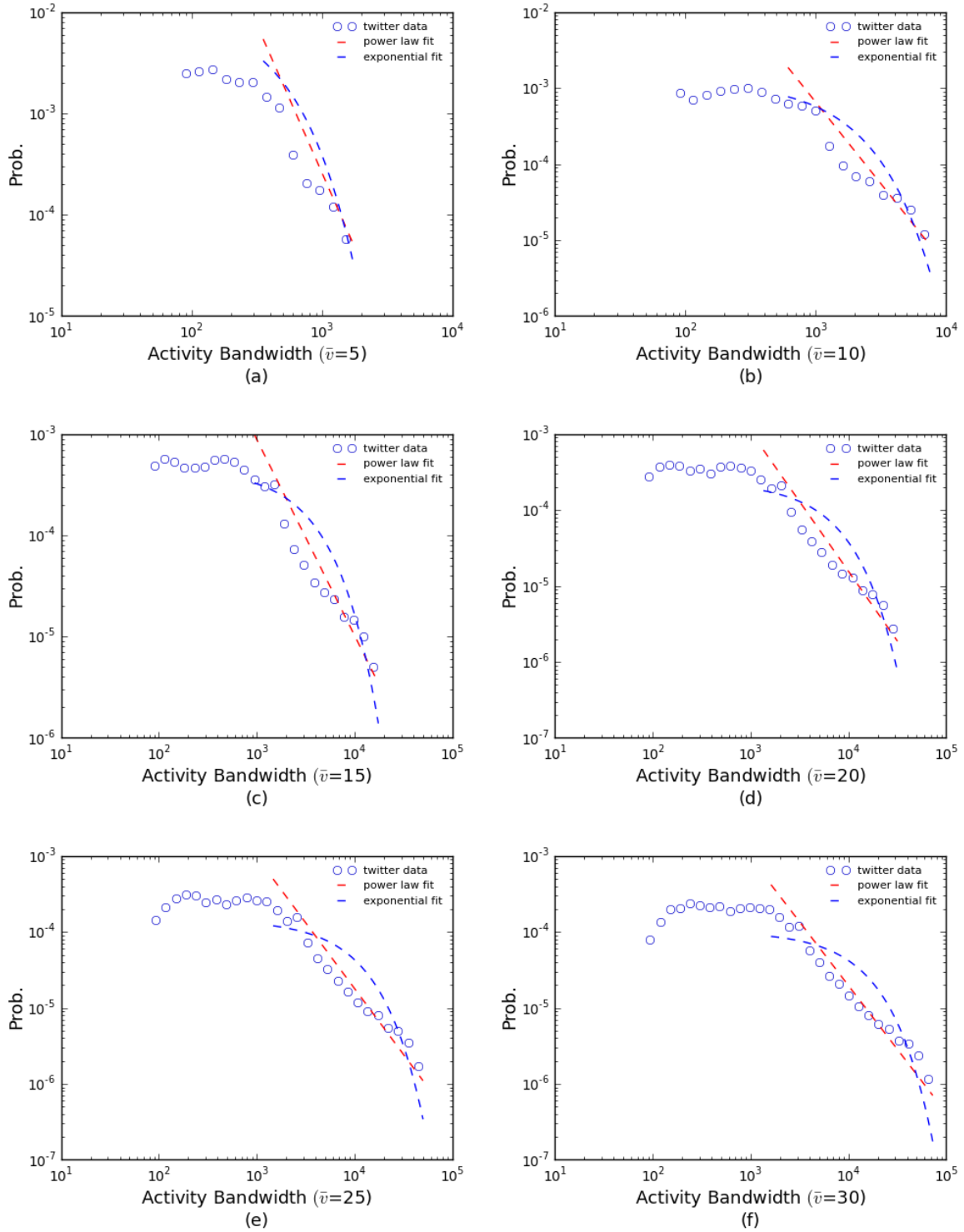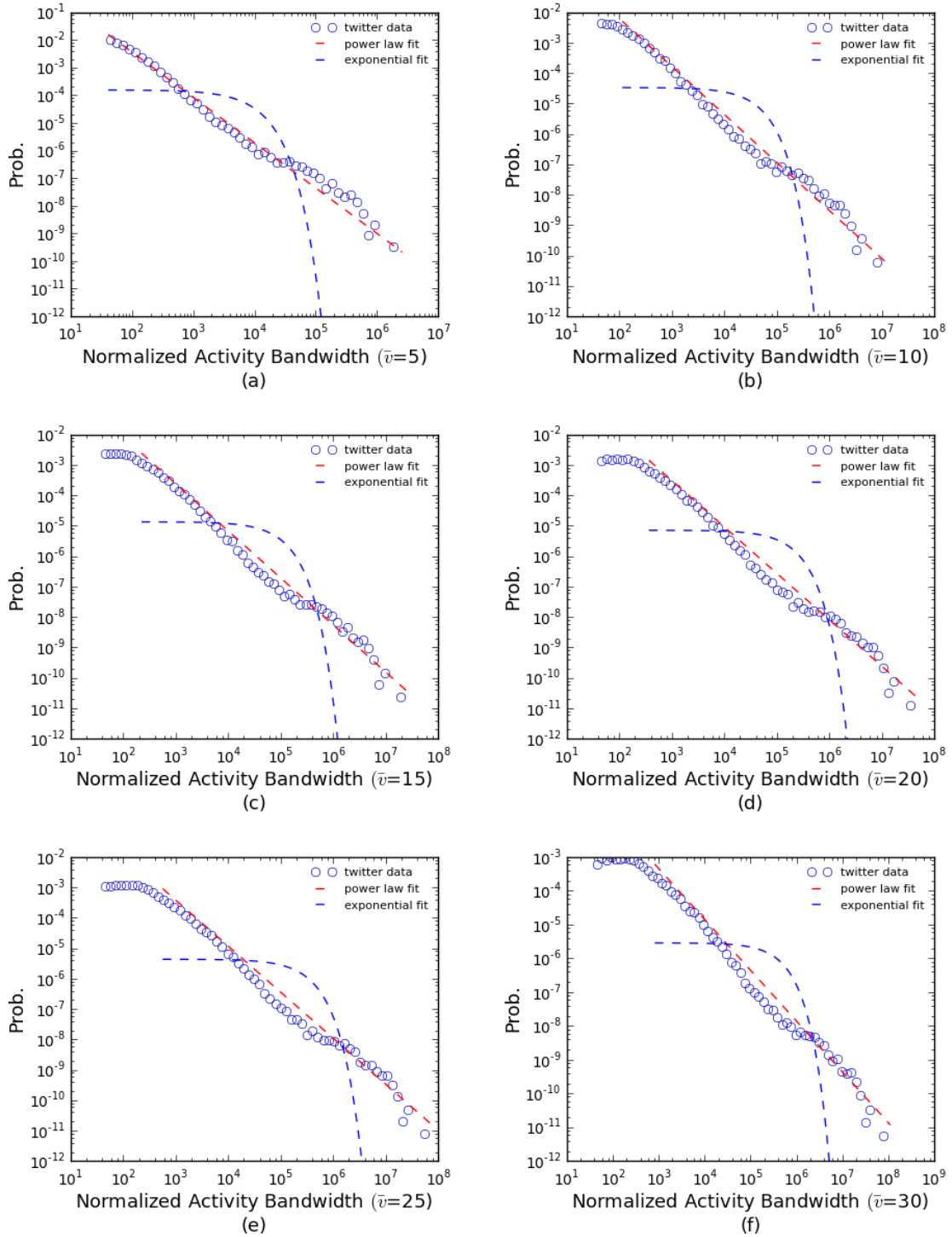
Figure 17 Power-law and exponential distribution fit for normalized activity bandwidth with different traveling speed limits (km/h). (a) $\bar{v} = 5$ (b) $\bar{v} = 10$ (c) $\bar{v} = 15$ (d) $\bar{v} = 20$ (e) $\bar{v} = 25$ (f) $\bar{v} = 30$.

Here, we only quantify the distributions of inaccuracy bounds, but identifying the best cutoff point is subjective to specifications (e.g. hourly versus monthly analysis) and inaccuracy tolerances of specific applications. With the gyration values calculated in equation (37), we calibrate the preceding activity bandwidths to the normalized activity bandwidths using equation (38). Figure 17 shows the distribution of normalized activity bandwidth. Again, by comparing the exponential and power-law distribution with the log likelihood ratio test we found that the normalized activity bandwidth is well approximated by a power-law distribution. Here, the increase of $\bar{v}$ from 5 km/h to 30 km/h will decrease the β from 1.64 to 1.50 that reveals less sensitivity to the speed limit. As before, the indication is that the majority of the tweet data have relatively large E_r in quantifying the mobility patterns of individuals, but there remain a large number of individuals around the head of the distribution that can provide relatively accurate mobility pattern information. For a given $\bar{v}$, the β value for the normalized activity bandwidth ($1.50 < \beta < 1.64$) is relatively smaller than that of the activity bandwidth ($1.68 < \beta < 2.93$), which implies that less data are useful for quantifying relative mobility patterns. However, as explained in section 4.2.2, we believe that normalized activity bandwidth is a better value for evaluating individuals' mobility patterns. The variation of NB with the average tweet density and the variance of tweet density is investigated in Figure 18. We can see that as the average number of tweets per day decreases and the variance increases, NB increases. Figure 18 could be useful in selecting a cutoff point for the minimum number of sample points required for mobility analysis for more specific applications.

Sensitivity analyses are conducted to draw insights on how the traveling speed limit ($\bar{v}$) as a key input parameter affect the activity bandwidth and normalized activity bandwidth. As shown in Figure 19, for both of these measures the mean and standard deviation will increase as we

increase the $\bar{v}$. In Figure 19(b) the standard deviation of normalized activity bandwidth will increase dramatically by increasing $\bar{v}$, which suggests a smaller $\bar{v}$ provides more rigorous results for mobility pattern analysis.
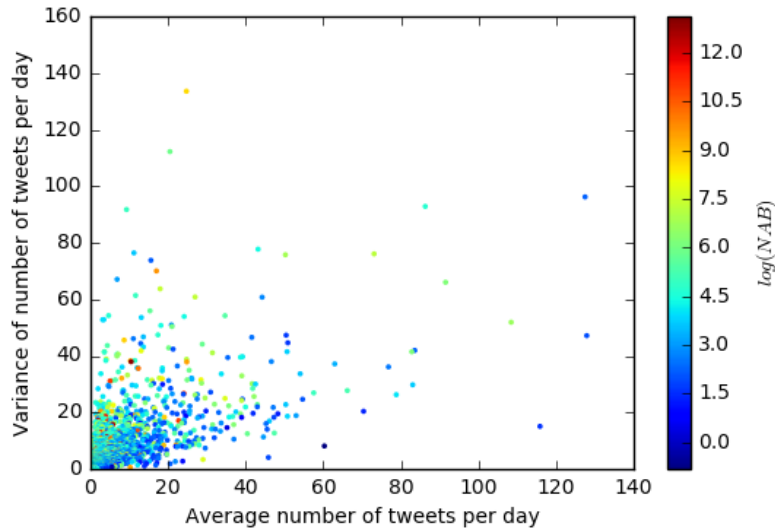
Figure 18 The variation of *NB* with average and variance of number of tweets per day.

(a)   (b)
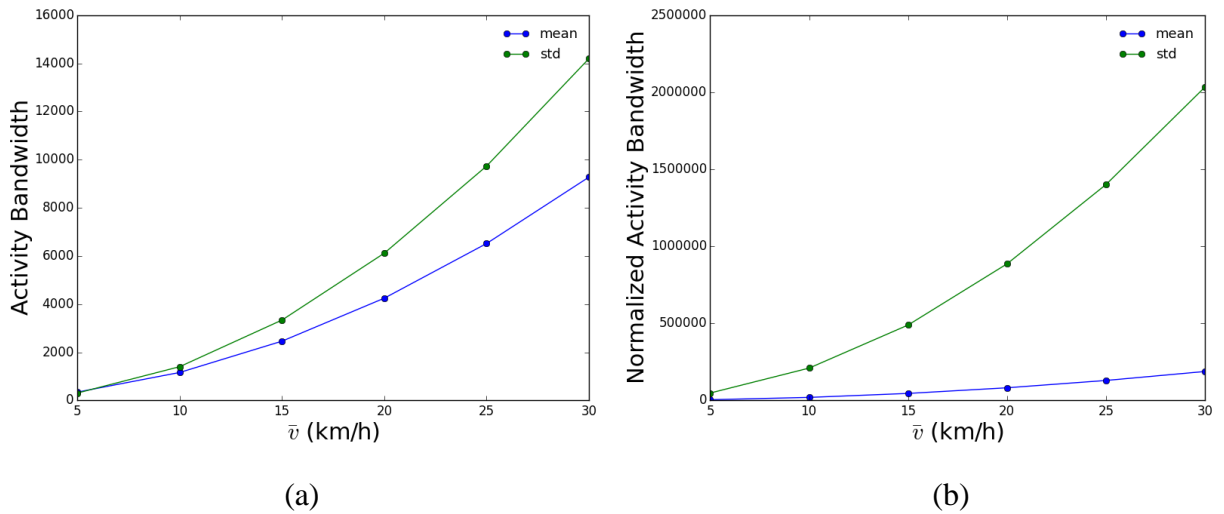
Figure 19 Sensitivity analyses with regard to traveling speed limit $\bar{v}$. (a) activity bandwidth (b) normalized activity bandwidth.

Finally, the application of the proposed methodology is illustrated for three cases with different time analysis intervals: case 1 for one day, case 2 for 16 hours and case 3 for 8 hours of

analysis. Different criteria and cutoff points are defined regarding the topology of an individual's mobility pattern and the location estimation accuracy in relation to the geographical region of the study area. From a topological point of view, if $NB < 0.1$ the activity pattern information is much higher than the estimation inaccuracy and the data is considered very good. If $0.1 \leq NB \leq 1$, the information is still higher than the inaccuracy and acceptable (good data). If $NB > 1$, the inaccuracy is greater than activity pattern information, and we recommend not to use such data for mobility pattern analysis.

Table 12 Percentage of data within the recommended cutoff points.

| Criteria	Case 1 (one day)	Case 2 (16 hours)	Case 3 (8 hours)
$NB \leq 0.1$	0.2	1.1	0.2
$0.1 \leq NB \leq 1$	0.6	0.2	1.6
$NB > 1$	99.2	98.7	98.2
$\dfrac{B(\mathbf{S})}{R_{NYC}} \leq 0.01$	0.02	0	0
$0.01 \leq \dfrac{B(\mathbf{S})}{R_{NYC}} \leq 0.1$	0.41	0.5	0.7
$\dfrac{B(\mathbf{S})}{R_{NYC}} > 0.1$	99.57	99.5	99.3

From the perspective of location estimation in relation to the study area, if $B(\mathrm{S})$ is significantly smaller than the radius of the studied area (denoted by $R_{NYC} = 16$ km for NYC with land area of 790 km^2), e.g. less than 1%, then the data is very useful for mobility analysis (very good data). If $B(\mathrm{S})$ is somehow smaller than the radius (e.g. greater than 0.01 but smaller than 0.1) the data is still considered useful for mobility analysis (good data), and all other data are not recommended (bad data). According to the results in Table 12, as we decrease the analysis time

interval (i.e. from one day to 8 hours) there are more data to be selected for mobility pattern analysis. For instance, in a one day analysis, 0.8 % of data are very good or good in case 1 based on the first criterion ($NB \leq 1$) compared with 1.8 % in case 3 when we decrease the time interval to 8 hours. The results are similar if we take the second criterion where 0.43 % very good or good data in case 1 increases to 0.7% in case 3. Considering the massive amount of geo-tagged data available in different platforms (e.g. social media), these findings suggest that there are many individuals with relatively accurate mobility information that could be targeted for future mobility pattern analysis. However, cautions are needed in screening these data for specific applications, given that the majority of such data contains overwhelming noise and may not be suitable for high-definition mobility pattern analysis.

4.3.2. Cellphone Data

The proposed measures are also applied to cellphone data collected in Shenzhen, China, for two days (Jan 14, 2014 and Jan 15, 2014) with 1,786,077 unique users and 18,485,979 geo-tagged sample points. The data consists of five fields including "User's ID", "Date" "Time", "Latitude" and "Longitude". A similar analysis is performed to the cellphone data and the results are compared with the Twitter data.

At first, to calculate $B(\mathbf{S})$, the lookup tables are prepared for the $U^c(t)$, $D^p(c,t)$ and $U^p(c,t)$. The lookup tables for cellphone data are alike Twitter data using the same parameters (see Table 8 to Table 10 for more details), except for ticks on c and t that are distributed using equation (36). Note that the variable's domain differs from Twitter data. For instance, cellphone data is collected for two days, therefore time ticks on t are distributed on 48 hours and the corresponding values in the lookup table are calculated accordingly.

65

Similar cutoff point analysis is presented in Table 13. The results show that 1.19% of data are very good or good in 8 hours analysis based on the first criterion ($NB \leq 1$), and 0.98% of data are very good or good considering the second criterion $\frac{B(\mathbf{S})}{R_S} < 0.1$ (Here $R_S = 25.54\ km$ is the average radius of Shenzhen with area of $2050\ km^2$).

Table 13 Percentage of cellphone data within the recommended cutoff points.

