Error Measures for Trajectory Estimations With Geo-Tagged Mobility Sample Data

Mohsen Parsafard, Guangqing Chi, Xiaobo Qu, Xiaopeng Li, and Haizhong Wang

***Abstract*— Although geo-tagged mobility data (e.g., cell phone data and social media data) can be potentially used to estimate individual space–time travel trajectories, they often have low sample rates that only tell travelers' whereabouts at the sparse sample times while leaving the remaining activities to be estimated with interpolation. This paper proposes a set of time geography-based measures to quantify the accuracy of the trajectory estimation in a robust manner. A series of measures including activity bandwidth and normalized activity bandwidth are proposed to quantify the possible absolute and relative error ranges between the estimated and the ground truth trajectories that cannot be observed. These measures can be used to evaluate the suitability of the estimated individual trajectories from sparsely sampled geo-tagged mobility data for travel mobility analysis. We suggest cutoff values of these measures to separate useful data with low estimation errors and noisy data with high estimation errors. We conduct theoretical analysis to show that these error measures decrease with sample rates and peoples' activity ranges. We also propose a lookup table-based interpolation method to expedite the computational time. The proposed measures have been applied to 2013 geo-tagged tweet data in New York City, USA, and 2014 cell-phone data in Shenzhen, China. The results illustrate that the proposed measures can provide estimation error ranges for exceptionally large datasets in much shorter times than the benchmark method without using lookup tables. These results also reveal managerial results into the quality of these data for human mobility studies, including their distribution patterns.**

***Index Terms*— Geo-tagged data, social media, cellphone, time geography, trajectory estimation, activity range.**

I. Introduction

THE rapid developments of geo-tagged human mobility data, such as cell phone data [1]–[3] and social media data [4]–[6], enable us to investigate space-time travel patterns of humans at the individual trajectory level with high-resolution detail. In general, there are a number of challenges to collecting the ground truth trajectory data, such as sampling frequency and error, user privacy, budget constraints, technological limitation, etc. 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 [1], [5], [7]–[9], travel behaviors, and lifestyles [10], [11] and safety analysis [12]. 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 [13]. For example, many researchers and practitioners have started applying geo-tagged mobility data to provide alternatives for traditional travel mobility surveys [14]–[16]. Such data are also used to characterize traffic flow patterns for better management of road networks [17]. Further, they have been also used to predict real-time travel demand for more responsive and convenient mobility operations [18].

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 target 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 [19]. 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 [20]. Other than the available sample locations, the remaining portions of trajectories are subject to interpolation-based estimation. However, connecting these discrete samples may not always exactly reconstruct the ground-truth trajectories. It is easy to imagine that sparse samples likely yield higher estimation errors. This paper aims to address the granularity issue while the representativeness issue would be a separate topic out of this paper's scope.

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

Manuscript received May 18, 2017; revised March 27, 2018; accepted August 28, 2018. This work was supported in part by the U.S. National Science Foundation under Grant CMMI 1558889, Grant CMMI 1541130, Grant CMMI 1453949, Grant CMMI 1541136, and Grant SES 1823633, and in part by the Eunice Kennedy Shriver National Institute of Child Health and Human Development under Grant P2C HD041025. The Associate Editor for this paper was W. Lin. *(Corresponding author: Xiaopeng Li.)*

M. Parsafard is with Coyote Logistics, Chicago, IL 60647 USA (e-mail: mohsen.parsafard@coyote.com).

G. Chi is with the Department of Agricultural Economics, Sociology, and Education, the Population Research Institute, and the Social Science Research Institute, Pennsylvania State University, University Park, PA 16802 USA (e-mail: gchi@psu.edu).

X. Qu is with the Department of Architecture and Civil Engineering, Chalmers University of Technology, 412 96 Gothenburg, Sweden (e-mail: drxiaoboqu@gmail.com).

X. Li is with the Department of Civil and Environmental Engineering, University of South Florida, Tampa, FL 33620 USA (e-mail: xiaopengli@usf.edu).

H. Wang is with the Department of Civil and Construction Engineering, Oregon State University, Corvallis, OR 97331 USA (e-mail: haizhong.wang@oregonstate.edu).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TITS.2018.2868182

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 [21], [22] uses a space-time prism to bound an object's activity range between two consecutive space-time samples as a prism, i.e., intersection of two cones oriented in opposite directions. This concept is further generalized to incorporate transportation network structures [23]. The second type assumes stochastic underlying patterns of individual movements. With such stochastic settings, developments on probability distributions [24] and stochastic processes [25] can be applied to describe uncertainties of trajectories estimated with geo-tagged sample points.

Despite these developments in modeling trajectory uncertainties, there is 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 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 this gap, this research proposes a set of quantitative measures for the errors 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 error measures to quantify the accuracy of trajectory estimation in a robust manner, one on the absolute errors between the estimated and ground truth trajectories and the other on relative errors with respect to an individual's overall activity area. 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 generally applicable to different types of geo-tagged mobility data. 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. 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 quantify the associated mobility estimation errors 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 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 quality of massive amounts of geo-tagged mobility data and screening useful information from such data for mobility studies.

The organization of this paper is as follows. Section 2 reviews the relevant literature. Section 3 describes some basic concepts of the time geography theory, such as space-time path and space-time prism. Section 4 formulates the proposed measures based on the time geography framework to quantify trajectory bounds. Section 5 presents case studies to illustrate the application of these measures and draw managerial insights with multiple types of geo-tagged mobility data. Section 6 concludes this paper and discusses possible directions for future research.

II. Prior Research

Studies on human mobility and activity patterns have been applied to several fields, including epidemic modeling [26], traffic prediction [3], urban planning [27], [28], and social networks [29], [30]. With the emergence of rapidly growing geo-tagged Big Data from various sources [1], [7], tremendous efforts have been made in the attempt to understand human mobility and activity patterns over time and across space. To name just a few seminal studies, Brockmann *et al.* [31] discovered a power-law distribution of human travel distances from anonymous one-dollar bill transactions. González *et al.* [1] 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 *et al.* [32] 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 [31], [33].

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 [3], sociology [34], and urban and transportation planning [4], [27], [28], [35]. 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 [36]. 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 [6], [35], [37], [38]. 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 [38]–[40].

While studies on theoretical path modeling are mature (e.g., [41], [42]), tracing and predicting an individual's activity patterns using geo-tagged mobility data are still in the exploratory stage. This paper 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 [43]. This concept has been recently applied to transportation network design in innovative ways [44], [45]. 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 basically quantifies an object's activity range given its mobility capability and sometimes geographical barriers as well [21], [22], [46].

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.) [47], [48]. Considering these factors, researchers applied time geography to investigate human activity patterns by integrating time geography concepts with geographic information systems (GIS) [49], [50], three-dimensional geovisualization of activity-travel patterns [51], and some analytical measurement, such as the space-time path, space-time prism, or station [22].

The space-time path and space-time prism are two fundamental concepts in time geography literature, and they serve as the basis for the proposed measures in this study [21], [22], [52]–[54]. The space-time path has been applied to mobile phone logs to study human movement behavior [54] and individual access to urban opportunities [55]. 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 [22], [56], [57].

Recently, probabilistic model 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 [58] and error propagation [59]. Researchers have also developed mathematical foundations for modeling the distribution of visit probabilities within the space-time prisms using the Random Walk theory [60], [61], the truncated Brownian Bridges method [62] and the moment-design method [63].

Previous studies on trajectory analysis have investigated motion pattern description [64], kernel density estimation to interpolates point data to a continuous surface in activity spaces [65], and interpolation methods to investigate the uncertain trajectory of moving objects [66]. If the real human trajectories are available, one can evaluate various interpolation methods such as linear, nearest and cubic interpolations by subsampling data set (see [67]–[69] for more information). 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.

III. Time Geography Review

Since the proposed measures in this study are based on certain time geography concepts, we briefly review them in this section. We consider a time period $\mathbf{T} := [\mathbf{0}, T]$ (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 1a. 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 $(\bar{c}_1, \bar{t}_1)$, $(\bar{c}_2, \bar{t}_2)$, …, $(\bar{c}_N, \bar{t}_N)$ 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:

$$\bar{\boldsymbol{P}}(t) := \begin{cases} \frac{1}{\bar{t}_{i+1} - \bar{t}_i}((\bar{t}_{i+1} - t)\bar{c}_i \\ +(t - \bar{t}_i)\bar{c}_{i+1}), & \text{if } \bar{t}_i \le t \le \bar{t}_{i+1}, 1 \le i \le N; \\ \bar{c}_1, & \text{if } 0 < t < \bar{t}_1; \\ \bar{c}_N, & \text{if } \bar{t}_N < t < T; \end{cases} \quad (1)$$

For the ease of notation, we denote this space-time path by $\bar{\mathbf{P}} := \{\bar{\boldsymbol{P}}(t), \forall t \in \mathbf{T}\}$. 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 geo-tagged mobility data (FIGURE 1b). 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 1b. 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 error 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 2a, 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 he/she travels backward at the maximum speed and $c_i + \bar{v}(t - t_i)$ if he/she travels forward at the

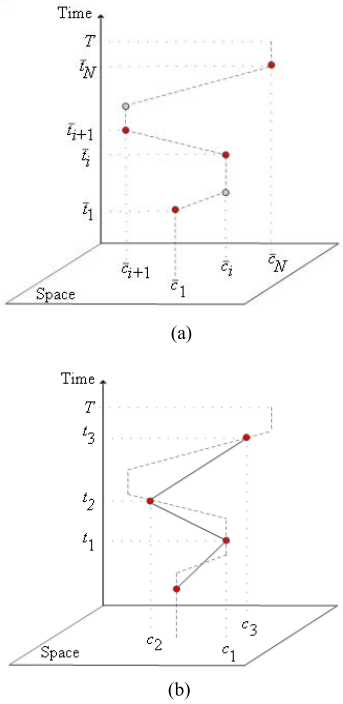


Fig. 1. (a) Space-time path and (b) accessible control points on space-time path.

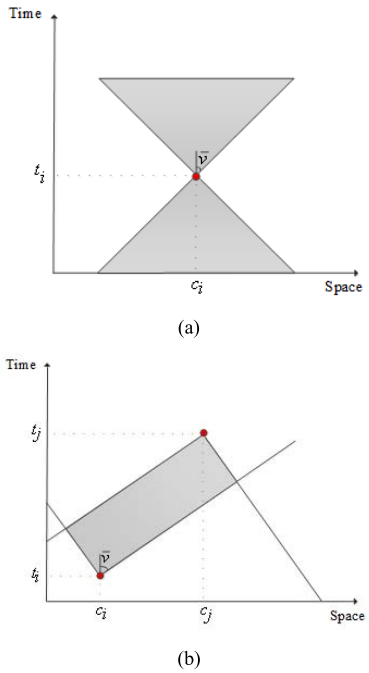


Fig. 2. (a) Space-time cone and (b) space-time prism in one-dimensional space.

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:

$$\mathbf{O}^{+}_{(c_i,t_i)} := \{(c,t)\,||c - c_i| \le \bar{v}\,(t - t_i)\,, t \in [t_i, T]\}. \quad (2)$$

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:

$$\mathbf{O}^{-}_{(c_i,t_i)} := \{(c,t)\,||c - c_i| \le \bar{v}\,(t_i - t)\,, t \in [0, t_i]\}\,. \quad (3)$$

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

$$\begin{aligned}\mathbf{O}_{(c_i,t_i)} &:= \mathbf{O}^{+}_{(c_i,t_i)} \cup \mathbf{O}^{-}_{(c_i,t_i)} \\ &= \{(c,t)\,||c - c_i| \le \bar{v}\,|t_i - t|\,, t \in [0, T]\} \quad (4)\end{aligned}$$

Now suppose that we observe two sample points (c_i, t_i) and (c_j, t_j) of $\bar{\mathbf{P}}$. The space-time range this individual can potentially reach during time period $[t_i, t_j]$ is shown as the shaded area in FIGURE 2b. We call this area a space-time prism, which is essentially the intersection between $\mathbf{O}^{+}_{(c_i,t_i)}$ and $\mathbf{O}^{-}_{(c_j,t_j)}$, i.e.,

$$\begin{aligned}\mathbf{R}_{(c_i,t_i)(c_j,t_j)} := \Big\{ (c,t)\,|\,|c - c_i| \le \bar{v}\,(t - t_i)\,, |c - c_j| \\ \le \bar{v}\,(t_j - t)\,, t \in [t_i, t_j]\Big\}. \quad (5)\end{aligned}$$

These definitions for a one-dimensional space can easily be extended to a two-dimensional space. FIGURE 3a illustrates the space-time cones, and FIGURE 3b shows the space-time prism in a 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 (2)-(5) can be adapted as follows:

$$\mathbf{O}^{+}_{(c_i,t_i)} := \{(c,t)\,|\|c - c_i\| \le \bar{v}\,(t - t_i)\,, t \in [t_i, T]\}\,, \quad (6)$$

$$\mathbf{O}^{-}_{(c_i,t_i)} := \{(c,t)\,|\|c - c_i\| \le \bar{v}\,(t_i - t)\,, t \in [0, t_i]\}\,, \quad (7)$$

$$\mathbf{O}_{(c_i,t_i)} := \{(c,t)\,|\|c - c_i\| \le \bar{v}\,|t_i - t|\,, t \in [0, T]\}\,, \quad (8)$$

$$\begin{aligned}\mathbf{R}_{(c_i,t_i)(c_j,t_j)} := \Big\{ (c,t)\,|\,\|c - c_i\| \le \bar{v}\,(t - t_i)\,, \\ \|c - c_j\| \le \bar{v}\,(t_j - t), t \in [t_i, t_j]\Big\}. \quad (9)\end{aligned}$$

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

IV. Proposed Mobility Measures

A. Activity Bandwidth

Given M consecutive sample points and $\mathbf{S} := \{(c_i, t_i)\}_{i=1,2,\ldots,M}$ along an unknown underlying space-time path $\bar{\mathbf{P}}$ 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 4 (i.e., the series of space-time cones and prisms determined by **S**:, as follows:

$$\begin{aligned}\mathbf{H}\,(\mathbf{S}) := \Big[\mathbf{O}^{-}_{(c_1,t_1)}, \mathbf{R}_{(c_1,t_1)(c_2,t_2)}, \mathbf{R}_{(c_2,t_2)(c_3,t_3)}, \ldots, \\ \mathbf{R}_{(c_{M-1},t_{M-1})(c_M,t_M)}, \mathbf{O}^{+}_{(c_M,t_M)}\Big]. \quad (10)\end{aligned}$$

Note that the size of $\mathbf{H}(\mathbf{S})$ bounds the space-time region for any possible ground truth $\bar{\mathbf{P}}$. Depending on the size of $\mathbf{H}(\mathbf{S})$,

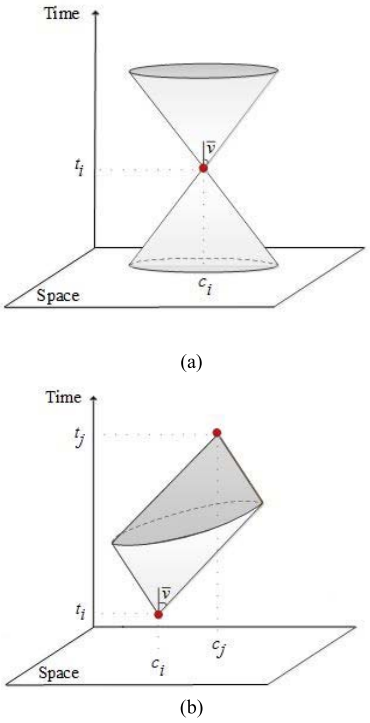


Fig. 3. (a) Space-time cones and (b) space-time prism in two-dimensional space.

there may exist many possible $\bar{\mathbf{P}}$ values falling in the space-time prism, and therefore an estimated path may differ from the ground truth with some error.