Criteria	Case 1 (one day)	Case 2 (16 hours)	Case 3 (8 hours)
$NB \leq 0.1$	0.06	0.06	0.07
$0.1 \leq NB \leq 1$	0.24	0.42	1.12
$NB > 1$	99.70	99.52	98.81
$\dfrac{B(\mathbf{S})}{R_S} \leq 0.01$	0.002	0.001	0.001
$0.01 \leq \dfrac{B(\mathbf{S})}{R_S} \leq 0.1$	0.094	0.103	0.978
$\dfrac{B(\mathbf{S})}{R_S} > 0.1$	99.904	99.896	99.021
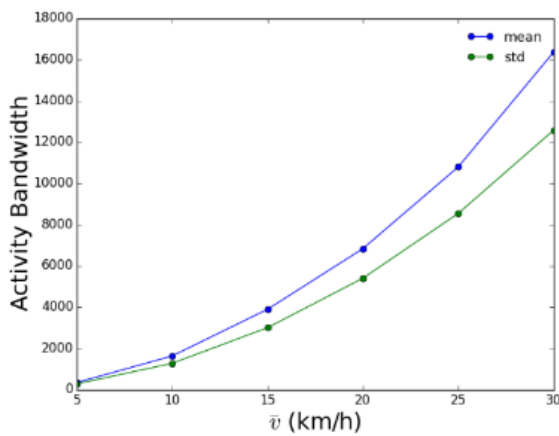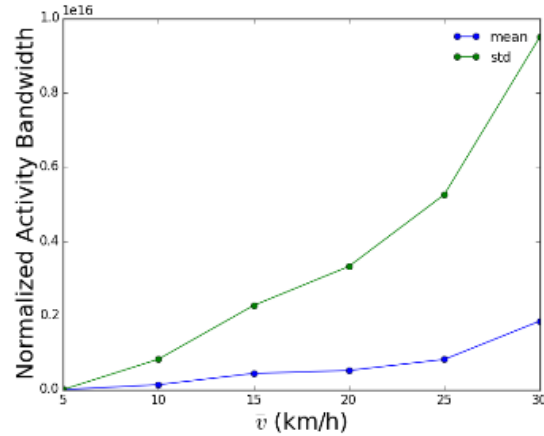
Figure 20 Sensitivity analyses on $\bar{v}$ for cellphone data. (a) activity bandwidth (b) normalized activity bandwidth.

Again, we can see for shorter analysis time (i.e. 8 hours) there are more data to be selected for mobility pattern analysis. Note that although the percentage of good data is slightly smaller

than Twitter data, the large number of cellphone users (1,786,077 individuals) means that more individuals can be selected for future mobility analysis. Figure 20 represents the sensitivity analysis results on effect of $\bar{v}$ on both $B(\mathbf{S})$ and $NB(\mathbf{S})$ and similar to the Twitter data, both mean and standard deviation increase by increasing $\bar{v}$.

Finally, Figure 21 and Figure 22 represent the distribution fit results for $B(\mathbf{S})$ and $NB(\mathbf{S})$. Unlike Twitter data, the distribution of $B(\mathbf{S})$ for all $\bar{v}$ values can be well fitted with an exponential distribution ($\lambda < 0.003$ for all $\bar{v}$). Note that exponential distribution has thinner tail than power-law distribution but still around the head of the distribution there are many users that can be selected for building accurate space-time trajectories. For $NB(\mathbf{S})$, again power-law distribution can be fitted to the data and the increase of $\bar{v}$ from 5 km/h to 30 km/h decrease the exponent (β) from 1.26 to 1.10 (less sensitivity to $\bar{v}$). This conveys consistent findings for both cellphone and Twitter data, where a small portion of individuals around the head of the distribution that can provide relatively accurate mobility pattern information.

4.4. Results and Discussion

This study proposes a set of parsimonious measures based on time geography concepts to answer an important question about mobility studies using geo-tagged mobility sample data: "How accurate it would be to utilize such samples in estimating continuous individual mobility trajectories?" In this study, the estimated trajectory between a set of limited space-time sample points is obtained by connecting these sample points with linear segments. However, since the estimated trajectory may differ from an individual's unknown ground truth trajectory, a set of fundamental measures is proposed to quantify the accuracy of the estimation inaccuracy in a robust manner. The estimation inaccuracy depends on the density of the sample points. In the proposed methodology, an individual's activity range around the estimated trajectory is constructed by a

Figure 21 Power-law and exponential distribution fit for activity bandwidth for cellphone data with different $\bar{v}$ (km/h). (a) $\bar{v} = 5$ (b) $\bar{v} = 10$ (c) $\bar{v} = 15$ (d) $\bar{v} = 20$ (e) $\bar{v} = 25$ (f) $\bar{v} = 30$.
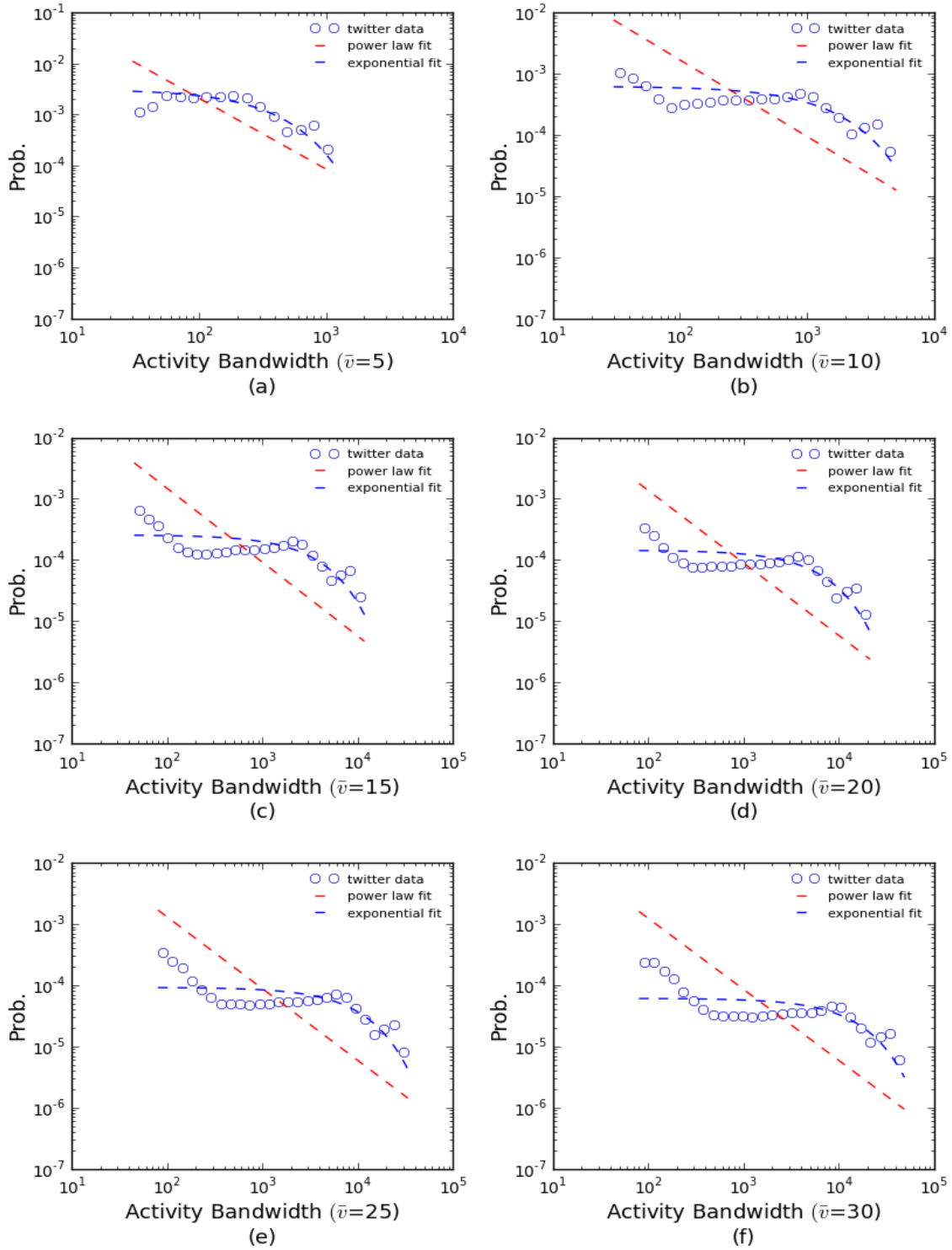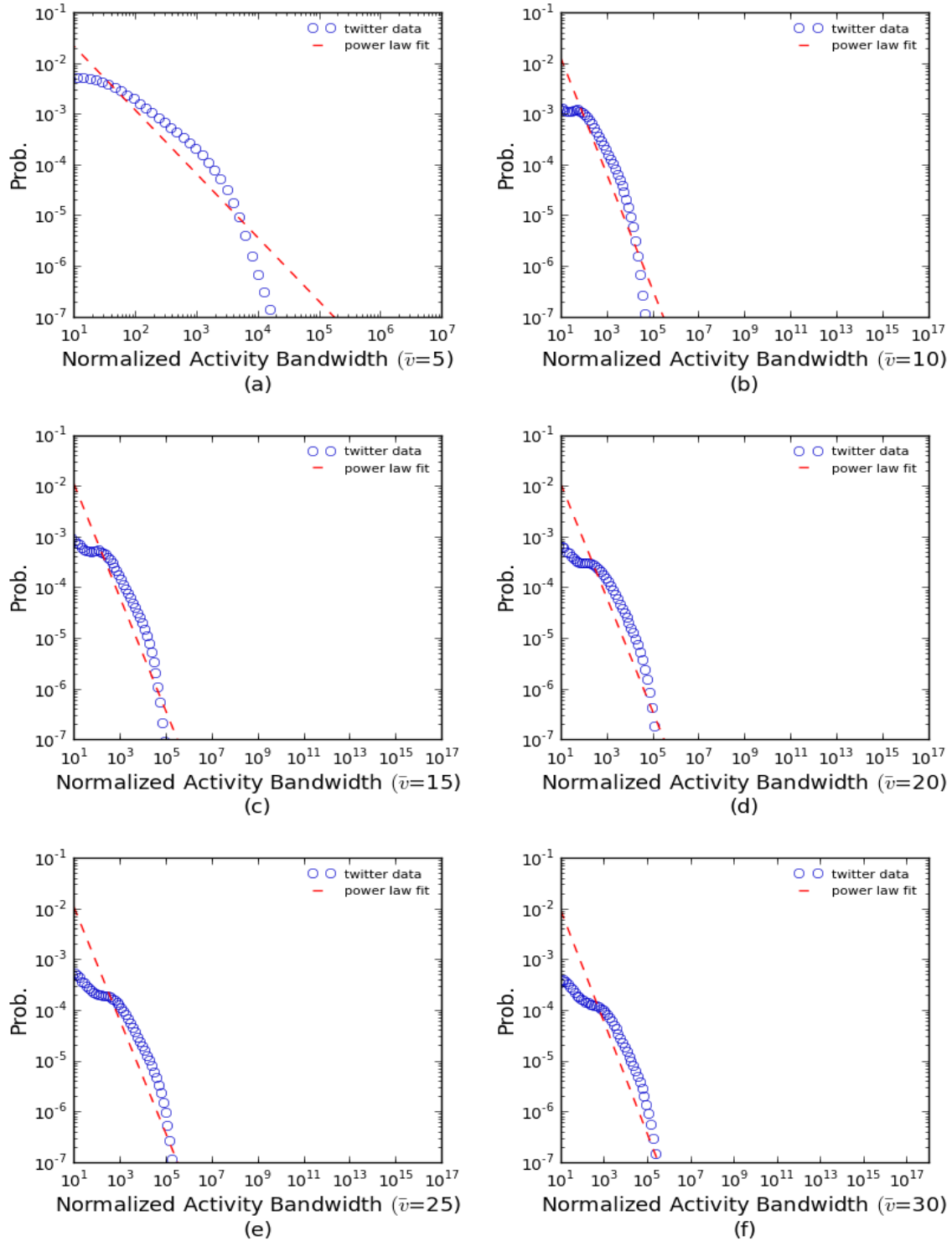
Figure 22 Power-law distribution fit for normalized activity bandwidth for cellphone data with different $\bar{v}$ (km/h). (a) $\bar{v} = 5$ (b) $\bar{v} = 10$ (c) $\bar{v} = 15$ (d) $\bar{v} = 20$ (e) $\bar{v} = 25$ (f) $\bar{v} = 30$.

69

chain of space-time prisms. Then the proposed measures, including activity bandwidth and normalized activity bandwidth, are calculated on this chain. The activity bandwidth quantifies the possible absolute inaccuracy range between the estimated and the ground truth trajectories, while the normalized activity bandwidth measures the relative difference between the mobility pattern of the estimated trajectory and that of the ground truth trajectory.

In travel mobility analysis, these measures can be used to evaluate the suitability of estimated individual trajectories from generic geo-tagged data. Since it is time consuming to calculate these measures working with massive mobility data, we also propose a lookup table–based interpolation method to expedite the calculation. The proposed measures and the associated lookup table method have been tested with two sets of real-world geo-tagged mobility data from social-media and cell phone logs. These cases studies demonstrate that the proposed measures can efficiently quantify inaccuracies of using sample individual mobility data in estimating their continuous trajectories. These case studies also draw a number of interesting managerial insights. For the Twitter data, we find that both measures proposed in this study follow power-law distributions at different traveling speed limits ($\bar{v}$). For the cellphone data, the normalized activity bandwidth again follows a power-law distribution, although the activity bandwidth measure can be better described with an exponential distribution. Sensitivity analyses are conducted to draw insights on how $\bar{v}$ can affect the proposed measures. Our findings show that most individuals in these data sets likely yield high estimation inaccuracies and may not be suitable for mobility studies with high accuracy requirement. However, because of the massive amount of geo-tagged data available, there are still a good number of individuals with relatively accurate mobility information for mobility pattern analysis. Nonetheless, cautions should be taken in screening these data for specific applications.

This study provides a methodological foundation for analyzing inaccuracy bounds of mobility measures for emerging geo-tagged mobility data, which can be extended in several directions. For specific applications, it will be interesting to use our proposed measures for investigating different cutoff points for the separation between useful data with less noise and valueless data with large noise based on the needs of these applications. When other geo-tagged data sources are available, it is useful to apply the proposed methodology to these data sets to draw implications of their mobility patterns. While the measured geo-coordinates have significant inaccuracies, this proposed methodology needs to be properly calibrated to account for such sampling data noise.

CHAPTER 5: NETWORK DESIGN PROBLEM – A SPACE-TIME PROBABILISTIC FRAMEWORK

5.1. Introduction

Sensor network design highly depends on the application considered e.g. network surveillance, traffic, pollution, agriculture or infrastructure monitoring [143]. Usually, sensors are designed to observe the environment and to collect data from target population to make a better decision. The observation could help us to detect specific events or track individual targets and interact with them to transfer valuable information. For instance, in advertisement industry, billboards are known as the most common outdoor advertising strategy in cities. Geographical locations of billboards can be predetermined if the target population and the environment are well known. Otherwise, caution may be exercised to determine the best locations for billboards. In transportation, traffic management centers and transportation planning organizations deploy traffic sensors at various places across a network to observe the traffic conditions and electronic digital signs are distributed in the network to inform drivers about real time traffic conditions and provide route guidance. Due to limited resources and investment constraints, smartly deploying surveillance sensors is essential to maximize the network coverage and increase the number of targets that are detected through the region,

In this chapter, a space-time network-based modeling framework is proposed to integrate human mobility into network design problem. A probabilistic network structure is constructed to quantify human's presence probability in a space-time framework. Within the space-time network, the presence probability of individual targets can be quantified in space-time nodes and links.

Then, we propose a Mixed Integer Nonlinear Programming (MINLP) model to maximize the spatial and temporal coverage of target population. To achieve near optimal solutions for large-scale networks, the greedy heuristic, lagrangian relaxation and simulated annealing algorithms are proposed to solve this problem.

5.2. Space-Time Network Structure - A Probabilistic Framework

Following the motivations on CHAPTER 4, in order to construct a space-time network, we can augment the physical network into multiple layers for different time steps. For instance, if we divide the time horizon into a series of intervals with the same duration of σ, we can reconstruct the space-time network as shown in Figure 23 [56]. Note that the network reconstruction is challenging for large scale geographical spaces and small values of σ. This reconstruction process is adapted from [126] that is explained in CHAPTER 3.
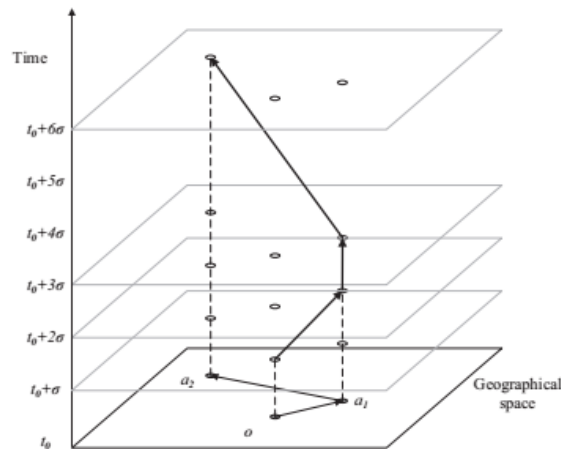
Figure 23 Reconstruction of a physical network into a space-time framework [56].