As illustrated in FIGURE 4, we define $\mathbf{P_S}$ as the centerline between sample points $\mathbf{S}$ in prism chain $\mathbf{H}(\mathbf{S})$, or, equivalently, the estimated trajectory obtained with $\mathbf{S}$ using linear interpolation[1]:

$$\mathbf{P_S}(t) := \begin{cases} \frac{1}{t_{i+1}-t_i}(t_{i+1}-t)\,c_i \\ \quad +(t-t_i)\,c_{i+1}, & if\ t_i \le t \le t_{i+1}, 1 \le i < M; \\ c_1, & if\ 0 < t < t_1; \\ c_M, & if\ t_M < t < T; \end{cases} \tag{11}$$

If we consider $\mathbf{P_S}$ as an estimated trajectory (which is likely since it is the shortest path and a good approximation of the ground truth), then the estimated error is highly related to the size of prisms in FIGURE 4. Therefore, the amount of error can be quantified with $\bar{v}$, the sample points $\mathbf{S}$, and the prism chain $\mathbf{H}(\mathbf{S})$ (i.e., smaller errors when the prisms are narrower and larger errors when we have a wider prism chain). The objective of this study is to quantify the estimation error. 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 4:

$$B(\mathbf{S}) = \frac{\int_{(c,t)\in\mathbf{H}(\mathbf{S})} \|c - \mathbf{P_S}(t)\|\, dcdt}{\int_{(c,t)\in\mathbf{H}(\mathbf{S})} dcdt} \tag{12}$$

[1]Note that results from other interpolation methods shall also fall within the space-time prism chain. Therefore, the implication of the proposed error bounds is applicable to other interpolation methods as well.

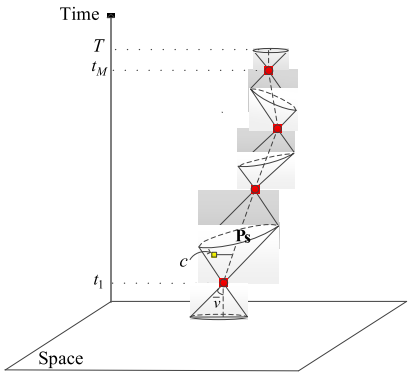


Fig. 4. Space-time prism chain.

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 the integral is held on $(c,t) \in \mathbf{H}(\mathbf{S})$, which is the entire prism chain).

Note that a large $B(\mathbf{S})$ indicates that the error between $\mathbf{P_S}$ and $\bar{\mathbf{P}}$ is likely high, whereas a small $B(\mathbf{S})$ value implies that $\mathbf{P_S}$ is likely close to the $\bar{\mathbf{P}}$. The *activity bandwidth* measures on average how far a point on the prism chain is apart from the center line path. A smaller $B(\mathbf{S})$ value means a narrower activity range (narrower prism chain) and better estimation. In other words, it implies that centerline $\mathbf{P_S}$ is likely close to the ground truth with less estimation error.

Despite the compact formulation, the integral equation (12) 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 the numerator and denominator in equation (12) are decomposed to $M+1$ components, where each component corresponds to a single cone or prism. In other words, $B(\mathbf{S})$ itself can be considered as a simple summation of several integral terms as explained in equation (13).

If U_m^{c} and D_m^{c} $(m = 1, M)$ specify the cone components in the numerator and denominator, respectively, and U_m^{p} and D_m^{p} $(m = 2, 3, \ldots, M)$ correspond to the prism components, then equation (12) could be rewritten as follows:

$$B(\mathbf{S}) = \frac{U_1^{\mathrm{c}} + \sum_{m=2}^{M} U_m^{\mathrm{p}} + U_M^{\mathrm{c}}}{D_1^{\mathrm{c}} + \sum_{m=2}^{M} D_m^{\mathrm{p}} + D_M^{\mathrm{c}}}, \tag{13}$$

where

$$D_1^{\mathrm{c}} = \int_{(c,t)\in\mathbf{O}^-_{(c_1,t_1)}} dcdt, \tag{14}$$

$$D_m^{\mathrm{p}} = \int_{(c,t)\in\mathbf{R}_{(c_{m-1},t_{m-1})(c_m,t_m)}} dcdt, \tag{15}$$

$$D_M^{\mathrm{c}} = \int_{(c,t)\in\mathbf{O}^+_{(c_M,t_M)}} dcdt, \tag{16}$$

and

$$U_1^{\mathrm{c}} = \int_{(c,t)\in \mathbf{O}^-_{(c_1,t_1)}} \|c - \mathbf{P_S}(t)\|\, dcdt, \tag{17}$$

$$U_m^{\mathrm{p}} = \int_{(c,t)\in \mathbf{R}_{(c_{m-1},t_{m-1})(c_m,t_m)}} \|c - \mathbf{P_S}(t)\|\, dcdt, \tag{18}$$

$$U_M^{\mathrm{c}} = \int_{(c,t)\in \mathbf{O}^+_{(c_M,t_M)}} \|c - \mathbf{P_S}(t)\|\, dcdt. \tag{19}$$

In equation (13), D_1^{c} and U_1^{c} are associated with the lower cone $\mathbf{O}^-_{(c_1,t_1)}$, D_M^{c} and U_M^{c} are associated with the upper cone $\mathbf{O}^+_{(c_M,t_M)}$, and D_m^{p} and $U_m^{\mathrm{p}}, \forall m = 2,\ldots,M$ 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 A for the proofs).

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

Proposition 2: Given (c_1, t_1) and (c_M, t_M), we have $U_1^{\mathrm{c}} = U^{\mathrm{c}}(t_1)$ and $U_M^{\mathrm{c}} = U^{\mathrm{c}}(T - t_M)$, where function

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

Proposition 3: Given two consecutive control points (c_{m-1}, t_{m-1}) and (c_m, t_m), $D_m^{\mathrm{p}} = D^{\mathrm{p}}(\|c_m - c_{m-1}\|, |t_m - t_{m-1}|), \forall 2 \le m \le M$, where function

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

Proposition 4: Given (c_{m-1}, t_{m-1}) and (c_m, t_m), $U_m^{\mathrm{p}} = U^{\mathrm{p}}\left(||c_m - c_{m-1}||, |t_m - t_{m-1}|\right), \forall 2 \le m \le M$, where

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

and

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

We see that only function $D^{\mathrm{c}}(t)$ can be solved analytically in a closed form equation defined in Proposition 1, and all other functions defined in Propositions 2–4 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 k^{th} 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

$$\omega_k = \frac{\sqrt{(A_k f_k)}}{\sum_{j\in K}\sqrt{(A_j f_j)}}\Omega, \forall k \in K, \tag{20}$$

where Ω is the total number of ticks selected based on the computational resource. 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 1 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 errors, our case study shows that those errors are well controlled. More details on the lookup table method are provided in the case study section.

Discussion: Note that $B(\mathbf{S})$ is a quantitative measure to judge whether an individual's sample points $\mathbf{S}$ will yield tolerable errors or not in estimating this individual's continuous trajectory. As expected, a threshold (or a cutoff point) is required to differentiate low-error $\mathbf{S}$ vs. high error $\mathbf{S}$. This threshold may differ across different applications depending on the specific error tolerance requirements. In general, we propose the following guideline for the threshold settings. From the perspective of location estimation in relation to the study area, if $B(\mathbf{S})$ is significantly smaller than the radius of the studied area, which we denote by R, say, $B(\mathbf{S})/\mathrm{R} \le 1\%$, then the data

TABLE I
A SCHEMATIC FORM OF LOOKUP TABLE FOR ESTIMATION OF $z = f(c, t)$

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

is very useful for mobility analysis (very good data). If $B(\mathbf{S})$ is somehow smaller than the radius, say, $0.01 < B(\mathbf{S})/\mathrm{R} \leq 0.1$, then the data is still considered useful for mobility analysis (good data). All other data with $B(\mathbf{S})/\mathrm{R} > 0.1$ are not recommended (bad data).

B. Normalized Activity Bandwidth

Activity bandwidth $B(\mathbf{S})$ indicates the absolute error between an estimated path (e.g., centerline $\mathbf{P_S}$) and ground truth path $\bar{\mathbf{P}}$. A relatively small activity bandwidth means that $\mathbf{P_S}$ is likely close to $\bar{\mathbf{P}}$ (note that the ground truth $\bar{\mathbf{P}}$ is not estimated itself, but the objective is to estimate the errors and how $\mathbf{P_S}$ differs from $\bar{\mathbf{P}}$in terms of errors). However, when control points $\mathbf{S}$ are spatially close to each other (clustered around one location e.g. home), even if the absolute activity bandwidth value is small, it is still difficult to discern an individual's activities. In this case, the person has been either stationary or participated in some activities without generating a control point (undiscovered activities). The former information is valuable (we know the person is stationary), however that is not the case in many occasions. It is more likely that we have a person with undiscovered activities even if there are hundreds of control points clustered around one location (e.g. his/her home). In case the clustered points are representing a mixed land use, the individual could conduct a series of different activities that may not be reflected by the clustered control points. Let consider two different individuals with clustered and non-clustered sample points in FIGURE 5.

As illustrated in FIGURE 5, although the activity bandwidth of the chain on the left is smaller than that on the right (which indicates smaller estimation errors), 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. On the other hand, although the activity bandwidth of the chain on the right is larger (which indicates larger estimation errors), 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. In another word, for studying travel patterns and understanding unobserved activities using geo-tagged sample points, the smaller activity bandwidth does not necessarily indicate a better performance. There is a trade-off between the estimation error and detecting individual's participation in more activities.

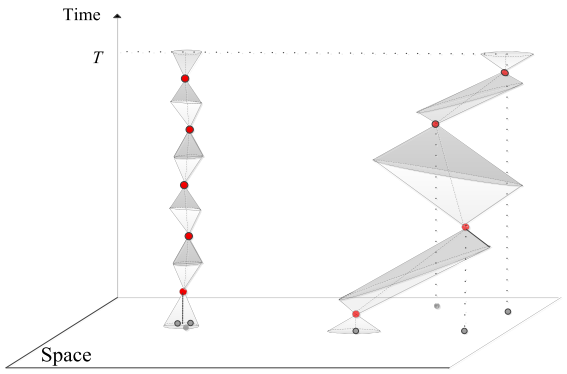


Fig. 5. Space-time prism chain with small and large radii of gyration.

To solve this challenge, we normalize the activity bandwidth by the radius of gyration [70], 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 5), 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:

$$g := \sqrt{\frac{\sum_{m=1}^{M} (c_m - \bar{c})^2}{M}}. \tag{21}$$

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:

$$NB(\mathbf{S}) := \frac{B(\mathbf{S})}{g}, \tag{22}$$

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). In the following case study, we present the proposed measures for a set of collected geo-tagged tweet data (see APPENDIX C for more clarification on *activity bandwidth* and *normalized activity bandwidth*).

Discussion: Note that similar to the $B(\mathbf{S})$, it is practically useful to setup similar thresholds to differentiate data qualities with $NB(\mathbf{S})$ as well. We recommend two threshold values, which may be subject to changes depending on the actual error tolerances in specific applications. From a topological point of view, if $NB(\mathbf{S}) \leq 0.1$ the activity pattern information is much higher than the estimation error and the data is considered very good. If $0.1 \leq NB(\mathbf{S}) \leq 1$, the information is still higher than the error and acceptable (good data). If $NB(\mathbf{S}) > 1$, the error is greater than activity pattern information, and we recommend not to use such data for mobility pattern analysis.

V. Case Studies

This section illustrates applications of the proposed measures with two sets of geotagged mobility data, i.e., geotagged twitter data and cell phone call log data. Each data set includes location samples for a number of individuals over a period of time. We apply the proposed lookup table method to efficiently solve the relevant entries for the corresponding data set. Based on the look-up table results, we solve the proposed trajectory estimation error measures for each individual. We used the proposed cutoff points to classify the data into different categories with different trajectory estimation confidences. Further, we investigate the distribution of these error measures. The results reveal interesting power-law distribution patterns.

A. Geo-Tagged 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. The data set was downloaded from the public data stream provided at https://dev.twitter.com/. 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:

1) 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.
2) 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: $\|c_m - c_{m-1}\| < 0.1$km, $|t_m - t_{m-1}| < 0.01$hr, or $||c_m - c_{m-1}||/|t_m - t_{m-1}|$ is greater than or equal to the corresponding $\bar{v}$.
3) 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.

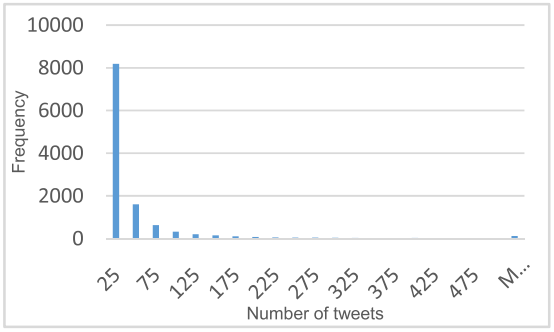


Fig. 6. The distribution of number of tweets.

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 6. 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 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.

To calculate $B(\mathbf{S})$ for each individual, we first prepare the lookup tables for all the terms in equation (13) that do not have closed form formulations (functions defined in Propositions 2–4). 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 1–4, 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 (20), 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 2 illustrates the lookup table for function $U^{\mathrm{c}}(t)$ in Proposition 2, where t is the only variable in this function. Also, TABLE 3 and Table 4 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 2 to TABLE 4 we can see how different terms in $B(\mathbf{S})$ are changing with respect to c or t. If we integrate these

TABLE II
LOOKUP TABLE FOR $U^{c}(t)$

t (hour)	0.205	0.409	0.614	…	455.4	451.95	458
$U^{c}(t)$	1693.77	26836.89	13630 5.2	⋯	3.77E+16	4.00E+16	4.24E+16

TABLE III
LOOKUP TABLE FOR $D^{p}(c,t)$

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

TABLE IV
LOOKUP TABLE FOR $U^{p}(c,t)$

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

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.

To illustrate the efficiency of the proposed lookup tables in calculating $B(\mathbf{S})$, Table 5 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@hyphetagged 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 5, the lookup table approach dramatically reduces the solution time compared to the numerical approach: the ratio of the

TABLE V
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

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).