The advantages of the network reconstruction are that we can characterize different attributes of moving targets (e.g. presence probability) considering space-time nodes and links. For ease of explanation, let consider a one-dimensional space network with nodes j_1 to j_5 as shown in Figure 24(a). Then, the reconstruction process starts with duplication of space nodes j_1 to j_5 in different time series $t_1, t_2, \dots, T$.
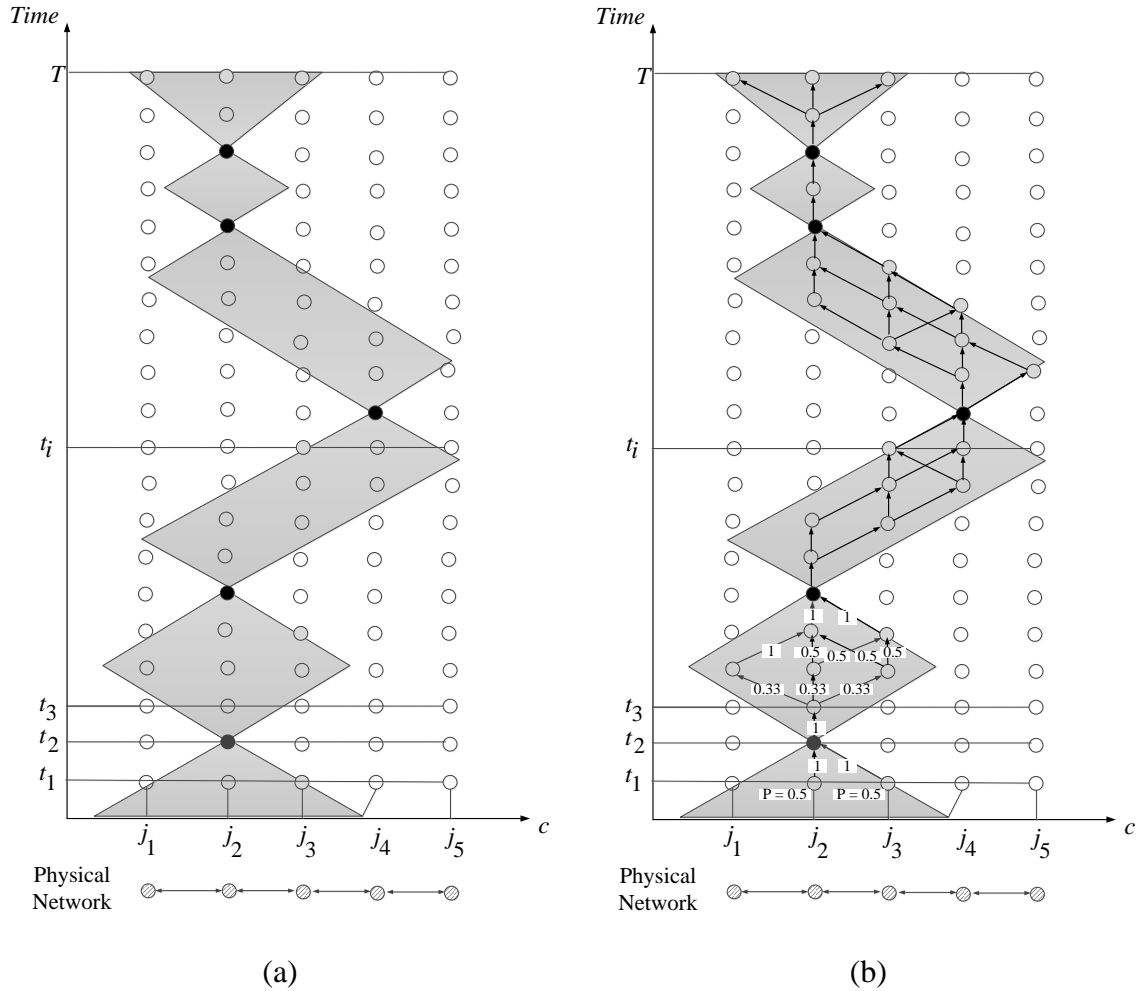
Figure 24 Illustration of probabilistic space-time network structure. a) prism chain in one dimensional space b) space-time node and link probabilities

Using the time geography concepts and trajectory analysis explained in CHAPTER 3, now we can consider the movement boundary of individuals based on their known space-time sample points (black points in Figure 24), and quantify their presence probability in various locations by constructing their prism chains. The space-time links are added next considering the node connections in the physical network Figure 24(b). Then, for a single target, the probabilities of all the space-time node and links can be easily quantified as shown in Figure 24(b). For instance, according to the prism chain in Figure 24(a), the target cannot be present in nodes j_1, j_4 and j_5 at time t_1 and its presence probability of being at either node j_2 or j_3 is 0.5. This probabilistic network

structure could be used for all the targets and their prism chains can be constructed based on their known space-time sample points. Then, by aggregating all the individuals we can identify the spots that are more likely to be visited by different targets to maximize the network coverage.

5.3. Model Formulation

This section proposes a probabilistic framework to integrate geo-tagged location data generated by targets into sensor location design problem. Table 14 provides a summary of the notation used in the formulations. We deploy sensors in transportation network to maximize the expected benefit from covering targets while having a budget constraint. Let H be the set of targets, T be the set of time stamps and $G(N, A)$ be the transportation network where N and A represent the set of space nodes and links, respectively. Also, let u_{hit} be the benefit (utility) of covering target $h \in H$ at node $i \in N$ and time $t \in T$ (If the targets are homogenous and the benefits of covering a target in different space-time points are identical, then $u_{hit} = u$). We can formulate the location problem by using a number of known space-time point (geo-tagged data) that are visited by each target $h \in H$. By implying time-geography concepts on these known space-time points, the probability of visiting all other space-time points can be calculated for all the targets. We denote this space-time point probability by $p_{hit}, \forall h \in H, i \in N, t \in T$.

For all the physical links $(j, i) \in A$, let (j, i, s, t) be the index of space-time link that indicates a movement from node j to node i ($j, i \in N$), and from time s to time t ($s, t \in T$ and $s < t$). Note that targets are moving in a space-time network and we need to distinguish between a space node (link) and a space-time point (link). The (j, i, s, t) index can be used to define p_{hjist}, which represents the probability of visiting a space-time link (j, i, s, t) by target $h \in H$. For simplicity, we assume $s = t - 1$ that clarifies the space-time links are defined for two consecutive time stamps. Finally, we denote the installation cost of deploying a new sensor at node $i \in N$ by

f_i. It is assumed that sensors are fixed after installation and they cannot be moved over time. With this assumption, we can define the primal decision variable $y := \{y_i\}$ to determine where to install sensors:

Table 14 Sample transportation networks.

H	Set of targets
T	Set of time stamps
N	Set of space-nodes
A	Set of space links
u_{hit}	Benefit (utility) of covering target $h \in H$ at node $i \in N$ and time $t \in T$
p_{hit}	Probability of visiting node $i \in N$ at time $t \in T$ by target $h \in H$
j, i	Indices of space-nodes, $j, i \in N$
s, t	Indices of time stamps, $s, t \in T$
(j, i, s, t)	Index of space-time link that indicates a movement from node j to node i and from time s to time t, where $(j, i) \in A$ and $s < t$
p_{hjist}	Probability of visiting space-time link (j, i, s, t) by target $h \in H$
f_i	Installation cost of deploying a new sensor at node $i \in N$
y_i	1 if a sensor (facility) is installed at node $i \in N$, 0 otherwise
q_{hit}	Probability of target h to visit node i at time t, without being intercepted by any sensor before time t, $\forall h \in H, i \in N, t \in T$
v_{hit}	Auxiliary variable ($v_{hit} = q_{hit}y_i$) defined for linearization, $\forall h \in H, i \in N, t \in T$
B	Total budget (number of available sensors)

$$y_i = \begin{cases} 1 & if\ a\ sensor\ is\ installed\ at\ location\ i \in N \\ 0 & otherwise \end{cases} \tag{40}$$

Also, we define q_{hit} the probability of target h to visit node i at time t, without being intercepted by any sensor before time t, $\forall h \in H, i \in N, t \in T$. Due to limited budget, no more than B sensors can be built on the network. Eventually, the space-time sensor location problem (STS) can be formulated as follows:

$$(STS)\ max\ Z(y_i, q_{hit}) := \sum_{h \in H} \sum_{i \in N} \sum_{t \in T} u_{hit} q_{hit} y_i \tag{41}$$

$$s.t.\ q_{hi1} = p_{hi1}, \forall h \in H, i \in N \tag{42}$$

$$q_{hit} = \sum_{(j,i)\in A} q_{hj(t-1)} p_{hji(t-1)t}(1-y_j), \forall h \in H, i \in N, t \in T\backslash\{1\} \tag{43}$$

$$\sum_{i\in N} f_i y_i \leq B \tag{44}$$

$$y_i \in \{0,1\}, \forall i \in N \tag{45}$$

$$q_{hit} \geq 0, \forall h \in H, i \in N, t \in T \tag{46}$$

The objectives function (41) maximizes the utility of covering targets in a spatial-temporal framework. Constraint (42) ensures q_{hi1} is equal to p_{hi1} at $t = 1$, where p_{hi1} is the given probability of visiting node i by individual h using time-geography (since we are in the beginning of the time ($t = 1$), we assume targets are not intercepted yet and $q_{hi1} = p_{hi1}$). According to the constraint (43), when $t > 1$, the q_{hit} is calculated for each target based on the historical information we know about her coverage at his parent nodes. A space-time point $(j, t - 1)$ is predecessor of space-time point (i, t) where $\exists (j, i) \in A$ (again note that space-time links are defined for two consecutive time stamps $t - 1$ and t). Constraint (44) enforces the budget limit and constraint (45) and (46) postulate binary and continuous decision variables. Because of the nonlinear terms $q_{hit}y_i$ in (STS) formulation, the solution time complexity is NP-hard. Therefore, we try to linearize (STS) by defining an auxiliary variable $v_{hit} = q_{hit}y_i$. Using v_{hit} and without affecting the optimal solution, constraints (47) to (50) can be added to (STS) to ensure that $v_{hit} = q_{hit}$ when $y_i = 1$, and $v_{hit} = 0$ when $y_i = 0$ (in another word, $v_{hit} = q_{hit}y_i$ is equivalent to $v_{hit} = \min(q_{hit}, y_i)$):

$$v_{hit} \leq q_{hit}, \forall h \in H, i \in N, t \in T \tag{47}$$

$$v_{hit} \leq y_i, \forall h \in H, i \in N, t \in T \tag{48}$$

$$v_{hit} \geq q_{hit} + (y_i - 1), \forall h \in H, i \in N, t \in T \tag{49}$$

$$v_{hit} \geq 0, \forall h \in H, i \in N, t \in T \tag{50}$$

Therefore, the linearized space-time sensor location problem (LSTS) can be written as:

$$(LSTS): max\ Z(y_i, q_{hit}, v_{hit}) := max \sum_{h \in H} \sum_{i \in N} \sum_{t \in T} u_h v_{hit} \tag{51}$$

$$s.t.\ q_{hi1} = p_{hi1}, \forall h \in H, i \in N \tag{52}$$

$$q_{hit} = \sum_{(j,i) \in A} p_{hji(t-1)t}(q_{hj(t-1)} - v_{hj(t-1)}), \forall h \in H, i \in N, t \in T \backslash \{1\} \tag{53}$$

$$v_{hit} \le q_{hit}, \forall\ h \in H, i \in N, t \in T \tag{54}$$

$$v_{hit} \le y_i, \forall\ h \in H, i \in N, t \in T \tag{55}$$

$$v_{hit} \ge q_{hit} + (y_i - 1), \forall\ h \in H, i \in N, t \in T \tag{56}$$

$$\sum_{i \in N} f_i y_i \le B \tag{57}$$

$$y_i \in \{0,1\}, \forall\ i \in N \tag{58}$$

$$q_{hit}, v_{hit} \ge 0, \forall h \in H, i \in N, t \in T \tag{59}$$

5.4. Solution Algorithm

The linearized space-time sensor location problem (LSTS) can be potentially solved by commercial solvers such as CPLEX or Gurobi for small-scale instances. However, as we increase the network size (see section 5.5 for more discussion), the computational burden is greatly exacerbated and it takes a large amount of computation time for the solvers to obtain even a feasible solution. Therefore, we develop heuristic algorithms to obtain near-optimum solutions in a reasonable amount of time. The first algorithm is based on a simple greedy heuristic, which can yield good solutions for many realistic applications. Since the greedy heuristic does not provide information on how close the solutions are from the true optima, Lagrangian Relaxation (LR) and Simulated Annealing (SA) are also implemented. LR provides not only good feasible solutions but also it provides the optimality gap.

5.4.1. Greedy Heuristic

The greedy heuristic is widely applied to many practical problems because of its simplicity and reasonable practical performance. The greedy algorithm for (STS) selects sensors sequentially based on the best marginal increase of $Z(y_i, q_{hit})$ until all the budget is exhausted. The exact steps are as follows.

Step 0: Initialization: Let $\Omega = \emptyset$ be the set of selected space nodes for sensor installation and $F = 0$ be the total installation cost. Set $y_i = 0, \forall i \in N$ and. $q_{hit} = 0, \forall h \in H, i \in N, t \in T \setminus \{1\}$.

Step 1: Search for the location that brings the maximum marginal improvement of objective (51); i.e., select

$$i^* = \underbrace{argmax}_{k \in N \setminus \Omega} \{Z(\mathbf{y}', \mathbf{q}') - Z(\mathbf{y}, \mathbf{q}) : y_i' = 1 \; iff \; i \in \Omega \cup \{k\}\}.$$

where $Z(\mathbf{y}', \mathbf{q}') - Z(\mathbf{y}, \mathbf{q})$ represents the marginal improvement (note that given $\mathbf{y}$, both $\mathbf{q}$ and $Z(\mathbf{y}, \mathbf{q})$ can be calculated.).

Step 2: Let $y_{i^*} = 1$, $F = F + f_{i^*}$ and $\Omega = \Omega \cup i^*$.

Step 3: If adding an additional sensor is possible according to the remaining budget, i.e. $\exists i \in N \setminus \Omega : f_i < B - F$, go to step 1; otherwise stop and return $\mathbf{y}$ and the corresponding $\mathbf{q}$ and $Z(\mathbf{y}, \mathbf{q})$.

Given $\mathbf{y}$ variable, both $\mathbf{q}$ and $Z(\mathbf{y}, \mathbf{q})$ in (STS) can be calculated. Therefore, we can assume Z is a monotone set function defined over subsets of $\mathbf{y}$ as a ground set (It is monotone since adding a new sensor does not decrease the objective value). Also, the problem of choosing sensor locations from a given set of candidate location is submodular [144]. Therefore, we can prove that the objective value of any greedy solution is no smaller than $(1 - \frac{1}{e})$ of the true optimum.

Definition 1: The Z function is monotone iff:

$$\forall S \subseteq T \subseteq N, Z(S) \leq Z(T) \tag{60}$$

Corollary 1. Let Z be a submodular function defined over subsets of a ground set N. Then,

$$\forall S \subseteq T \subseteq N, Z(T) \leq Z(S) + \sum_{e \in T \setminus S} Z(S \cup \{e\}) - Z(S) \tag{61}$$

Proposition 6: In the STS problem, the objective value of the greedy algorithm solution is no smaller than $(1 - 1/e)$ of the true optimum.

Proof. Since given $\mathbf{y}$ variable, both $\mathbf{q}$ and $Z(\mathbf{y}, \mathbf{q})$ can be calculated, we ignore the $\mathbf{q}$ for ease of explanation. Let $\mathbf{y}^*$ be the optimal solution and $Z(\mathbf{y}^*)$ be the true optimum and $|\mathbf{y}^*| \leq K$ due to the budget constraint. Also, let $\mathbf{y_i}$ be the set of selected locations for sensor installation returned by greedy algorithm after i iteration where $\mathbf{y_i} = \{y_1, y_2, \dots, y_i\}$. Then we have:

$$
\begin{aligned}
Z(\mathbf{y}^*) &\leq Z(\mathbf{y_{i-1}}) + \sum_{y \in \mathbf{y}^* \setminus \mathbf{y}_{i-1}} Z(\mathbf{y_{i-1}} \cup \{y\}) - Z(\mathbf{y_{i-1}}) \\
&\leq Z(\mathbf{y_{i-1}}) + \sum_{y \in \mathbf{y}^* \setminus \mathbf{y}_{i-1}} Z(\mathbf{y_i}) - Z(\mathbf{y_{i-1}}) \\
&\leq Z(\mathbf{y_{i-1}}) + K\big(Z(\mathbf{y_i}) - Z(\mathbf{y_{i-1}})\big) \tag{62}
\end{aligned}
$$

The first inequality used Corollary 1. The second inequality is correct because of the greediness of algorithm. Moving from iteration $i - 1$ to iteration i, greedy will choose a node with best marginal contribution to Z, therefor, $\sum_{y \in \mathbf{y}^* \setminus \mathbf{y_{i-1}}} Z(\mathbf{y_{i-1}} \cup \{y\}) - Z(\mathbf{y_{i-1}}) \leq \sum_{y \in \mathbf{y}^* \setminus \mathbf{y_{i-1}}} Z(\mathbf{y_i}) - Z(\mathbf{y_{i-1}})$, therefore the second inequality is correct. The last inequality is correct because $|\mathbf{y}^*| \leq K$ and at most set of $\mathbf{y}^* \setminus \mathbf{y_{i-1}}$ has K members.