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:

$$E_r := \frac{|B_{num}(\mathbf{S}) - B_{lookup}(\mathbf{S})|}{B_{num}(\mathbf{S})}. \tag{23}$$

TABLE 5 reports the average E_r value and the 95 percentile 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 5 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.

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 7, for both measures the mean and standard deviation will increase as we increase the $\bar{v}$. In FIGURE 7b the standard deviation of normalized activity bandwidth will increase dramatically by increasing $\bar{v}$. Therefore, ideally, $\bar{v}$ should be selected close to an average individual's mobility speed. If such information is not available, the $\bar{v}$value can be selected on the slightly higher end. This way, the estimated errors are likely upper bounds to the actual

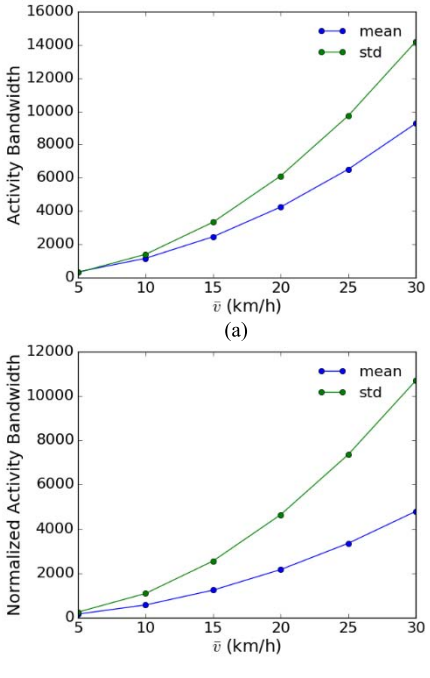


Fig. 7. Sensitivity analyses on $\bar{v}$ for twitter data (a) activity bandwidth (b) normalized activity bandwidth.

errors and this ensures that the data with low estimated errors are with a high probability of good quality.

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 considered as discussed in section IV. According to the results in TABLE 6, 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 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. These findings suggest that although the twitter data contain massive amount of geo-tagged records, only a very small portion of the data have relatively accurate individual mobility information. Therefore, in addition to the representativeness issue, we find that the granularities of most twitter records are too coarse for high-definition travel pattern analysis. Having said this, the small portion of the twitter data with good B and NB values may still contain some useful information on mobility patterns of this particular group of users. Cautions are needed in identifying such useful twitter data depending on the requirements of specific applications.

The following analysis provides a more detailed understanding of the distributions of the proposed measure values in the twitter data. The distribution of $B(\mathbf{S})$ for all $\bar{v}$ values is shown in FIGURE 8. Interestingly, this distribution can be well fit with a power-law distribution $B^{-\beta}$. We used a python package called "power law" [71] for fitting the power-law distributions and compare it with the other alternative heavy

TABLE VI

PERCENTAGE OF TWITTER 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
$\frac{B(\mathbf{S})}{R_{NYC}} \leq 0.01$	0.02	0	0
$0.01 \leq \frac{B(\mathbf{S})}{R_{NYC}} \leq 0.1$	0.41	0.5	0.7
$\frac{B(\mathbf{S})}{R_{NYC}} > 0.1$	99.57	99.5	99.3

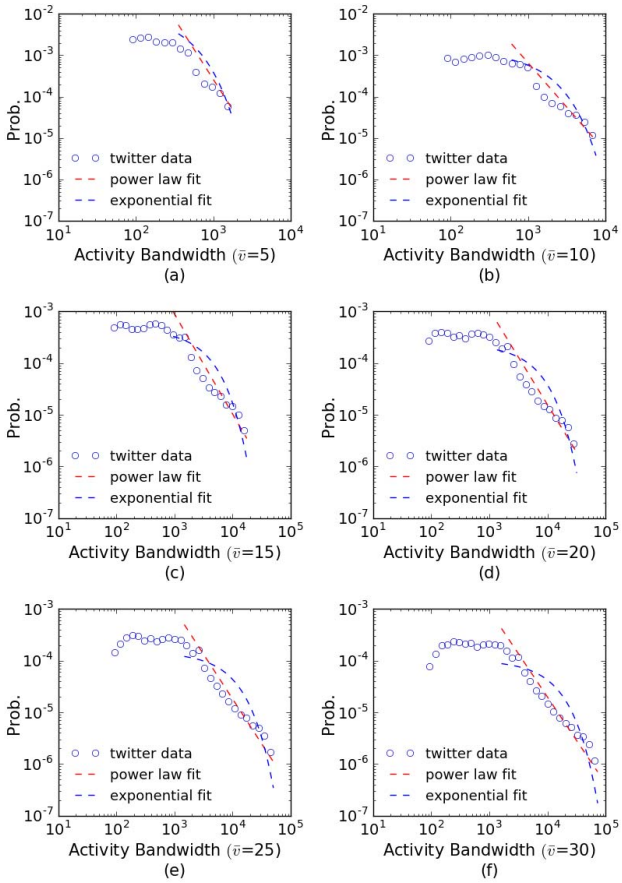


Fig. 8. Power-law and exponential distribution fit for activity bandwidth for twitter data with different $\bar{v}$ (km/h): (a) $\bar{v} = 5$ (b) $\bar{v} = 10$ (c) $\bar{v} = 15$ (d) $\bar{v} = 20$ (e) $\bar{v} = 25$ (f) $\bar{v} = 30$.

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 log likelihood 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 B). This finding reveals an interesting pattern about how frequently people travel and tweet. The long tail of this power-law

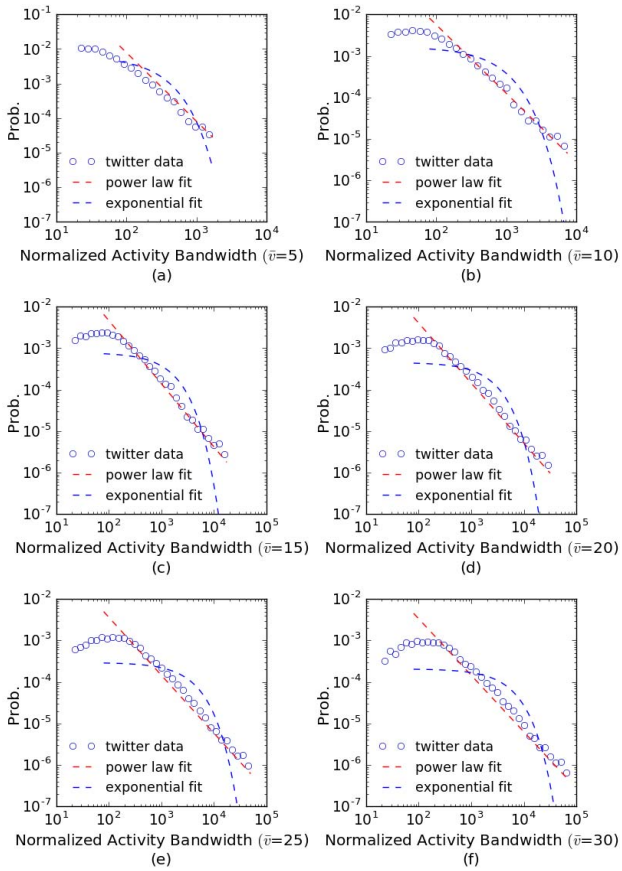


Fig. 9. Power-law and exponential distribution fit for normalized activity bandwidth for twitter data with different $\bar{v}$ (km/h): (a) $\bar{v} = 5$ (b) $\bar{v} = 10$ (c) $\bar{v} = 15$ (d) $\bar{v} = 20$ (e) $\bar{v} = 25$ (f) $\bar{v} = 30$.

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 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.

With the gyration values calculated in equation (21), we calibrate the preceding activity bandwidths to the normalized activity bandwidths using equation (22). V-B 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, and there only remains a small portion 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.

TABLE VII
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
$\frac{B(\mathbf{S})}{R_S} \leq 0.01$	0.002	0.001	0.001
$0.01 \leq \frac{B(\mathbf{S})}{R_S} \leq 0.1$	0.094	0.103	0.978
$\frac{B(\mathbf{S})}{R_S} > 0.1$	99.904	99.896	99.021

B. 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 $U^{\mathrm{c}}(t)$, $D^{\mathrm{p}}(c,t)$ and $U^{\mathrm{p}}(c,t)$. The lookup tables for cellphone data are alike Twitter data using the same parameters (see Table 2 to TABLE 4 for more details), except for ticks on c and t that are distributed using equation (20). 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. Similar cutoff point analysis is presented in Table 7. 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).

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 10 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, VI and FIGURE 12 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

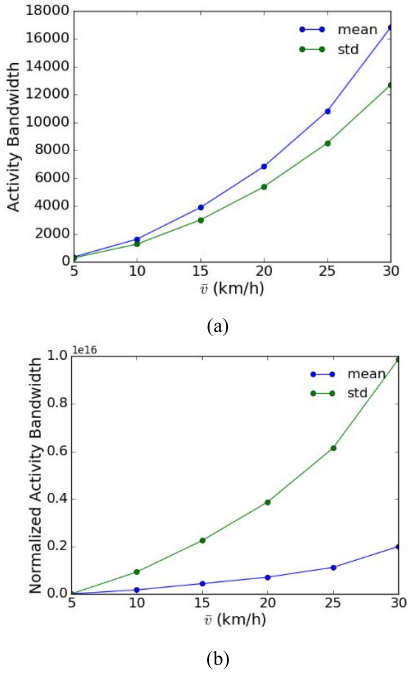


Fig. 10. Sensitivity analyses on $\bar{v}$ for cellphone data (a) activity bandwidth (b) normalized activity bandwidth.

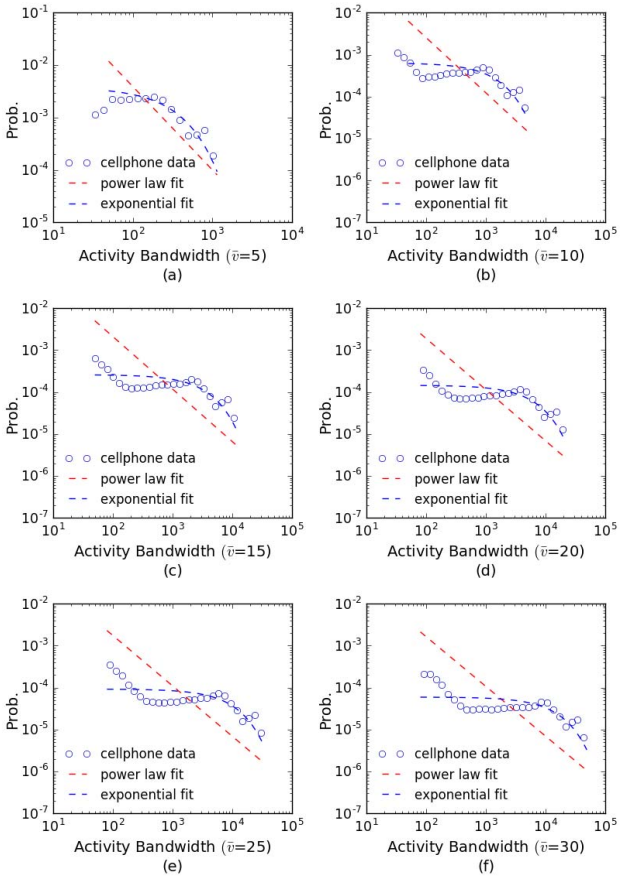


Fig. 11. Power-law and exponential distribution fit for activity bandwidth for cellphone data with different $\bar{v}$ (km/h): (a) $\bar{v} = 5$ (b) $\bar{v} = 10$ (c) $\bar{v} = 15$ (d) $\bar{v} = 20$ (e) $\bar{v} = 25$ (f) $\bar{v} = 30$.

Twitter data, where a small portion of individuals around the head of the distribution that can provide relatively accurate mobility pattern information.

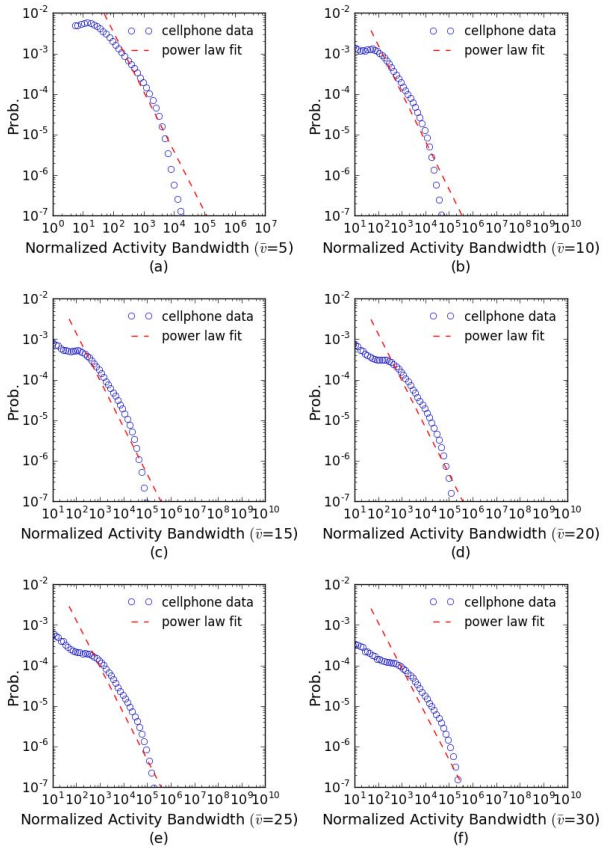


Fig. 12. Power-law distribution fit for normalized activity bandwidth for cellphone data with different $\bar{v}$ (km/h): (a) $\bar{v} = 5$ (b) $\bar{v} = 10$ (c) $\bar{v} = 15$ (d) $\bar{v} = 20$ (e) $\bar{v} = 25$ (f) $\bar{v} = 30$.

VI. Conclusion

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 the unknown ground truth trajectory, a set of fundamental measures are proposed to quantify the accuracy of the estimation error in a robust manner. The estimation error 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 error 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 errors 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 errors 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 error 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. In particular, it will be interesting to compare the proposed error bounds with actual errors when ground truth trajectory data are available. While the measured geo-coordinates have significant errors, this proposed methodology needs to be properly calibrated to account for such sampling data noise.

APPENDIX A

This section discusses the proofs of the proposition in Section 4.