Subtracting $K * Z(\mathbf{y}^*)$ from both sides gives $Z(\mathbf{y_i}) - Z(\mathbf{y}^*) \geq (1 - \frac{1}{K})(Z(\mathbf{y_{i-1}}) - Z(\mathbf{y}^*))$ which implies by induction:

$$Z(\mathbf{y_i}) \geq \left(1 - \left(1 - \frac{1}{K}\right)^i\right) Z(\mathbf{y}^*)$$

Note that if $a_i = Z(\mathbf{y_i}) - Z(\mathbf{y}^*)$, then we have $a_i \leq \left(1 - \frac{1}{K}\right) a_{i-1} \leq \left(1 - \frac{1}{K}\right)^2 a_{i-2} \leq \ldots \leq$

$\left(1 - \frac{1}{K}\right)^i a_0$. If $i = K$ and using $\left(1 - \frac{1}{K}\right)^K \leq 1/e$, we have $Z(\mathbf{y_i}) \geq \left(1 - \frac{1}{e}\right) Z(\mathbf{y}^*)$ which

concludes the proof. □

5.4.2. Lagrangian Relaxation

As we discuss in Section 5.5, commercial optimization solvers (e.g., CPLEX) run into difficulty when problem size increases. In this section, we present an algorithm based on Lagrangian Relaxation (LR) and decompose the (LSTS) into a set of sub-problems that are easy to solve. The LR heuristic provides a feasible solution and an upper bound for the maximization problem to evaluate the optimality gap.

5.4.2.1. Upper Bound

We relax constraints (55) and (56) of (LSTS) formulation and add them to the objective (51) with nonnegative Lagrangian multipliers λ and γ. Then, the relaxed (LSTS) becomes:

$$(\text{RLSTS}): \underbrace{\min_{\lambda,\gamma}}_{} Z_R(\lambda,\gamma) := \underbrace{\max_{y,v,q}}_{} \sum_{h\in H}\sum_{i\in N}\sum_{t\in T} u_h v_{hit} + \sum_{h\in H}\sum_{i\in N}\sum_{t\in T} \lambda_{hit}(y_i - v_{hit}) \tag{63}$$
$$+ \sum_{h\in H}\sum_{i\in N}\sum_{t\in T} \gamma_{hit}(v_{hit} - q_{hit} - y_i + 1)$$

$$s.t. \ q_{hi1} = p_{hi1}, \forall h \in H, i \in N \tag{64}$$

$$q_{hit} = \sum_{(j,i)\in A} p_{hji(t-1)t}(q_{hj(t-1)} - v_{hj(t-1)}), \forall h \in H, i \in N, t \in T\backslash\{1\} \tag{65}$$

$$v_{hit} \leq q_{hit}, \forall h \in H, i \in N, t \in T \tag{66}$$

$$\sum_{i\in N} f_i y_i \leq B \tag{67}$$

$$y_i \in \{0,1\}, \forall i \in N \tag{68}$$

$$q_{hit}, v_{hit} \geq 0, \forall h \in H, i \in N, t \in T \tag{69}$$

For any given λ and γ, the exact value of $Z_R(\lambda, \gamma)$ provides an upper bound on the objective function (51). The (RLSTS) formulation can be decomposed into two sub-problems on targets (h) and space nodes (i). For sub-problem 1, we set $\lambda_i := [\lambda_{hit} \geq 0]_{h \in H, t \in T}, \gamma_i := [\gamma_{hit} \geq 0]_{h \in H, t \in T}, \forall i \in N$, then decompose $Z_R(\lambda, \gamma)$ as follows:

$$Z_{R1}(\lambda_i, \gamma_i) := \underbrace{max}_{y_i} \left(\sum_{h \in H} \sum_{t \in T} (\lambda_{hit} - \gamma_{hit}) \right) y_i, \forall i \in N \tag{70}$$

$$s.t. \quad \sum_{i \in N} f_i y_i \leq B \tag{71}$$

$$y_i \in \{0,1\}, \forall\, i \in N \tag{72}$$

For sub-problem 2, we set $\lambda_h := [\lambda_{hit} \geq 0]_{i \in N, t \in T}, \gamma_h := [\gamma_{hit} \geq 0]_{i \in N, t \in T}, \forall h \in H$, then:

$$Z_{R2}(\lambda_h, \gamma_h) := \underbrace{max}_{q,v} \sum_{i \in N} \sum_{t \in T} \left((u_h - \lambda_{hit} + \gamma_{hit}) v_{hit} + \gamma_{hit} (1 - q_{hit}) \right), \forall h \in H \tag{73}$$

$$s.t. \quad q_{hi1} = p_{hi1}, \forall\, i \in N \tag{74}$$

$$q_{hit} = \sum_{(j,i) \in A} p_{hji(t-1)t} (q_{hj(t-1)} - v_{hj(t-1)}), \forall\, i \in N, t \in T \backslash \{1\} \tag{75}$$

$$v_{hit} \leq q_{hit}, \forall\, i \in N, t \in T \tag{76}$$

$$q_{hit}, v_{hit} \geq 0, \forall\, i \in N, t \in T \tag{77}$$

Note that $Z_R(\lambda, \gamma) = Z_{R1}(\lambda_i, \gamma_i) + Z_{R2}(\lambda_h, \gamma_h)$. After decomposition, sub-problem 1 becomes a knapsack problem, which can be solved practically using efficient knapsack algorithms such as dynamic programming. Also, for sub-problem 2, the LP model can be solved for each target h, and the results can be aggregated for all the targets.

5.4.2.2. Lower Bound

By solving (RLSTS) at each iteration, we have all the decision variables y_i, q_{hit} and v_{hit} and the corresponding upper bound value. Considering the original problem, y_i satisfies all the

82

constraints (i.e. budget and integrality constraint). However, q_{hit} and v_{hit} are probably not feasible to the original (LSTS) problem and heuristic methods are used to construct a feasible solution. At first, we fix y_i to the values obtained from (RLSTS), then update q_{hit} and v_{hit} $\forall h \in H, i \in N$ and $t \in T \backslash \{1\}$ accordingly. Note that since q_{hi1} is known from constraint (52), for given y_i, both q_{hit} and v_{hit} can be obtained from constraints (53) to (56). Finally, deploying all the decision variables into the objective function (51) provides a lower bound at each iteration. If the lower and upper bounds are equal at an iteration, then the optimal solution is found and algorithm is terminated. Otherwise, the difference between the upper and lower bounds provides the optimality gap (which means the difference between the true optimum and the feasible solution is no larger than this gap). The optimality gap could be used as a stopping criterion to terminate the algorithm. Note that if the Lagrangian multipliers are near convergence, the obtained lower bound yields a good feasible solution of the original objective with small optimality gap.

5.4.2.3. Updating Lagrangian Multiplier

Lagrangian Relaxation is an iterative algorithm and we aim to minimize the upper bound while updating λ and γ at each iteration $(\underbrace{min}_{\lambda, \gamma} Z_R(\lambda, \gamma))$. Since the $Z_R(\lambda, \gamma)$ is convex over lagrangian multipliers, we can use a subgradient optimization method and update the multiplier as follows:

$$\lambda_{hit}^{k+1} = max(0, \lambda_{hit}^k + t^k \Delta \lambda_{hit}^k), \forall\, h \in H, i \in N, t \in T \tag{78}$$
$$\gamma_{hit}^{k+1} = max(0, \gamma_{hit}^k + t^k \Delta \gamma_{hit}^k), \forall\, h \in H, i \in N, t \in T \tag{79}$$

where the subgradients are $\Delta \lambda_{hit}^k = v_{hit} - y_i$ and $\Delta \gamma_{hit}^k = q_{hit} + y_i - 1 - v_{hit}$. Note that the max operator is used to take care of the negative outcomes while updating multipliers. The step size is usually set to

$$t^k = \frac{\mu^k(Z_R(\lambda^k, \gamma^k) - Z^{LB})}{\sum_{h \in H} \sum_{i \in N} \sum_{t \in T} \left[\left(\Delta\lambda_{hit}^k \right)^2 + \left(\Delta\lambda_{hit}^k \right)^2 \right]} \tag{80}$$

where Z^{LB} is the objective value of the best-known feasible solution and $Z_R(\lambda^k, \gamma^k)$ is the upper bound at k^{th} iteration and μ^k is the control scalar. In practice, the scalar parameter is set to a value between 0 and 2 in the first iteration ($0 < \mu^0 < 2$) ant it will be updated by comparing the best and worst values of $Z_R(\lambda^k, \gamma^k)$ in every certain number of iterations (e.g., 20). In section 5.5, we decrease μ^k ($k > 0$) if the difference between best and worst values is greater than 1%, and increase μ^k if the difference is less than a 0.1%.

The LR algorithm is terminated if one of the following conditions is satisfied: (i) the optimality gap is less than a threshold (e.g. 0.0001), (ii) the solution time exceeds a reasonable limit (or a certain number of iterations). The LR steps could be summarized as follows:

Step 0: Initialization: Let K be the maximum number of iterations (e.g. $K = 1000$), $k = 0$ be the iteration index and ϵ be the acceptable optimality gap. Initialize λ_{hit}^k and γ_{hit}^k, for all $h \in H, i \in N, t \in T$ and let y^{LB}, q^{LB} and Z^{LB} be the best feasible solution of STS obtained by greedy algorithm and $Z^{UB} = Inf$.

Step 1: Upper Bound: Solve sub-problem 1 using dynamic programing and store $Z_{R1}^k(\lambda_i, \gamma_i)$, and y^k. Also, solve sub-problem 2 using CPLEX and store $Z_{R2}^k(\lambda_h, \gamma_h)$, q^k and v^k. Then, the upper bound objective value is $Z_R^k(\lambda, \gamma) = Z_{R1}^k(\lambda_i, \gamma_i) + Z_{R2}^k(\lambda_h, \gamma_h)$.

Step 2: Lower Bound: using y^k from step 1, calculate the corresponding q^D, v^D and Z^D for the STS.

Step 3: If $Z^D > Z^{LB}$, update $Z^{LB} = Z^D$, $q^{LB} = q^D$, $v^{LB} = v^D$ and $y^{LB} = y^k$. If $Z^k < Z^{UB}$, then $Z^{UB} = Z^k$. Also, update the optimality gap as $opt\_gap = \frac{Z^{UB} - Z^{LB}}{Z^{LB}}$

Step 4: If $k = K$ or $opt\_gap < \epsilon$, stop and return *opt-gap*, y^{LB}, q^{LB} and Z^{LB}; otherwise, go to step 5.

Step 5: Update lagrangian multipliers at iteration k as follows. Then, $k = k + 1$ and go to step 1.

5.4.3. Simulated Annealing

The simulate annealing (SA) algorithm is used for approximating the optimal value (finding a near optimal solution) and comparing the results with LR and the exact solutions. As explained in section 3.4.2, SA is a random neighborhood search technique improves the objective value iteratively. As an important advantage, SA can escape from local optima by accepting a worse solution with a probability function that decreases in the higher iterations. In each iteration, the algorithm compares the new objective value $Z^{new}(.)$, with the current objective $Z^{cur}(.)$ obtained from previous iterations. In a maximization problem, if $Z^{new}(.) > Z^{cur}(.)$, then $Z^{cur}(.) = Z^{new}(.)$, which means the new solution is accepted, otherwise the new solution which is worse than the current solution ($Z^{new}(.) < Z^{cur}(.)$) will be accepted with the probability of $p = \exp\left(\frac{Z^{new}(.) - Z^{cur}(.)}{T}\right)$ where T is temperature parameter. Please see section 3.4.2 for more discussion on SA and the pseudo code of the algorithm is presented in Figure 4.

For initialization, the algorithm starts with the solution obtained from Greedy algorithm. To find a new solution by the neighborhood search, remove one of the current locations i with an installed sensor (set $y_i = 0$). Then, search all the first level neighbors j (first level neighbors are connected to i with a space-link, i.e. $(i,j) \in A$) and add them to the new solution (set $y_j = 1$) as long as the budget constraint is satisfied. If none of the first level neighbors of node i can be selected due to the budget constraint, set $y_i = 1$ with 0.5 probability, and repeat the neighborhood search for another location in the current solution that has a sensor installed. If all the locations are investigated and still a new solution has not been found, the neighborhood search is extended to a

second level neighbors (second level neighbors are indirectly connected to i where there is one node in between them.). This process continues by extending the neighborhood search to third level neighbors until a new solution can be found. Again, note that for the new locations selected for installation $\boldsymbol{y}$, both $\mathbf{q}$ and $Z^{new}(\mathbf{y}, \mathbf{q})$ can be calculated.

5.5. Case Studies

The proposed mathematical model and the solution approaches are applied to different numerical examples to demonstrate their applicability and performance. In the first example, a series of hypothetical space-time grid networks of various sizes are constructed and the solution results are compared. The hypothetical networks are generated in a one-dimensional space as shown in Figure 25 and the space-time sample points for each target are randomly generated. The second numerical example, on the other hand, is based on real geo-tagged twitter data collected in NYC (see section 4.3 for more details).
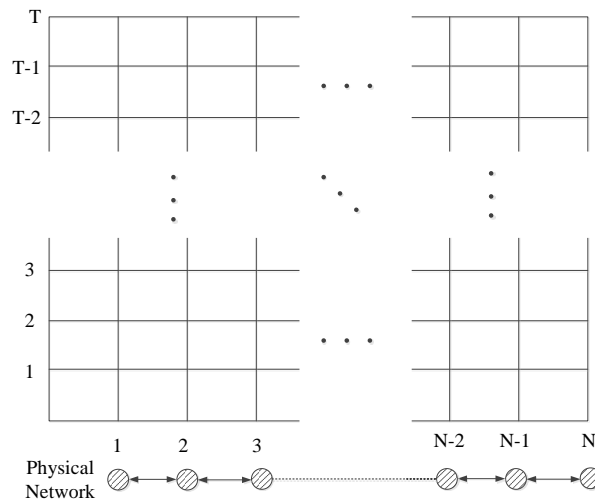

Figure 25 $N \times T$ grid network.

The proposed model and solution algorithms are implemented in Julia Language [145] using high performance computing with 8 processors and 40 GB of RAM. The (RLSTS) model is

also solved directly by commercial solver (CPLEX). Table 15 shows the tuned parameters used for SA and LR algorithm.

Table 15 The SA parameters and their values.

Algorithm	Parameter	Value
SA	T_{max}	40000
	T_{min}	0.0001
	$Iter_{max}$	10
	$\propto$	0.80
LR	μ^0	0.5
	$Iter_{max}$	500

5.5.1. Hypothetical Grid Networks

For the hypothetical networks, we use a solution time limit of 1800 seconds for all the algorithms and run a series of instances with different numbers of individuals (H), network sizes (N) and time periods (T). Let denote the optimal objective value obtained by CPLEX, its solution time and optimality gap by Z^*, t^* and ϵ^*, respectively. We also denote the objective values found by the greedy, LR and SA algorithms by Z^G, Z^{LR}, and Z^{SA}, respectively. The solution time for LR and SA are denoted by t^{LR} and t^{SA} respectively, and ϵ^{LR} represents the optimality gap for LR. Overall, 24 hypothetical networks are generated and the results are summarized in Table 16. For the first 15 cases, the exact objective value with 0 optimality gap is found by CPLEX. LR returns the exact solution as well, but the gap is not promising. This high gap is because of the nature of the relaxed problem (we will explain this later in section 5.6). Also, SA is returning the exact value for more than 60% of cases and it is performing well in the other cases.

From case 16 to 24, CPLEX is not able to solve the problem in 1800 s (or it returns a high optimality gap) but both LR and SA can find a near optimal solution. In cases 22 and 24, SA objective value is slightly better than LR and that is because LR needs to solve an LP in sub-

problem 2 in each iteration which is time consuming. The LR terminates after 1800 seconds with less number of iterations compared to SA and cannot reach a better objective value.

5.5.2. Large Scale Application

The second numerical example is based on the real geo-tagged twitter data that are collected in NYC area. The data is cleaning and preprocessing is explained in section 4.3. We further filter the good twitter users based on the cutoff points that are proposed in Table 12, and narrow down the study area to Manhattan, NY (Figure 26).

Table 16 Results for hypothetical grid networks.

Case #	Network size			CPLEX			Greedy	LR			SA	
	H	N	T	Z^*	t^*	ϵ^*	Z^G	Z^{LR}	t^{LR}	ϵ^{LR}	Z^{SA}	t^{SA}
1	10	12	24	42.74	0.21	0.00	42.74	42.74	7.67	0.26	42.74	3.28
2	10	24	48	40.53	0.31	0.00	34.52	40.53	33.07	0.26	40.53	9.13
3	25	12	24	108.50	0.62	0.00	101.78	108.50	27.05	0.22	108.50	6.66
4	25	24	48	79.61	1.17	0.00	70.75	79.61	59.53	0.25	76.27	22.11
5	50	12	24	206.92	0.93	0.00	196.83	206.92	42.57	0.22	206.92	14.03
6	50	24	48	160.56	3.03	0.00	150.80	160.56	207.23	0.27	156.99	42.75
7	100	12	24	382.96	2.92	0.00	332.89	382.96	178.88	0.27	382.96	27.35
8	100	24	48	283.32	12.29	0.00	220.65	283.32	431.37	0.26	276.11	101.10
9	200	12	24	799.12	11.34	0.00	654.71	779.12	357.45	0.30	779.12	60.36
10	200	24	48	582.44	24.87	0.00	575.99	582.44	784.17	0.47	575.99	180.50
11	500	12	24	1906.62	64.28	0.00	1832.41	1906.62	787.95	0.30	1906.62	165.21
12	500	24	48	1363.75	109.28	0.00	947.42	1363.75	1530.83	0.50	1256.13	662.40
13	1000	12	24	3627.19	161.84	0.00	3100.53	3627.19	1800	0.33	3627.19	378.45
14	1000	24	48	2702.57	482.10	0.00	1847.72	2702.57	1800	0.51	2503.20	1101.78
15	2500	12	24	9198.45	1023.74	0.00	7698.97	9198.45	1800	0.34	9198.45	1014.03
16	2500	24	48	6923.86	1800	0.13	6037.35	6923.86	1800	0.51	6285.72	1800
17	5000	12	24	13487.73	1800	0.59	15205.76	18551.18	1800	0.33	18551.18	1800
18	5000	24	48	-	1800	-	9081.96	12497.94	1800	0.56	12007.73	1800
19	7500	12	24	-	1800	-	22953.22	28117.34	1800	0.32	28117.34	1800
20	7500	24	48	-	1800	-	16972.08	18403.00	1800	0.56	18177.63	1800
21	10000	12	24	-	1800	-	30626.28	37646.79	1800	0.31	37646.79	1800
22	20000	12	24	-	1800	-	61029.89	70106.36	1800	0.36	74481.90	1800
23	25000	12	24	-	1800	-	81529.33	87444.82	1800	0.36	87444.82	1800
24	30000	12	24	-	1800	-	97667.47	108667.47	1800	0.41	113085.16	1800

The Google API is used to select 84 major intersection in Manhattan. Then, by using a map matching technique, the location of each sample tweet (longitude, latitude of each sample point) is matched to the nearest neighbor intersection if it falls within a radius of 100 meter, otherwise the sample point is dropped from the dataset. To construct the space-time network, 37 time stamps in 30 minute intervals are considered from 6 Am to 12 AM. Then, the time of each sample point is matched to its nearest time stamp and the corresponding space-time nodes are characterized by the longitude, latitude and time stamp.
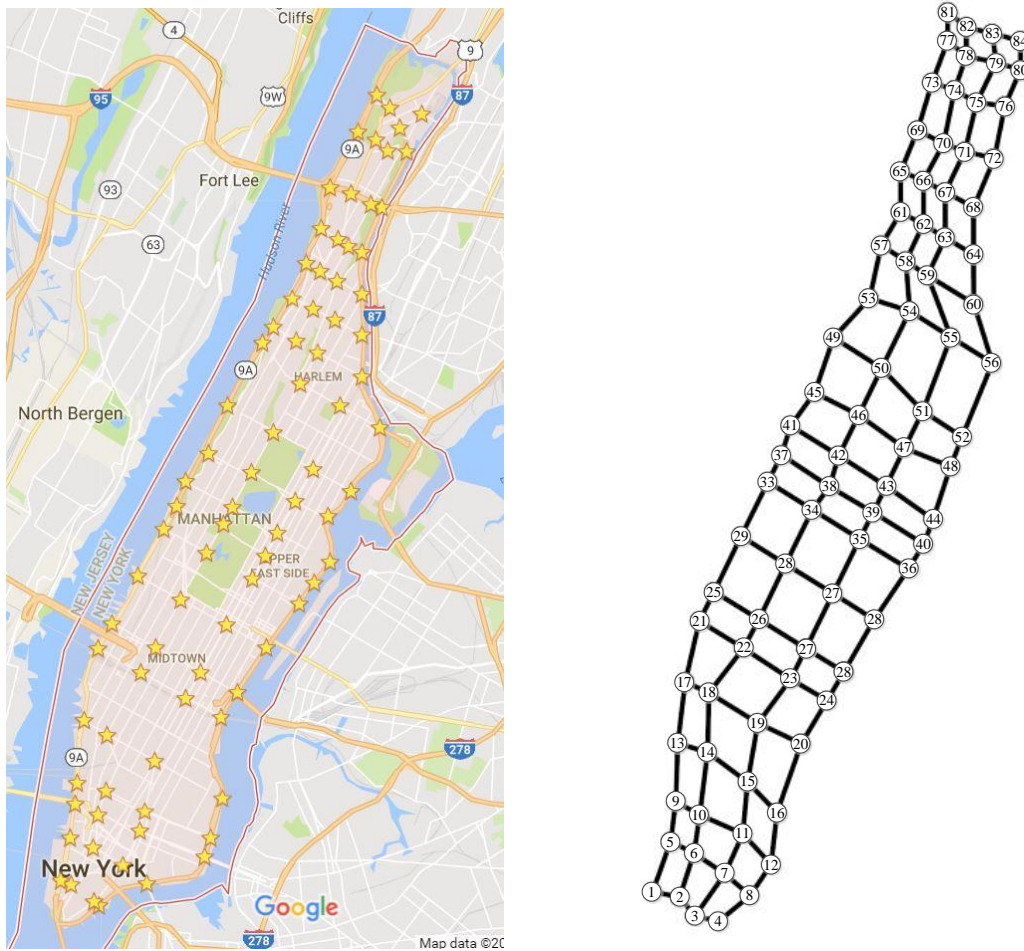

Figure 26 Manhattan Network.

The twitter data is collected from June 28 to July 18. In each day, the prism chains are formed for each individual target based on sample points from 6 Am to 12 AM and all days are

aggregated to increase the sample size, Eventually, 1022 individuals are selected and three cases are considered to apply the proposed model and compare the performance of solution approaches: (1) the first 50 individuals, (2) by first 500 individuals and (3) all 1022 individuals. The solution results are presented in Table 17. We set a solution time limit of 3600 seconds and due to the increased problem size, CPLEX and Gurobi are not able to find a meaningful feasible solution.

Table 17 Results for Manhattan network.

Case #	Network size			Greedy	LR			SA	
	H	N	T	Z^G	Z^{LR}	t^{LR}	ϵ^{LR}	Z^{SA}	t^{SA}
1	50	84	37	2860.30	3128.290365	1039	0.21	3079.21	448
2	500	84	37	27553.92	30028.4056	3600	0.30	28105.82	3600
3	1022	84	37	54656.05	62127.06621	3600	0.39	60511.57	3600

The greedy algorithm finds good feasible solutions that are approximately no smaller than 63% $(1 - \frac{1}{e})$ of the true optimum (see section 5.4.1 for more details). The results in Table 17 show that SA can improve the objective value found by greedy for cases 1 to 3 by 8%, 2% and 11%, respectively. On the other hand, LR improves greedy objective values by 9%, 9% and 14%. The relative gaps still are not promising in LR and that is because of the problem structure, as explained in section 5.6. However, from our experiments on hypothetical networks, the difference between the near-optimal solution and the true optimum is often much smaller than the relative gap and the solutions are reasonable in practice.

5.6. Results and Discussion

In this chapter, we try to integrate human travel characteristics and mobility patterns into a sensor network design problem. A Mixed Integer Nonlinear Programming (MINLP) model is proposed in a probabilistic space-time network structure. The proposed model is linearized and greedy, Lagrangian Relaxation and simulated annealing algorithms are designed to obtain near-optimum solutions. The results show that Greedy algorithm can yield a good feasible solution and

both LR and SA algorithms have a good performance in finding a near optimal solution for hypothetical grid networks (in fact, when the network size is small, they return the exact optimal solution found by CPLEX). Despite that, the optimality gap obtained by LR is not promising. The reason is that when we relax constraints (55) and (56) in (LSTS) formulation, the dummy variable v_{hit} does not relate to y_i any more (note that before relaxation $v_{hit} = q_{hit}$ when $y_i = 1$, and $v_{hit} = 0$ when $y_i = 0$). Therefore, since u_h is the predominant positive coefficient of v_{hit} in the objective function, the algorithm tries to maximize v_{hit}, i.e. make $v_{hit} = q_{hit}$ (considering constraint (66)). It means $\sum_{i \in N} \sum_{t \in T} v_{hit} = 1, \forall h \in H$, whereas in (LSTS) formulation when $y_i = 0$ we have $v_{hit} = 0$ and $\sum_{i \in N} \sum_{t \in T} v_{hit} \ll 1$ for some of the targets. This structure affects the optimality gap and the algorithm cannot improve the upper bound although it is able to find the exact solution in smaller case studies.

Future work can be conducted in several directions. First, a branch and bound framework can be implemented to slightly improve the optimality gap. The current study assumes that the space-time network is homogeneous and space-time node and links are independent with an equal visiting probability. However, in real world different weight factors may be considered regarding the land use associated with each node. For instance, the probability of moving to a space-time node in a residential area at 10 pm, is relatively higher than moving to a node in shopping centers at the same time. The target utilities (u_h) are randomly generated for in this study, however there are many factors that can be considered in the future for assigning different utilities to each target h e.g. sociodemographic variables. Exploring social media targets using text mining and machine learning techniques can also provide insights into formulating the defining utility functions.

CHAPTER 6: CONCLUSION

Facility network design has long been recognized as one of the most difficult and challenging problems in transportation and urban planning. In this study, at first we investigates a new framework for multimodal network design problems, by considering safety and flow interactions between different modes of transport. A mixed-integer nonlinear programming model is developed for optimally locating sidewalks and crosswalks in the network. The greedy heuristic and simulated annealing algorithms are implemented for finding a near optimal solution and numerical examples are conducted to evaluate the algorithms and draw managerial insights. According to the results, sidewalk and crosswalk installation can reduce the total transportation cost by 12%, 5% and 59% respectively for the Small, Hearn and Sioux Falls networks, respectively. The proposed model in task 1 only considers the long-term stationary travel patterns (e.g., fixed OD demand) and short-term variations of human mobility are ignored. Therefore, we defined our second task towards human mobility and travel characteristics.

In the second task of this research, a set of parsimonious measures based on time geography concepts are proposed to answer an important question about mobility studies using geo-tagged mobility sample data: How accurate it would be to utilize such samples in estimating continuous individual mobility trajectories? The estimated trajectory between a set of limited space-time sample points is obtained by connecting these sample points with linear segments. However, since the estimated trajectory may differ from an individual's unknown ground truth trajectory, a set of fundamental measures are proposed to quantify the accuracy of the estimation inaccuracy in a

robust manner. The estimation inaccuracy depends on the density of the sample points. In the proposed methodology, an individual's activity range around the estimated trajectory is constructed by a chain of space-time prisms. Then the proposed measures, including activity bandwidth and normalized activity bandwidth, are calculated on this chain. The activity bandwidth quantifies the possible absolute inaccuracy range between the estimated and the ground truth trajectories, while the normalized activity bandwidth measures the relative difference between the mobility pattern of the estimated trajectory and that of the ground truth trajectory.

In travel mobility analysis, these measures can be used to evaluate the suitability of estimated individual trajectories from generic geo-tagged social media data. Since it is time consuming to calculate these measures working with massive mobility data, we also propose a lookup table–based interpolation method to expedite the calculation. The proposed measures and the associated lookup table method have been tested with two sets of real-world geo-tagged mobility data from social-media and cell phone logs. These cases studies demonstrate that the proposed measures can efficiently quantify inaccuracies of using sample individual mobility data in estimating their continuous trajectories. They also draw a number of interesting managerial insights. For the Twitter data, we find that both measures proposed in this study follow power-law distributions at different traveling speed limits ($\bar{v}$). For the cellphone data, the normalized activity bandwidth again follows a power-law distribution, although the activity bandwidth measure can be better described with an exponential distribution. Our findings show that most individuals in these data sets likely yield high estimation inaccuracies and may not be suitable for mobility studies with high accuracy requirement. However, because of the massive amount of geo-tagged data available, there are still a good number of individuals with relatively accurate mobility

93

information for mobility pattern analysis. Nonetheless, cautions should be taken in screening these data for specific applications.

The final and predominant goal of this study is to integrate human travel characteristics and mobility patterns into traditional network design problems. We propose a probabilistic space-time network structure to quantify the presence probability of individual targets at different locations and times. Then, a new Mixed Integer Nonlinear Programming (MINLP) model is developed to maximize the network coverage. Customized solution approaches including greedy heuristic, Lagrangian relaxation and simulated annealing are designed to achieve near optimal solutions. A set of hypothetical space-time networks with different sizes and parameter settings and a real-world case study are designed to demonstrate the performance of the algorithms. The results show that for small networks, LR has the best performance and it yields the exact optimal solution found by CPLEX. As we increase the size of the hypothetical network or for the real-world case study, all algorithms can find good feasible solutions. In this case, the optimality gap in LR is not promising because of the problem structure as explained in section 5.6. However, from our experiments, the difference between the near-optimal solution and the true optimum is often much smaller than the optimality gap. This study can be extended in several directions in the future. A branch and bound framework can be implemented to slightly improve the optimality gap in LR. The current model rely on homogenous space-time network assumption and different weight factors may be considered regarding land use and other important factors. Finally, exploring different utility functions using text mining and machine learning techniques would be desirable in real world problems.

REFERENCES

[1]   M. S. Daskin, *Network and discrete location: Models, algorithms and applications:* Wiley New York, ISBN 0 47101 897, 1995.

[2]   Z. Drezner, *Facility location: a survey of applications and methods*. Springer, 1995.

[3]   C. F. Daganzo, *Logistics systems analysis*. Springer Science & Business Media, 2005.

[4]   A. Langevin, P. Mbaraga, and J. F. Campbell, "Continuous approximation models in freight distribution: An overview," *Transp. Res. Part B Methodol.*, vol. 30, no. 3, pp. 163–188, 1996.

[5]   R. Batta, J. M. Dolan, and N. N. Krishnamurthy, "The maximal expected covering location problem: Revisited," *Transp. Sci.*, vol. 23, no. 4, pp. 277–287, 1989.

[6]   M. S. Daskin, "A maximum expected covering location model: formulation, properties and heuristic solution," *Transp. Sci.*, vol. 17, no. 1, pp. 48–70, 1983.

[7]   J. Cui, M. Zhao, X. Li, M. Parsafard, and S. An, "Reliable design of an integrated supply chain with expedited shipments under disruption risks," *Transp. Res. Part E Logist. Transp. Rev.*, vol. 95, pp. 143–163, 2016.

[8]   R. Z. Farahani, E. Miandoabchi, W. Y. Szeto, and H. Rashidi, "A review of urban transportation network design problems," *Eur. J. Oper. Res.*, vol. 229, no. 2, pp. 281–302, 2013.

[9]   B. J. Campbell, C. V Zegeer, H. H. Huang, and M. J. Cynecki, "A Review of Pedestrian Safety Research in the United States and Abroad," 2004.

[10]   FHWA, "Safety Benefits of Walkways, Sidewalks, and Paved Shoulders. Technical report, U.S. Dept. of Transportation, Federal Highway Administration, Office of Safety, Washington, D.C.," 2010. [Online]. Available: http://safety.fhwa.dot.gov/ped_bike/tools_solve/walkways_trifold/. [Accessed: 22-Apr-2014].

[11]   M. C. González, C. A. Hidalgo, and A.-L. Barabási, "Understanding individual human mobility patterns," *Nature*, vol. 453, no. 7196, pp. 779–782, 2008.

[12]   S. Hasan, C. M. Schneider, S. V. Ukkusuri, and M. C. González, "Spatiotemporal patterns of urban human mobility," *J. Stat. Phys.*, vol. 151, no. 1–2, pp. 304–318, 2013.

[13]   J. Candia, M. C. González, P. Wang, T. Schoenharl, G. Madey, and A.-L. Barabási, "Uncovering individual and collective human dynamics from mobile phone records," *J. Phys. Math. Theor.*, vol. 41, no. 22, p. 224015, 2008.

[14]   T. M. T. Do and D. Gatica-Perez, "Contextual conditional models for smartphone-based human mobility prediction," in *Proceedings of the 2012 ACM Conference on Ubiquitous Computing*, 2012, pp. 163–172.

[15]   S. Hasan, X. Zhan, and S. V. Ukkusuri, "Understanding urban human activity and mobility patterns using large-scale location-based data from online social media," in *Proceedings of the 2nd ACM SIGKDD International Workshop on Urban Computing*, 2013, p. 6.

[16] I. Salomon and M. Ben-Akiva, "The use of the life-style concept in travel demand models," *Environ. Plan. A*, vol. 15, no. 5, pp. 623–638, 1983.

[17] R. Kitamura, "Life-style and travel demand," *Transportation*, vol. 36, no. 6, pp. 679–710, 2009.

[18] M. M. Haque and S. Washington, "The impact of mobile phone distraction on the braking behaviour of young drivers: a hazard-based duration model," *Transp. Res. Part C Emerg. Technol.*, vol. 50, pp. 13–27, 2015.

[19] D. McFadden, "The measurement of urban travel demand," *J. Public Econ.*, vol. 3, no. 4, pp. 303–328, 1974.

[20] J. Kim and H. S. Mahmassani, "Trajectory Clustering for Discovering Spatial Traffic Flow Patterns in Road Networks," in *Transportation Research Board 94th Annual Meeting*, 2015.

[21] J. Yuan, Y. Zheng, L. Zhang, Xi. Xie, and G. Sun, "Where to find my next passenger," in *Proceedings of the 13th international conference on Ubiquitous computing*, 2011, pp. 109–118.

[22] L. Sloan, J. Morgan, P. Burnap, and M. Williams, "Who tweets? Deriving the demographic characteristics of age, occupation and social class from Twitter user meta-data," *PloS One*, vol. 10, no. 3, p. e0115545, 2015.

[23] Y. Zheng, "Trajectory data mining: an overview," *ACM Trans. Intell. Syst. Technol. TIST*, vol. 6, no. 3, p. 29, 2015.