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

Proof: It can be seen from FIGURE 4 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

$$D_1^{\mathrm{c}} = \frac{1}{3}\pi(\bar{v}t_1)^2 t_1 = \frac{1}{3}\pi\bar{v}^2 t_1^3 = D^{\mathrm{c}}(t_1).$$

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,

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

□

Proposition 2: Given (c_1, t_1) and (c_M, t_M), we have $U_1^{\mathrm{c}} = U^{\mathrm{c}}(t_1)$ and $U_M^{\mathrm{c}} = U^{\mathrm{c}}(T - t_M)$, where function

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

Proof: Essentially, function $U^{\mathrm{c}}(t)$ 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

$$U^{\mathrm{c}}(t) := \int_0^t \int_{-\bar{v}z}^{\bar{v}z} \int_{-\sqrt{\bar{v}^2z^2 - x^2}}^{\sqrt{\bar{v}^2z^2 - x^2}} \sqrt{x^2 + y^2}\, dy dx dz. \quad (24)$$

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

$$U^{\mathrm{c}}(t) := \int_0^{2\pi}\int_0^{tan^{-1}(\bar{v})}\int_0^{\frac{t}{cos\varphi}} \times \sqrt{(\rho sin\varphi cos\theta)^2 + (\rho sin\varphi sin\theta)^2}\rho^2 sin\varphi d\rho d\varphi d\theta. \quad (25)$$

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 $(x \in (-\bar{v}t, \bar{v}t), y \in (-\sqrt{\bar{v}^2t^2 - x^2}, \sqrt{\bar{v}^2t^2 - x^2}))$. Then, apparently $U^{\mathrm{c}}(t_1)$ and $U^{\mathrm{c}}(T - t_M)$ defined in (25) are equivalent to equations (17) and (19), respectively. This completes the proof. □

Proposition 3: Given two consecutive control points (c_{m-1}, t_{m-1}) and (c_m, t_m), $D_m^{\mathrm{p}} = D^{\mathrm{p}}(\|c_m - c_{m-1}\|, |t_m - t_{m-1}|), \forall 2 \leq m \leq M$, where function

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

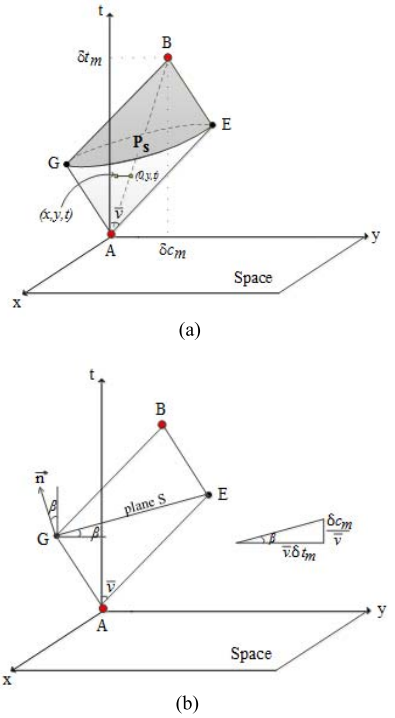


Fig. 13. (a) Space-time prism after shifting and rotating coordinates and (b) position of points in $y-t$ plane and specification of angle β.

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 $\delta c_m = \|c_m - c_{m-1}\| = \sqrt{(x_m - x_{m-1})^2 + (y_m - y_{m-1})^2}$ and $\delta t_m = |t_m - t_{m-1}|$ (FIGURE 13a). Note that now both control points are in the $y-t$ plane.

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):

$$\begin{aligned}
\text{AE:}y &= \bar{v}t,\ \text{BE:}y = -\bar{v}t + (\delta c_m + \bar{v}.\delta t_m) \\
&\Rightarrow \mathrm{E} = \left(0, \mathrm{E}_y := \frac{\delta c_m + \bar{v}.\delta t_m}{2}, \mathrm{E}_t := \frac{\bar{v}.\delta t_m + \delta c_m}{2\bar{v}}\right) \\
\text{AG:y} &= -\bar{v}\text{t},\ \text{BG:y} = \bar{v}\text{t} + (\delta c_m - \bar{v}.\delta t_m) \\
&\Rightarrow \mathrm{G} = (0, \mathrm{G}_y := \frac{\delta c_m - \bar{v}.\delta t_m}{2}, \mathrm{G}_t := \frac{\bar{v}.\delta t_m - \delta c_m}{2\bar{v}}).
\end{aligned}$$

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 13b) as follows:

$$-\sin\beta\left(y - \mathrm{E}_y\right) + \cos\beta\left(t - \mathrm{E}_t\right) = 0.$$

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

$$\begin{aligned}
&-\sin\beta\left(\rho sin\varphi sin\theta - \mathrm{E}_y\right) + \cos\beta\left(\rho cos\varphi - \mathrm{E}_t\right) = 0 \\
&\Rightarrow \rho = \frac{\cos\beta \mathrm{E}_t - \sin\beta \mathrm{E}_y}{\cos\beta cos\theta - \sin\beta sin\theta sin\varphi}.
\end{aligned}$$

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

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

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

Proof: This proof follows the same notation defined in Proposition 3. 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 13a. 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:

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

which is identical to Q defined in the proposition statement. Therefore, similar to Step 4 in Proposition 3, 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, $U_m^{\mathrm{p}} = U^{\mathrm{p}}(\delta c_m, \delta t_m), \forall 2 \le m \le M$. □

APPENDIX B

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 [72]:

$$KS = \max|\mathrm{F} - \mathrm{G}|, \tag{26}$$

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

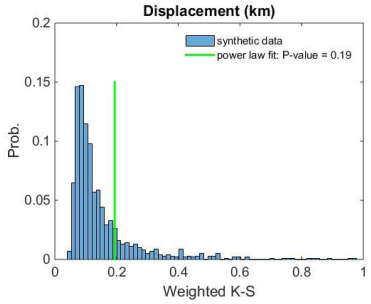


Fig. 14. Goodness-of-fit test based on KS_w for displacement.

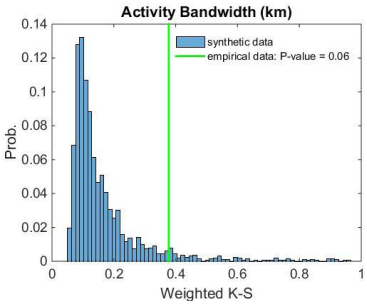


Fig. 15. Goodness-of-fit test based on KS_w for activity bandwidth.

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

$$KS_w = max \frac{|F - G|}{\sqrt{G\,(1 - G)}}. \tag{27}$$

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 14 through FIGURE 16 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.

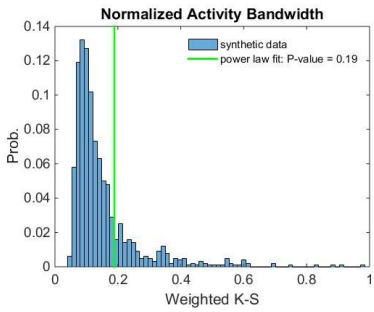


Fig. 16. Goodness-of-fit test based on KS_w for normalized activity bandwidth.

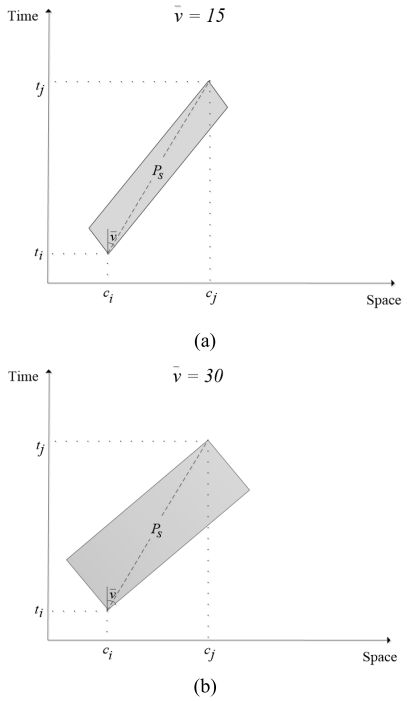


Fig. 17. (a) Space-time prism, $\bar{v} = 15$ and (b) space-time prism, $\bar{v} = 30$.