[24] T. Hägerstraand, "What about people in regional science?," *Pap. Reg. Sci.*, vol. 24, no. 1, pp. 7–24, 1970.

[25] H. J. Miller, "A measurement theory for time geography," *Geogr. Anal.*, vol. 37, no. 1, pp. 17–45, 2005.

[26] B. Kuijpers and W. Othman, "Modeling uncertainty of moving objects on road networks via space–time prisms," *Int. J. Geogr. Inf. Sci.*, vol. 23, no. 9, pp. 1095–1117, 2009.

[27] R. Cheng, D. V. Kalashnikov, and S. Prabhakar, "Querying imprecise data in moving object environments," *IEEE Trans. Knowl. Data Eng.*, vol. 16, no. 9, pp. 1112–1127, 2004.

[28] S. Qiao *et al.*, "PutMode: prediction of uncertain trajectories in moving objects databases," *Appl. Intell.*, vol. 33, no. 3, pp. 370–386, 2010.

[29] D. Liu *et al.*, "SmartAdP: Visual Analytics of Large-scale Taxi Trajectories for Selecting Billboard Locations," *IEEE Trans. Vis. Comput. Graph.*, vol. 23, no. 1, pp. 1–10, 2017.

[30] G. F. Newell, "A simplified car-following theory: a lower order model," *Transp. Res. Part B Methodol.*, vol. 36, no. 3, pp. 195–205, Mar. 2002.

[31] X. Li, J. Cui, S. An, and M. Parsafard, "Stop-and-go traffic analysis: Theoretical properties, environmental impacts and oscillation mitigation," *Transp. Res. Part B Methodol.*, vol. 70, pp. 319–339, 2014.

[32] M. Parsafard, X. Li, J. Cui, and S. An, "A Describing Function Method for Traffic Oscillation Analysis: Theoretical Properties," in *Transportation Research Board 94th Annual Meeting*, 2015.

[33] M. Parsafard, X. Li, J. Cui, and S. An, "A Describing Function Method for Traffic Oscillation Analysis: Environmental Impacts and Oscillation Mitigation," in *Transportation Research Board 94th Annual Meeting*, 2015.

[34] D. Brockmann, V. David, and A. M. Gallardo, "Human mobility and spatial disease dynamics," *Rev. Nonlinear Dyn. Complex.*, vol. 2, pp. 1–24, 2009.

[35] U. Paul, A. P. Subramanian, M. M. Buddhikot, and S. R. Das, "Understanding traffic dynamics in cellular data networks," in *INFOCOM, 2011 Proceedings IEEE*, 2011, pp. 882–890.

[36] X. Lu, E. Wetter, N. Bharti, A. J. Tatem, and L. Bengtsson, "Approaching the limit of predictability in human mobility," *Sci. Rep.*, vol. 3, 2013.

[37] C. Song, T. Koren, P. Wang, and A.-L. Barabási, "Modelling the scaling properties of human mobility," *Nat. Phys.*, vol. 6, no. 10, pp. 818–823, 2010.

[38] J. K. Hackney, J. K. Hackney, J. K. Hackney, and S. N. zur F. der Wissenschaftlichen, *Coevolving social and transportation networks*. ETH, Eidgenössische Technische Hochschule Zürich, IVT, Institut für Verkehrsplanung und Transportsysteme, 2005.

[39] K. W. Axhausen, "Social networks, mobility biographies, and travel: survey challenges," *Environ. Plan. B Plan. Des.*, vol. 35, no. 6, p. 981, 2008.

[40] D. Brockmann, L. Hufnagel, and T. Geisel, "The scaling laws of human travel," *Nature*, vol. 439, no. 7075, pp. 462–465, 2006.

[41] C. Song, Z. Qu, N. Blumm, and A.-L. Barabási, "Limits of predictability in human mobility," *Science*, vol. 327, no. 5968, pp. 1018–1021, 2010.

[42] R. Metzler and J. Klafter, "The random walk's guide to anomalous diffusion: a fractional dynamics approach," *Phys. Rep.*, vol. 339, no. 1, pp. 1–77, 2000.

[43] T. Arentze and H. Timmermans, "Social networks, social interactions, and activity-travel behavior: a framework for microsimulation," *Environ. Plan. B Plan. Des.*, vol. 35, no. 6, p. 1012, 2008.

[44] S. Hasan, "Modeling urban mobility dynamics using geo-location data," PURDUE UNIVERSITY, 2013.

[45] Y. Gu, Z. S. Qian, and F. Chen, "From Twitter to detector: Real-time traffic incident detection using social media data," *Transp. Res. Part C Emerg. Technol.*, vol. 67, pp. 321–342, 2016.

[46] P. Krishnamurthy and K. Pelechrinis, "Location based social networks," *Adv. Locat.-Based Technol. Serv.*, p. 127, 2013.

[47] H. Gao, J. Tang, and H. Liu, "Exploring Social-Historical Ties on Location-Based Social Networks.," in *ICWSM*, 2012.

[48] S. Scellato, A. Noulas, R. Lambiotte, and C. Mascolo, "Socio-Spatial Properties of Online Location-Based Social Networks.," *ICWSM*, vol. 11, pp. 329–336, 2011.

[49] S. Hasan and S. V. Ukkusuri, "Urban activity pattern classification using topic models from online geo-location data," *Transp. Res. Part C Emerg. Technol.*, vol. 44, pp. 363–381, 2014.

[50] S. Bregman and K. E. Watkins, *Best Practices for Transportation Agency Use of Social Media*. CRC Press, 2013.

[51] P. J. H. Daas, M. J. Puts, B. Buelens, and P. A. M. Van den Hurk, "Big Data and official statistics," in *Paper for the 2013 New Techniques and Technologies for Statistics conference, Brussels, Belgium*, 2013.

[52] W. Dong, H. Vu, Y. Nazarathy, B. Vo, M. Li, and S. Hoogendoorn, "Shortest paths in Stochastic time-dependent networks with link travel time correlation," *Transp. Res. Rec. J. Transp. Res. Board*, no. 2338, pp. 58–66, 2013.

[53] Y. M. Nie and X. Wu, "Shortest path problem considering on-time arrival probability," *Transp. Res. Part B Methodol.*, vol. 43, no. 6, pp. 597–613, 2009.

[54] H. J. Miller, "Necessary space-time conditions for human interaction," *Environ. Plan. B Plan. Des.*, vol. 32, no. 3, pp. 381–401, 2005.

[55] J. Tang, Y. Song, H. J. Miller, and X. Zhou, "Estimating the most likely space–time paths, dwell times and path uncertainties from vehicle trajectory data: A time geographic method," *Transp. Res. Part C Emerg. Technol.*, vol. 66, pp. 176–194, May 2016.

[56] L. Tong, X. Zhou, and H. J. Miller, "Transportation network design for maximizing space–time accessibility," *Transp. Res. Part B Methodol.*, vol. 81, pp. 555–576, 2015.

[57] N. Thrift, "An introduction to time geography," 1977.

[58] R. G. Golledge, *Spatial behavior: A geographic perspective*. Guilford Press, 1997.

[59] H. Yu and S.-L. Shaw, "Revisiting Hägerstrand's time-geographic framework for individual activities in the age of instant access," in *Societies and cities in the age of instant access*, Springer, 2007, pp. 103–118.

[60] H. Yu and S.-L. Shaw, "Exploring potential human activities in physical and virtual spaces: a spatio-temporal GIS approach," *Int. J. Geogr. Inf. Sci.*, vol. 22, no. 4, pp. 409–430, 2008.

[61] S.-L. Shaw, H. Yu, and L. S. Bombom, "A space-time GIS approach to exploring large individual-based spatiotemporal datasets," *Trans. GIS*, vol. 12, no. 4, pp. 425–441, 2008.

[62] M.-P. Kwan, "Interactive geovisualization of activity-travel patterns using three-dimensional geographical information systems: a methodological exploration with a large data set," *Transp. Res. Part C Emerg. Technol.*, vol. 8, no. 1, pp. 185–203, 2000.

[63] G. Chi, J. R. Porter, A. G. Cosby, and D. Levinson, "The impact of gasoline price changes on traffic safety: a time geography explanation," *J. Transp. Geogr.*, vol. 28, pp. 1–11, 2013.

[64] H. J. Miller and S. A. Bridwell, "A field-based theory for time geography," *Ann. Assoc. Am. Geogr.*, vol. 99, no. 1, pp. 49–75, 2009.

[65] H. Tian, X. Ma, H. Wang, G. Song, and K. Xie, "A novel approach to estimate human space-time path based on mobile phone call records," in *Geoinformatics, 2010 18th International Conference on*, 2010, pp. 1–6.

[66] M.-P. Kwan, "Gender and individual access to urban opportunities: a study using space–time measures," *Prof. Geogr.*, vol. 51, no. 2, pp. 210–227, 1999.

[67] H. J. Miller, "Measuring space-time accessibility benefits within transportation networks: basic theory and computational procedures," *Geogr. Anal.*, vol. 31, no. 1, pp. 1–26, 1999.

[68] H.-M. Kim and M.-P. Kwan, "Space-time accessibility measures: A geocomputational algorithm with a focus on the feasible opportunity set and possible activity duration," *J. Geogr. Syst.*, vol. 5, no. 1, pp. 71–91, 2003.

[69] Y. Leung, J.-H. Ma, and M. F. Goodchild, "A general framework for error analysis in measurement-based GIS Part 1: The basic measurement-error model and related concepts," *J. Geogr. Syst.*, vol. 6, no. 4, pp. 325–354, 2004.

[70] T. Kobayashi, H. J. Miller, and W. Othman, "Analytical methods for error propagation in planar space–time prisms," *J. Geogr. Syst.*, vol. 13, no. 4, pp. 327–354, 2011.

[71] S. Winter and Z.-C. Yin, "The elements of probabilistic time geography," *GeoInformatica*, vol. 15, no. 3, pp. 417–434, 2011.

[72] S. Winter and Z.-C. Yin, "Directed movements in probabilistic time geography," *Int. J. Geogr. Inf. Sci.*, vol. 24, no. 9, pp. 1349–1365, 2010.

[73] Y. Song and H. J. Miller, "Simulating visit probability distributions within planar space-time prisms," *Int. J. Geogr. Inf. Sci.*, vol. 28, no. 1, pp. 104–125, 2014.

[74] Y. Leung, Z. Zhao, and J.-H. Ma, "Uncertainty analysis of space–time prisms based on the moment-design method," *Int. J. Geogr. Inf. Sci.*, vol. 30, no. 7, pp. 1336–1358, 2016.

[75] H. Farvaresh and M. M. Sepehri, "A branch and bound algorithm for bi-level discrete network design problem," *Netw. Spat. Econ.*, vol. 13, no. 1, pp. 67–106, 2013.

[76] Z. Khooban, R. Z. Farahani, E. Miandoabchi, and W. Y. Szeto, "Mixed network design using hybrid scatter search," *Eur. J. Oper. Res.*, vol. 247, no. 3, pp. 699–710, 2015.

[77] W. Y. Szeto and Y. Jiang, "Transit route and frequency design: Bi-level modeling and hybrid artificial bee colony algorithm approach," *Transp. Res. Part B Methodol.*, vol. 67, pp. 235–263, 2014.

[78] B. Yu, L. Kong, Y. Sun, B. Yao, and Z. Gao, "A bi-level programming for bus lane network design," *Transp. Res. Part C Emerg. Technol.*, vol. 55, pp. 310–327, 2015.

[79] Y. U. Seo, J. H. Park, H. Jang, and Y. I. Lee, "A study on setting-up a methodology and criterion of exclusive bus lane in urban area," in *Proceedings of the Eastern Asia Society for Transportation Studies*, 2005, vol. 5, pp. 339–351.

[80] M. Gallo, B. Montella, and L. D'Acierno, "The transit network design problem with elastic demand and internalisation of external costs: An application to rail frequency optimisation," *Transp. Res. Part C Emerg. Technol.*, vol. 19, no. 6, pp. 1276–1305, 2011.

[81] E. Miandoabchi, R. Z. Farahani, W. Dullaert, and W. Y. Szeto, "Hybrid evolutionary metaheuristics for concurrent multi-objective design of urban road and public transit networks," *Netw. Spat. Econ.*, vol. 12, no. 3, pp. 441–480, 2012.

[82] E. Miandoabchi, R. Z. Farahani, and W. Y. Szeto, "Bi-objective bimodal urban road network design using hybrid metaheuristics," *Cent. Eur. J. Oper. Res.*, vol. 20, no. 4, pp. 583–621, 2012.

[83] W. Y. Szeto, X. Jaber, and M. O'Mahony, "Time-Dependent Discrete Network Design Frameworks Considering Land Use," *Comput.-Aided Civ. Infrastruct. Eng.*, vol. 25, no. 6, pp. 411–426, 2010.

[84] R. Buehler and J. Pucher, "Walking and cycling in Western Europe and the United States: trends, policies, and lessons," *TR News*, no. 280, 2012.

[85] H. Millward, J. Spinney, and D. Scott, "Active-transport walking behavior: destinations, durations, distances," *J. Transp. Geogr.*, vol. 28, pp. 101–110, 2013.

[86] O. T. Stewart, A. V. Moudon, B. E. Saelens, C. Lee, B. Kang, and M. P. Doescher, "Comparing associations between the built environment and walking in rural small towns and a large metropolitan area," *Environ. Behav.*, vol. 48, no. 1, pp. 13–36, 2016.

[87] B. Beltran, S. Carrese, E. Cipriani, and M. Petrelli, "Transit network design with allocation of green vehicles: A genetic algorithm approach," *Transp. Res. Part C Emerg. Technol.*, vol. 17, no. 5, pp. 475–483, 2009.

[88] G. E. Cantarella, G. Pavone, and A. Vitetta, "Heuristics for urban road network design: Lane layout and signal settings," *Eur. J. Oper. Res.*, vol. 175, no. 3, pp. 1682–1695, 2006.

[89] W. Fan and R. B. Machemehl, "Tabu search strategies for the public transportation network optimizations with variable transit demand," *Comput.-Aided Civ. Infrastruct. Eng.*, vol. 23, no. 7, pp. 502–520, 2008.

[90] M. Mesbah, M. Sarvi, and G. Currie, "New methodology for optimizing transit priority at the network level," *Transp. Res. Rec. J. Transp. Res. Board*, no. 2089, pp. 93–100, 2008.

[91] H. G. Resat and M. Turkay, "Design and operation of intermodal transportation network in the Marmara region of Turkey," *Transp. Res. Part E Logist. Transp. Rev.*, vol. 83, pp. 16–33, 2015.

[92]   B. Yao, P. Hu, X. Lu, J. Gao, and M. Zhang, "Transit network design based on travel time reliability," *Transp. Res. Part C Emerg. Technol.*, vol. 43, pp. 233–248, 2014.

[93]   O. Berman and D. Krass, "On n-facility median problem with facilities subject to failure facing uniform demand," *Discrete Appl. Math.*, vol. 159, no. 6, pp. 420–432, 2011.

[94]   T. Cui, Y. Ouyang, and Z.-J. M. Shen, "Reliable facility location design under the risk of disruptions," *Oper. Res.*, vol. 58, no. 4-part-1, pp. 998–1011, 2010.

[95]   L. V. Snyder, M. S. Daskin, and C.-P. Teo, "The stochastic location model with risk pooling," *Eur. J. Oper. Res.*, vol. 179, no. 3, pp. 1221–1238, 2007.

[96]   L. V. Snyder and M. S. Daskin, "Reliability models for facility location: the expected failure cost case," *Transp. Sci.*, vol. 39, no. 3, pp. 400–416, 2005.

[97]   X. Li and Y. Ouyang, "A continuum approximation approach to reliable facility location design under correlated probabilistic disruptions," *Transp. Res. Part B Methodol.*, vol. 44, no. 4, pp. 535–548, 2010.

[98]   X. Li and Y. Ouyang, "Reliable Traffic Sensor Deployment Under Probabilistic Disruptions and Generalized Surveillance Effectiveness Measures," *Oper. Res.*, vol. 60, no. 5, pp. 1183–1198, Sep. 2012.

[99]   X. Li and Y. Ouyang, "Reliable sensor deployment for network traffic surveillance," *Transp. Res. Part B Methodol.*, vol. 45, no. 1, pp. 218–231, 2011.

[100] S. Hu, S. Peeta, and C. Chu, "Identification of vehicle sensor locations for link-based network traffic applications," *Transp. Res. Part B Methodol.*, 2009.

[101] J. A. Laval, "Hysteresis in traffic flow revisited: an improved measurement method," *Transp. Res. Part B Methodol.*, vol. 45, no. 2, pp. 385–391, 2011.

[102] X. Li, F. Peng, and Y. Ouyang, "Measurement and estimation of traffic oscillation properties," *Transp. Res. Part B Methodol.*, vol. 44, no. 1, pp. 1–14, 2010.

[103] H. Yang and J. Zhou, "Optimal traffic counting locations for origin–destination matrix estimation," *Transp. Res. Part B Methodol.*, vol. 32, no. 2, pp. 109–126, Feb. 1998.

[104] M. Gentili and P. B. Mirchandani, "Locating sensors on traffic networks: Models, challenges and research opportunities," *Transp. Res. Part C Emerg. Technol.*, vol. 24, pp. 227–255, 2012.

[105] E. Castillo, I. Gallego, J. M. Menéndez, and A. Rivas, "Optimal use of plate-scanning resources for route flow estimation in traffic networks," *Intell. Transp. Syst. IEEE Trans. On*, vol. 11, no. 2, pp. 380–391, 2010.

[106] S. He, "A graphical approach to identify sensor locations for link flow inference," *Transp. Res. Part B Methodol.*, vol. 51, pp. 65–76, 2013.

[107] M. Ng, "Synergistic sensor location for link flow inference without path enumeration: a node-based approach," *Transp. Res. Part B Methodol.*, vol. 46, no. 6, pp. 781–788, 2012.

[108] P. Chootinan, A. Chen, and H. Yang, "A bi-objective traffic counting location problem for origin-destination trip table estimation," *Transportmetrica*, vol. 1, no. 1, pp. 65–80, 2005.

[109] X. Fei, H. S. Mahmassani, and P. Murray-Tuite, "Vehicular network sensor placement optimization under uncertainty," *Transp. Res. Part C Emerg. Technol.*, vol. 29, pp. 14–31, 2013.

[110] X. Fei and H. S. Mahmassani, "Structural analysis of near-optimal sensor locations for a stochastic large-scale network," *Transp. Res. Part C Emerg. Technol.*, vol. 19, no. 3, pp. 440–453, 2011.

[111] X. Zhou and G. F. List, "An information-theoretic sensor location model for traffic origin-destination demand estimation applications," *Transp. Sci.*, vol. 44, no. 2, pp. 254–273, 2010.

[112] L. Bianco, G. Confessore, and M. Gentili, "Combinatorial aspects of the sensor location problem," *Ann. Oper. Res.*, vol. 144, no. 1, pp. 201–234, 2006.

[113] L. Bianco, G. Confessore, and P. Reverberi, "A network based model for traffic sensor location with implications on O/D matrix estimates," *Transp. Sci.*, vol. 35, no. 1, pp. 50–60, 2001.

[114] S. Eisenman, X. Fei, X. Zhou, and H. Mahmassani, "Number and location of sensors for real-time network traffic estimation and prediction: Sensitivity analysis," *Transp. Res. Rec. J. Transp. Res. Board*, no. 1964, pp. 253–259, 2006.

[115] P. B. Gentili, M. and Mirchandani, "Locating sensors on traffic networks: Models, challenges and research opportunities," *Transp. Res. Part C Emerg. Technol.*, vol. 24, pp. 227–255, 2012.

[116] Y. Liu, N. Zhu, S. Ma, and N. Jia, "Traffic sensor location approach for flow inference," *IET Intell. Transp. Syst.*, vol. 9, no. 2, pp. 184–192, 2015.

[117] R. Mínguez and S. Sánchez-Cambronero, "Optimal traffic plate scanning location for OD trip matrix and route estimation in road networks," *... Res. Part B ...*, 2010.

[118] F. Viti, M. Rinaldi, F. Corman, and C. M. Tampère, "Assessing partial observability in network sensor location problems," *Transp. Res. Part B Methodol.*, vol. 70, pp. 65–89, 2014.

[119] T. Xing, X. Zhou, and J. Taylor, "Designing heterogeneous sensor networks for estimating and predicting path travel time dynamics: An information-theoretic modeling approach," *Transp. Res. Part B Methodol.*, 2013.

[120] C. Cerrone, R. Cerulli, and M. Gentili, "Vehicle-ID sensor location for route flow recognition: Models and algorithms," *Eur. J. Oper. Res.*, 2015.

[121] G. Wang, G. Cao, T. La Porta, and W. Zhang, "Sensor relocation in mobile sensor networks," in *Proceedings IEEE 24th Annual Joint Conference of the IEEE Computer and Communications Societies.*, 2005, vol. 4, pp. 2302–2312.

[122] W. Hart, N. Gharaibeh, and others, "Use of micro unmanned aerial vehicles in roadside condition surveys," in *Transportation and Development Institute Congress 2011*, 2011, pp. 80–92.

[123] J. Zhang, L. Jia, S. Niu, F. Zhang, L. Tong, and X. Zhou, "A space-time network-based modeling framework for dynamic unmanned aerial vehicle routing in traffic incident monitoring applications," *Sensors*, 2015.

[124] N. Zhu, Y. Liu, S. Ma, and Z. He, "Mobile traffic sensor routing in dynamic transportation systems," *Intell. Transp. Syst. ...*, 2014.

[125] M. Parsafard, A. Esmaeel, K. Masoud, N. Mohammadreza, and X. Li, "Practical Approach for Finding Optimum Routes for Fuel Delivery Trucks in Large Cities," *Transp. Res. Rec. J. Transp. Res. Board*, vol. 2478, pp. 66–74, 2015.

[126] E. Rashidi, M. Parsafard, H. Medal, and X. Li, "Optimal traffic calming: A mixed-integer bi-level programming model for locating sidewalks and crosswalks in a multimodal transportation network to maximize pedestrians' safety and network usability," *Transp. Res. Part E Logist. Transp. Rev.*, vol. 91, pp. 33–50, 2016.

[127] M. Parsafard, E. Rashidi, X. Li, and H. Medal, "A Multi-modal User Equilibrium Traffic Assignment with Network Expansion," *Submitt. Transp. Lett.*, 2017.

[128] P. E. Gårder, "The impact of speed and other variables on pedestrian safety in Maine," *Accid. Anal. Prev.*, vol. 36, no. 4, pp. 533–542, 2004.

[129] Aashtiani, "The multi-modal traffic assignment problem (Doctoral dissertation).," Massachusetts Institute of Technology, 1979.

[130] O. Ben-Ayed and C. E. Blair, "Computational difficulties of bilevel linear programming," *Oper. Res.*, vol. 38, no. 3, pp. 556–560, 1990.

[131] J. Lofberg, "YALMIP: A toolbox for modeling and optimization in MATLAB," in *Computer Aided Control Systems Design, 2004 IEEE International Symposium on*, 2004, pp. 284–289.

[132] N. ~V. Sahinidis, "BARON 14.3.1: Global Optimization of Mixed-Integer Nonlinear Programs, User's Manual." 2014.

[133] S. Toobaie, H. Z. Aashtiani, M. Hamedi, and A. Haghani, "Application of a complementarity approach to large scale traffic equilibrium problems," in *Transportation Research Board*, 2010, pp. 1–29.

[134] M. Frank and P. Wolfe, "An algorithm for quadratic programming," *Nav. Res. Logist. Q.*, vol. 3, no. 1–2, pp. 95–110, 1956.

[135] K. Bouleimen and H. Lecocq, "A new efficient simulated annealing algorithm for the resource-constrained project scheduling problem and its multiple mode version," *Eur. J. Oper. Res.*, vol. 149, no. 2, pp. 268–281, 2003.

[136] D. W. Hearn and M. V. Ramana, *Solving congestion toll pricing models*. Springer US, 1998.

[137] A. Vandervalk, H. Louch, J. Guerre, and R. Margiotta, *Incorporating Reliability Performance Measures Into the Transportation Planning and Programming Processes: Technical Reference*, no. SHRP 2 Report S2-L05-RR-3. 2014.

[138] M. Parsafard, E. Rashidi, X. Li, and H. Medal, "A Multimodal User Equilibrium Traffic Assignment with Network Reconstruction," 2015.

[139] M. Southworth, "Designing the walkable city," *J. Urban Plan. Dev.*, vol. 131, no. 4, pp. 246–257, 2005.

[140] M. Parsafard, G. Chi, X. Qu, X. Li, and H. WANG, "Error Measures for Trajectories Estimations with Geo-tagged Mobility Sample Data," *Submitt. IEEE Trans. Intell. Transp. Syst.*, 2017.

[141] Z. Cheng, J. Caverlee, K. Lee, and D. Z. Sui, "Exploring Millions of Footprints in Location Sharing Services.," *ICWSM*, vol. 2011, pp. 81–88, 2011.

[142] J. Alstott, E. Bullmore, and D. Plenz, "powerlaw: a Python package for analysis of heavy-tailed distributions," *PloS One*, vol. 9, no. 1, p. e85777, 2014.

[143] T. Clouqueur, V. Phipatanasuphorn, P. Ramanathan, and K. K. Saluja, "Sensor deployment strategy for target detection," in *Proceedings of the 1st ACM international workshop on Wireless sensor networks and applications*, 2002, pp. 42–48.

[144] R. K. Iyer and J. A. Bilmes, "Submodular optimization with submodular cover and submodular knapsack constraints," in *Advances in Neural Information Processing Systems*, 2013, pp. 2436–2444.

[145] C. Kwon, *Julia Programming for Operations Research: A Primer on Computing*. 2016.

[146] A. Clauset, C. R. Shalizi, and M. E. Newman, "Power-law distributions in empirical data," *SIAM Rev.*, vol. 51, no. 4, pp. 661–703, 2009.

APPENDIX A: PROOF OF PROPOSITION 1

We will show that the objective function of the lower level problem, $\varphi_{l,m}(.)$, is a combination of convex functions and therefore convex. First we will show that the objective function for the auto link, $\varphi_{l,a}(.)$, (equation (9)) is convex. If we ignore ϑ and μ_{la} in (9) (as they are positive constant values), and denote $x_{lm} = \sum_{k \in K} X_{klm}$ for all links (l, m), we can rewrite (9) as:

$$\varphi_{l,a}(X,Y) = \frac{1}{(\gamma_{l,s})^{\beta_1}} \left(x_{l,s1}{}^{\beta_1} + x_{l,s2}{}^{\beta_1} \right) + \frac{1}{(\gamma_{l,c})^{\beta_2}} \left(x_{l,bx}{}^{\beta_2} + x_{l,ex}{}^{\beta_2} \right)$$
$$+ t_{la} \left(1 + \alpha_1 \left(\frac{x_{la} + \omega x_{l,t}}{\gamma_{la}} \right)^{\beta_3} \right) + \mu_{l,a}$$

As x^a is convex for $a \geq 1$ or $a < 0$, therefore, for $\beta_1, \beta_2 \geq 1$ or $\beta_1, \beta_2 < 0$ the first line (parts related to S&C) of (16) is convex. For simplicity we assume $\sum_{k \in K} X_{kla} = x$ and $\sum_{k \in K} X_{k,l,t} = y$; then, the Hessian matrix of the last part in equation (16) (the second line) is:

$$H = \frac{1}{\gamma_{la}{}^2} \begin{bmatrix} \beta_3(\beta_3-1)\alpha_1 t_{la}(\frac{x+\omega y}{\gamma_{la}})^{\beta_3-2} & \beta_3(\beta_3-1)\alpha_1 t_{la}\omega(\frac{x+\omega y}{\gamma_{la}})^{\beta_3-2} \\ \beta_3(\beta_3-1)\alpha_1 t_{la}\omega(\frac{x+\omega y}{\gamma_{la}})^{\beta_3-2} & \beta_3(\beta_3-1)\alpha_1 t_{la}\omega^2(\frac{x+\omega y}{\gamma_{la}})^{\beta_3-2} \end{bmatrix}$$

A $n \times n$ real symmetric matrix M is positive semi-definite if $z^T \times M \times z \geq 0$ for all non-zero vectors z with real entries. For an arbitrary non-zero vector $z = \begin{bmatrix} a & b \end{bmatrix}$, if we show that $z^T \times H \times z \geq 0$, then H is positive semidefinite.

$z^T \times H \times z = [a\beta_3(\beta_3-1)\alpha_1 t_{la}\left(\frac{x+\omega y}{\gamma_{la}}\right)^{\beta_3-2} + b\beta_3(\beta_3-1)\alpha_1 t_{la}\omega(\frac{x+\omega y}{\gamma_{la}})^{\beta_3-2}; a\beta_3(\beta_3-$
$1)\alpha_1 t_{la}\omega\left(\frac{x+\omega y}{\gamma_{la}}\right)^{\beta_3-2} + b\beta_3(\beta_3-1)\alpha_1 t_{la}\omega^2\left(\frac{x+\omega y}{\gamma_{la}}\right)^{\beta_3-2}] \times \begin{bmatrix} a \\ b \end{bmatrix} =$

$$a^2\beta_3(\beta_3-1)\alpha_1 t_{la}\left(\frac{x+\omega y}{\gamma_{la}}\right)^{\beta_3-2} + ab\beta_3(\beta_3-1)\alpha_1 t_{la}\omega\left(\frac{x+\omega y}{\gamma_{la}}\right)^{\beta_3-2} + ba\beta_3(\beta_3-$$
$$1)\alpha_1 t_{la}\omega\left(\frac{x+\omega y}{\gamma_{la}}\right)^{\beta_3-2} + b^2\beta_3(\beta_3-1)\alpha_1 t_{la}\omega^2\left(\frac{x+\omega y}{\gamma_{la}}\right)^{\beta_3-2}$$

We need to show that the above expression is non-negative:

$$a^2\beta_3(\beta_3-1)\alpha_1 t_{la}\left(\frac{x+\omega y}{\gamma_{la}}\right)^{\beta_3-2} + 2ab\beta_3(\beta_3-1)\alpha_1 t_{la}\omega\left(\frac{x+\omega y}{\gamma_{la}}\right)^{\beta_3-2} + b^2\beta_3(\beta_3-$$
$$1)\alpha_1 t_{la}\omega^2\left(\frac{x+\omega y}{\gamma_{la}}\right)^{\beta_3-2}$$
$$= \beta_3(\beta_3-1)\alpha_1 t_{la}\left(\frac{x+\omega y}{\gamma_{la}}\right)^{\beta_3-2}(a^2+2ab\omega+b^2\omega^2)$$
$$= \beta_3(\beta_3-1)\alpha_1 t_{la}\left(\frac{x+\omega y}{\gamma_{la}}\right)^{\beta_3-2}(a+b\omega)^2$$

Knowing that x, y, t_{la}, ω and γ_{la} are all positive, for $\beta_3 \geq 1$ or $\beta_3 < 0$, the Hessian will always be non-negative, and therefore $\varphi_{l,a}(.)$ (represented by equation (9)) in the objective function is convex. In a similar way we can show that equations (10) and (13) are also convex. For $\beta_5 \geq 1$ or $\beta_5 < 0$ and positive values of $\propto_3, \gamma_{lp}$ and σ, equation (11) is convex as long as $P_{la}(X_{ls})X_{ls}$ is convex. In our case, $P_{la}(X_{ls})$ is a linear expression with positive coefficients, and therefore $P_{la}(X_{ls})X_{ls}$ is convex. Equation (14) consists of positive constant parameters and equation (15) is zero. Thus, the objective function in the lower level problem is a sum of convex expressions, which indicates that it is convex.

This section discusses the proofs of the proposition in section 4.2.

Proposition 2: Given (c_1, t_1) and (c_M, t_M), we have $D_1^C = D^c(t_1)$ and $D_M^C = D^c(T - t_M)$, where

function $D^c(t) := \frac{1}{3}\pi\bar{v}^2 t^3, \forall t \in [0, \infty)$ (note that $\bar{v}$ and T are given parameters).

Proof. It can be seen from Figure 12 in the manuscript that the lower cone $\mathbf{O}^-_{(c_1,t_1)}$'s height and

the base radius are t_1 and $\bar{v}t_1$, respectively, and therefore the volume of $\mathbf{O}^-_{(c_1,t_1)}$ would be

$D_1^C = \frac{1}{3}\pi(\bar{v}t_1)^2 t_1 = \frac{1}{3}\pi\bar{v}^2 t_1^3 = D^c(t_1)$. For upper cone $\mathbf{O}^+_{(c_M,t_M)}$, again the height and the base

radius are $(T - t_M)$ and $\bar{v}(T - t_M)$, respectively, and as a result,

$D_M^C = \frac{1}{3}\pi(\bar{v}(T - t_M))^2(T - t_M) = \frac{1}{3}\pi\bar{v}^2(T - t_M)^3 = D^c(T - t_M)$. $\square$

Proposition 3: Given (c_1, t_1) and (c_M, t_M), we have $U_1^C = U^c(t_1)$ and $U_M^C = U^c(T - t_M)$, where

function

$$U^c(t) := \int_0^{2\pi} \int_0^{tan^{-1}(\bar{v})} \int_0^{\frac{t}{cos\varphi}} \sqrt{(\rho sin\varphi cos\theta)^2 + (\rho sin\varphi sin\theta)^2}\,\rho^2 sin\varphi \; d\rho d\varphi d\theta, \forall t$$
$$\in [0, \infty).$$

Proof. Essentially, function $U^c(t)$ solves the angular momentum of a cone with a height of t and

a base radius of $t\bar{v}$. We define a spherical coordinate system $\{(\rho, \theta, \varphi)|\rho \geq 0, \theta \in [0, \pi], \varphi \in$

$[0, 2\pi]\}$, where ρ is the radial distance, θ is the polar angle, and φ is the azimuthal angle. Let us

consider a cone in this spherical coordinate system while the cone vertex is placed at the origin

and its axis is on the radius with $\theta = 0$ (so the cone's base is facing up). Note that this spherical

coordinate system is equivalent to the orthogonal coordinate system $\{(x, y, z)|x, y, z \in (-\infty, \infty)\}$,

where $x = \rho sin\varphi cos\theta$, $y = \rho sin\varphi sin\theta$, and $z = \rho cos\varphi$. Then the angular momentum of this cone in the orthogonal coordinate system can be formulated as

$$U^c(t): = \int_0^t \int_{-\bar{v}z}^{\bar{v}z} \int_{-\sqrt{\bar{v}^2 z^2 - x^2}}^{\sqrt{\bar{v}^2 z^2 - x^2}} \sqrt{x^2 + y^2} \; dydxdz \; . \tag{81}$$

We can translate this expression into the spherical coordinate system as follows:

$$U^c(t):$$
$$= \int_0^{2\pi} \int_0^{tan^{-1}(\bar{v})} \int_0^{\frac{t}{cos\varphi}} \sqrt{(\rho sin\varphi cos\theta)^2 + (\rho sin\varphi sin\theta)^2} \rho^2 sin\varphi \; d\rho d\varphi d\theta \; . \tag{82}$$

Note that in equations (17) and (19) in the manuscript, either cone can be rotated and repositioned in the coordinate system as specified in equation (25) and the operations will not change the values of the functions we are looking for. This yields a cone with a height of t_1 for equation (17) or $T - t_M$ for equation (19). Here, in the integrand, $\mathbf{P_S}(t) = (0,0)$ and c is a general point $(x \in (-\bar{v}t, \bar{v}t), y \in (-\sqrt{\bar{v}^2 t^2 - x^2}, \sqrt{\bar{v}^2 t^2 - x^2}))$. Then, apparently $U^c(t_1)$ and $U^c(T - t_M)$ defined in (82) are equivalent to equations (17) and (19), respectively. This completes the proof. □

Proposition 4: Given two consecutive control points (c_{m-1}, t_{m-1}) and (c_m, t_m), $D_m^p = D^p(\|c_m - c_{m-1}\|, |t_m - t_{m-1}|)$, $\forall 2 \leq m \leq M$, where function

$$D^p(c, t) := 2 \int_0^{2\pi} \int_0^{tan^{-1}(\bar{v})} \int_0^{\frac{\bar{v}^2 t^2 - c^2}{(2\bar{v}^2 t)cos\theta - (2c)sin\theta sin\varphi}} \rho^2 sin\varphi \; d\rho d\varphi d\theta, \quad \forall c, t \in [0, \infty).$$

Proof. For the convenience of the presentation, we decompose $c_{m-1} = (x_{m-1}, y_{m-1})$ and $c_m = (x_m, y_m)$. Then we investigate the prism between two control points $(x_{m-1}, y_{m-1}, t_{m-1})$ and (x_m, y_m, t_m). Because shifting and rotating operations will not affect D_m^p, we can obtain the value of D_m^p with the following steps:

Step 1: Shift and rotate the original orthogonal coordinate system to translate points $(x_{m-1}, y_{m-1}, t_{m-1})$ and (x_m, y_m, t_m) into A:= (0,0,0) and B:= $(0, \delta c_m, \delta t_m)$, respectively, where

$$\delta c_m = \|c_m - c_{m-1}\| = \sqrt{(x_m - x_{m-1})^2 + (y_m - y_{m-1})^2} \quad \text{and} \quad \delta t_m = |t_m - t_{m-1}| \quad \text{(Figure}$$

B.1(a)). Note that now both control points are in the $y - t$ plane.

(a)   (b)
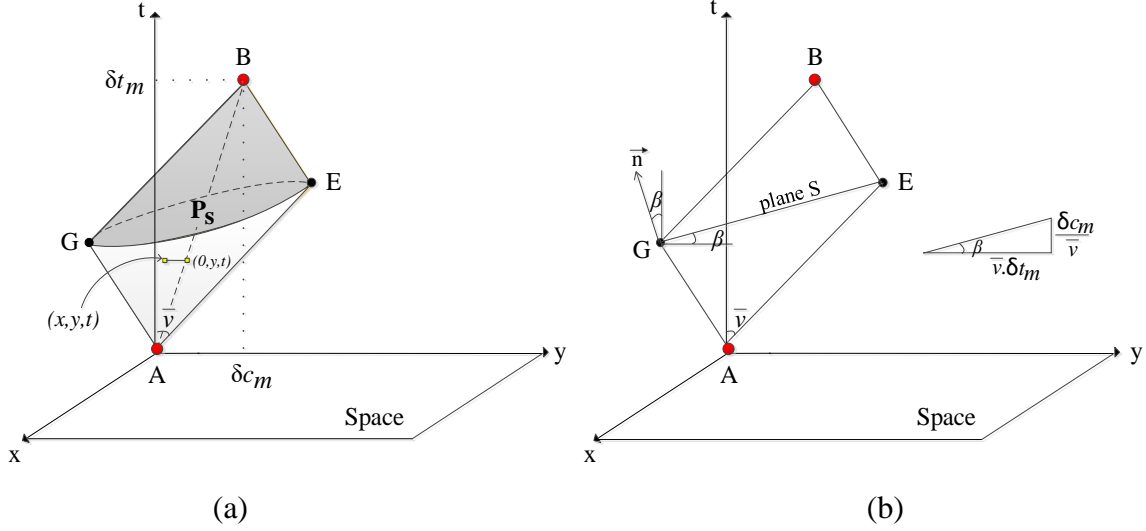
Figure B.1 (a) Space-time prism after shifting and rotating coordinates and (b) position of points in $y - t$ plane and specification of angle β.

Step 2: Find the coordinates of points G and E in $y - t$ plane by crossing the lines AE, BE and

AG, BG as follows (these lines are also in $y - t$ plane):

AE: $y = \bar{v}t$, BE: $y = -\bar{v}t + (\delta c_m + \bar{v}.\delta t_m) \Rightarrow E = \left(0, E_y := \frac{\delta c_m + \bar{v}.\delta t_m}{2}, E_t := \frac{\bar{v}.\delta t_m + \delta c_m}{2\bar{v}}\right)$

AG: $y = -\bar{v}t$, BG: $y = \bar{v}t + (\delta c_m - \bar{v}.\delta t_m) \Rightarrow G = \left(0, G_y := \frac{\delta c_m - \bar{v}.\delta t_m}{2}, G_t := \frac{\bar{v}.\delta t_m - \delta c_m}{2\bar{v}}\right).$

Step 3: Find the equation of plane S, the interphase between the two cones in the prism. This plane

can be formulated with point E and angle β (Figure B.1(b)) as follows:

$$-\sin\beta \left(y - E_y\right) + \cos\beta \left(t - E_t\right) = 0.$$

This indicates that the normal vector of plane S is $(0, -\sin\beta, \cos\beta)$. Again, in the equivalent

spherical coordinate system defined in the above proposition, plane S can be represented as

$$-\sin\beta\left(\rho sin\varphi sin\theta - E_y\right) + \cos\beta\left(\rho cos\varphi - E_t\right) = 0$$

$$\Rightarrow \rho = \frac{\cos\beta\, E_t - \sin\beta\, E_y}{\cos\beta\, cos\theta - \sin\beta\, sin\theta sin\varphi}.$$

Step 4: By substituting β, E_t, and E_y with their respective formulations, the preceding equation of plane S becomes

$$\rho = \frac{\bar{v}^2 \delta t_m{}^2 - \delta c_m{}^2}{(2\bar{v}^2\delta t_m)cos\theta - (2\delta c_m)sin\theta sin\varphi}.$$

Then, using the triple integral in the spherical coordinates, the volume between the plane S and truncated cone in the bottom of $\mathbf{R}_{(c_{m-1},t_{m-1})(c_m,t_m)}$ is

$$V := \int_0^{2\pi}\int_0^{tan^{-1}(\bar{v})}\int_0^{\frac{\bar{v}^2\delta t_m{}^2 - \delta c_m{}^2}{(2\bar{v}^2\delta t_m)cos\theta - (2\delta c_m)sin\theta sin\varphi}} \rho^2 sin\theta\ d\rho d\theta d\varphi.$$

Step 5: Because of symmetry, we obtain $D_m^{\mathrm{p}} = 2V = D^{\mathrm{p}}(\delta c_m, \delta t_m)$, $\forall 2 \le m \le M$. This completes the proof. □

Proposition 5: Given (c_{m-1}, t_{m-1}) and (c_m, t_m), $U_m^{\mathrm{p}} = U^{\mathrm{p}}(\|c_m - c_{m-1}\|, |t_m - t_{m-1}|)$, $\forall 2 \le m \le M$, where

$$U^{\mathrm{p}}(c,t) := 2\int_0^{2\pi}\int_0^{tan^{-1}(\bar{v})}\int_0^{\frac{\bar{v}^2 t^2 - c^2}{(2\bar{v}^2 t)cos\theta - (2c)sin\theta sin\varphi}} Q\rho^2 sin\varphi\ d\rho d\varphi d\theta, \quad \forall c, t \in [0, \infty),$$

and $Q := \sqrt{(\rho sin\varphi cos\theta)^2 + \left(\rho sin\varphi sin\theta - (\frac{c\rho cos\varphi}{t})\right)^2}$.

Proof. This proof follows the same notation defined in Proposition 4. Shift and rotate the prism $\mathbf{R}_{(c_{m-1},t_{m-1})(c_m,t_m)}$ to put it into the position specified in Figure B.1(a). Now the centerline becomes $\mathbf{P_S}(t) = (0, \delta c_m t / \delta t_m)$, and for a generic point $c = (x, y)$ and a generic time t, the integrand of equation (18) in the manuscript can be formulated in the equivalent spherical coordinates as follows:

$$\|c - \boldsymbol{P_S}(t)\| = \sqrt{x^2 + (y - (\frac{\delta c_m t}{\delta t_m}))^2} = \sqrt{(\rho sin\varphi cos\theta)^2 + (\rho sin\varphi sin\theta - (\frac{\delta c_m \rho cos\varphi}{\delta t_m}))^2}\,,$$

which is identical to Q defined in the proposition statement. Therefore, similar to Step 4 in

Proposition 4, we obtain $U_m^{\mathrm{p}} = 2 * \int_0^{2\pi} \int_0^{tan^{-1}(\bar{v})} \int_0^{\frac{\bar{v}^2 \delta t_m^2 - \delta c_m^2}{(2\bar{v}^2 \delta t_m)cos\theta - (2\delta c_m)sin\theta sin\varphi}} Q\rho^2 sin\theta \ d\rho d\theta d\varphi$. As

a result, $U_m^{\mathrm{p}} = U^{\mathrm{p}}(\delta c_m, \delta t_m), \ \forall 2 \le m \le M.$ $\square$

APPENDIX C: GOODNESS OF FIT TEST FOR DIFFERENT MEASURES

To investigate the goodness-of-fit test, we use the Kolmogorov-Smirnov (K-S) test, which simply measures the maximum distance between the cumulative distribution function (CDF) of the data and the fitted power-law distribution. To do that, we need to calculate a KS statistic as follows [146]:

$$KS = max|F - G|, \tag{83}$$

where F is the cumulative distribution of the best fit and G is the cumulative distribution of the synthetic data. The synthetic data are generated from the fitted distribution, and then the best fit for the empirical data can be tested by the KS value. González, Hidalgo, and Barabási (2008) proposed a weighted KS statistic (KS_w) because the regular KS is not very sensitive on the edges of the cumulative distribution. Hence, we also used KS_w, defined as

$$KS_w = max \frac{|F - G|}{\sqrt{G(1 - G)}}. \tag{84}$$

We calculate KS_w for the empirical data and its best fit and compare it with that obtained for 2,000 synthetic data sets generated from the best fit. Empirical data are statistically consistent with their best fit if their KS_w behaves as good as or better than those obtained for the synthetic data. For the goodness-of-fit test for each of the measures, the distribution of KS_w values generated with the synthetic data is compared with the distribution of the ones representing the empirical distribution.

We summarize the goodness-of-fit test by calculating the p-value based on the distribution of KS_w generated with the synthetic data and the value of KS_w representing the empirical distribution. The p-value quantifies the plausibility of the hypothesis. The p-value is defined to be the fraction of the synthetic data values that is larger than the empirical data values. We assume the critical p-value is equal to 0.05; that means if the resulting p-value is greater than 0.05, the power law is a plausible hypothesis for the data; otherwise, it is rejected. Figure C.1 through Figure C.3 show the goodness-of-fit test results for displacement, activity bandwidth, and normalized activity bandwidth, respectively. In all cases we find the p-value is greater than 0.05, and thus the empirical data passes the goodness-of-fit test.
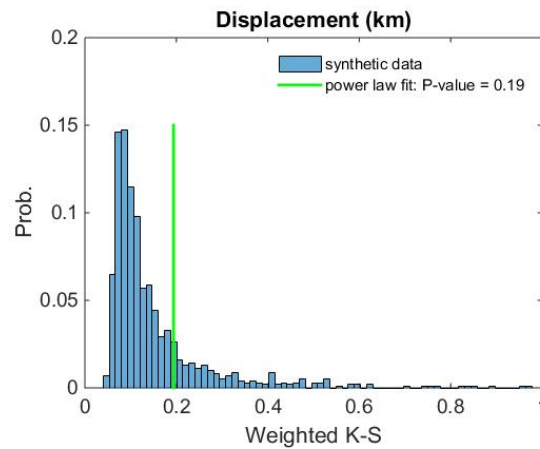
Figure C.1 Goodness-of-fit test based on KS_w for displacement.
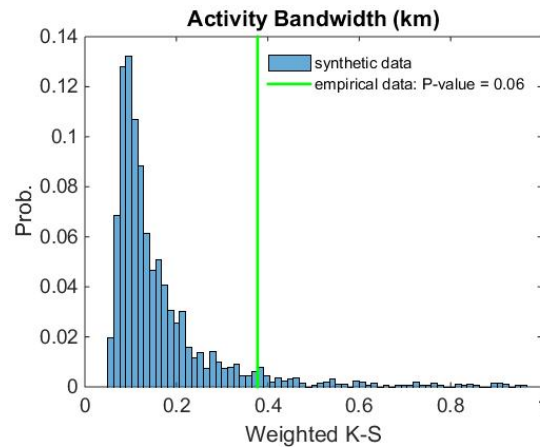
Figure C.2 Goodness-of-fit test based on KS_w for activity bandwidth.

111
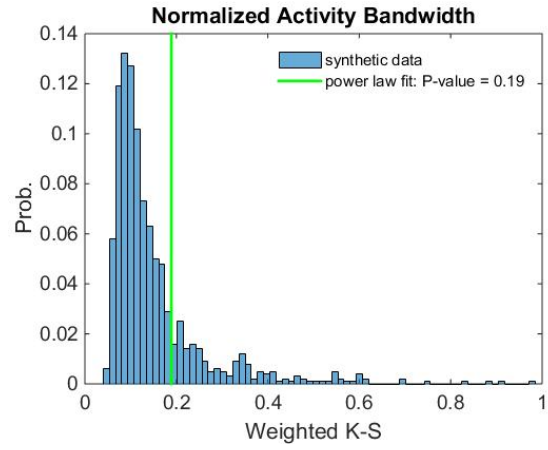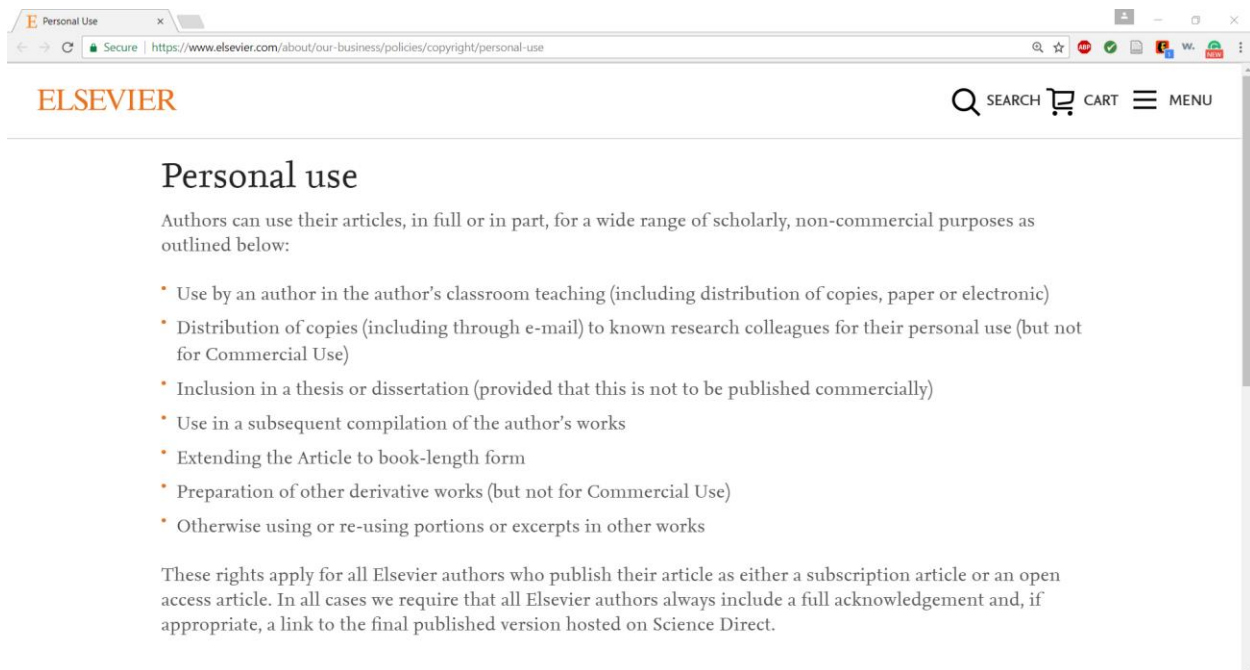
Figure C.3 Goodness-of-fit test based on KS_w for normalized activity bandwidth.

APPENDIX D: COPYRIGHT PERMISSIONS

D.1. Permission from ELSEVIER to Reuse [125],[126] in CHAPTER 3

The permission from ELSEVIER for material in CHAPTER 3 is below.