APPENDIX C

Let consider a simple example in a one-dimensional space with two sample points (c_i, t_i) and (c_j, t_j). Using two different speed limits (e.g. $\bar{v} = 15$ versus $\bar{v} = 30$) we can draw the prisms for these sample points as shown in FIGURE 17. Based on our discussion in section III, there are many feasible space-time paths falling in the space-time prism between these two sample points, and of course an estimated path may differ from the ground truth with some error. In general, the estimated trajectories in FIGURE 17a have relatively smaller errors because of narrower prisms compared with those in FIGURE 17b. To this end, the objective of this study is to quantify this error with the proposed measures, e.g. a*ctivity bandwidth*and *normalized activity bandwidth*. The *activity bandwidth* measures on average how far a point on the prism chain is apart from the center line path. A smaller $B(\mathbf{S})$ value means a narrower activity range (narrower prism chain) and better estimation. In other words, it implies that the centerline $\mathbf{P_S}$ is likely close to the ground truth with less estimation error.

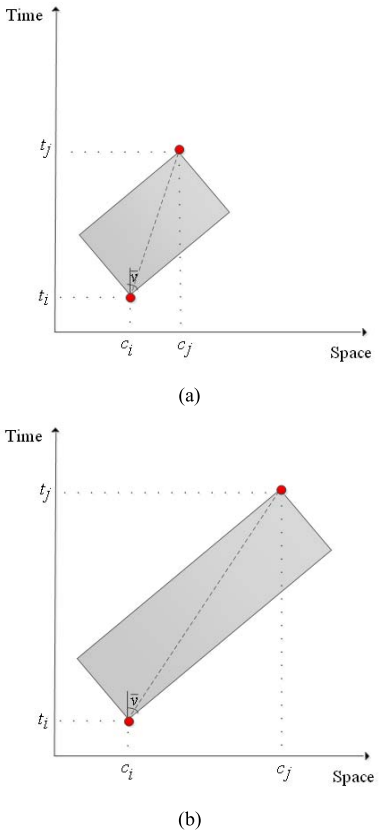


Fig. 18. Two trajectories with different t and c, but equal $B(\mathbf{S})$, (a) short trajectory, (b) long trajectory.

To clarify the normalized activity bandwidth, let consider two trajectories with different t and c in a one dimensional space as shown in FIGURE 18, and let assume they have the same $B(\mathbf{S})$. For mobility analysis, we prefer the trajectory in FIGURE 18b because of its potentials for capturing unobserved activities. Since both trajectories in FIGURE 18a and FIGURE 18b have the same $B(\mathbf{S})$, normalized activity bandwidth is defined to distinguish between these trajectories. In other words, there is a trade-off between the estimation error and detecting individual's participation in more activities. To address this trade-off, we normalize the activity bandwidth by the radius of gyration, which measures the spread of an individual's locations around his/her center of mass.

REFERENCES

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

[2] J. C. Herrera, D. B. Work, R. Herring, X. Ban, Q. Jacobsond, and A. M. Bayen, "Evaluation of traffic data obtained via GPS-enabled mobile phones: The mobile century field experiment," *Transp. Res. C, Emerg. Technol.*, vol. 18, no. 4, pp. 568–583, Aug. 2010.

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

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

[5] 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 *Proc. 2nd ACM SIGKDD Int. Workshop Urban Comput.*, 2013, Art. no. 6.

[6] S. Scellato, A. Noulas, R. Lambiotte, and C. Mascolo, "Socio-spatial properties of online location-based social networks," in *Proc. ICWSM*, vol. 11, 2011, pp. 329–336.

[7] S. Hasan, C. Schneider, S. V. Ukkusuri, and M. C. González, "Spatiotemporal patterns of urban human mobility," *J. Statist. Phys.*, vol. 151, nos. 1–2, pp. 304–318, Apr. 2013.

[8] 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.

[9] T. M. T. Do and D. Gatica-Perez, "Contextual conditional models for smartphone-based human mobility prediction," in *Proc. ACM Conf. Ubiquitous Comput.*, 2012, pp. 163–172.

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

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

[12] 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. C, Emerg. Technol.*, vol. 50, pp. 13–27, Jan. 2015.

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

[14] F. Calabrese, M. Diao, G. Di Lorenzo, J. Ferreira, and C. Ratti, "Understanding individual mobility patterns from urban sensing data: A mobile phone trace example," *Transp. Res. C, Emerg. Technol.*, vol. 26, pp. 301–313, Jan. 2013.

[15] H. Park and A. Haghani, "Optimal number and location of Bluetooth sensors considering stochastic travel time prediction," *Transp. Res. C, Emerg. Technol.*, vol. 55, pp. 203–216, Jun. 2015.

[16] P. Stopher, C. FitzGerald, and J. Zhang, "Search for a global positioning system device to measure person travel," *Transp. Res. C, Emerg. Technol.*, vol. 16, no. 3, pp. 350–369, 2008.

[17] J. Kim and H. S. Mahmassani, "Trajectory clustering for discovering spatial traffic flow patterns in road networks," in *Proc. Transp. Res. Board 94th Annu. Meeting*, 2015, Paper 15-5443.

[18] J. Yuan, Y. Zheng, L. Zhang, X. Xie, and G. Sun, "Where to find my next passenger," in *Proc. 13th Int. Conf. Ubiquitous Comput.*, 2011, pp. 109–118.

[19] 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.

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

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

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

[23] 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.

[24] 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, Sep. 2004.

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

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

[27] X. Lu, E. Wetter, N. Bharti, A. J. Tatem, and L. Bengtsson, "Approaching the limit of predictability in human mobility," *Sci. Rep.*, vol. 3, Oct. 2013, Art. no. 2923.

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

[29] J. K. Hackney, "Coevolving social and transportation networks," in *Arbeitsbericht Verkehrs- und Raumplanung*, vol. 335. 2005.

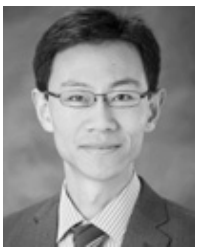
[30] K. W. Axhausen, "Social networks, mobility biographies, and travel: Survey challenges," *Environ. Planning B, Urban Anal. City Sci.*, vol. 35, no. 6, pp. 981–996, 2008.

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

[32] 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.

[33] 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.
[34] T. Arentze and H. Timmermans, "Social networks, social interactions, and activity-travel behavior: A framework for microsimulation," *Environ. Planning B, Urban Anal. City Sci.*, vol. 35, no. 6, pp. 1012–1027, 2008.
[35] S. Hasan, "Modeling urban mobility dynamics using geo-location data," Ph.D. dissertation, Lyles School Civil Eng., Purdue Univ., West Lafayette, IN, USA, 2013.
[36] P. Symeonidis, N. Dimitrios, and Y. Manolopoulos, "Location-based social networks," in *Recommender Systems for Location-based Social Networks*. New York, NY, USA: Springer, 2014, pp. 35–48.
[37] H. Gao, J. Tang, and H. Liu, "Exploring social-historical ties on location-based social networks," in *Proc. ICWSM*, 2012, pp. 114–121.
[38] S. Hasan and S. V. Ukkusuri, "Urban activity pattern classification using topic models from online geo-location data," *Transp. Res. C, Emerg. Technol.*, vol. 44, pp. 363–381, Jul. 2014.
[39] S. Bregman and K. E. Watkins, *Best Practices for Transportation Agency Use of Social Media*. Boca Raton, FL, USA: CRC Press, 2013.
[40] P. J. H. Daas, M. J. Puts, B. Buelens, and P. A. M. van den Hurk, "Big data and official statistics," *J. Off. Statist.*, vol. 31, no. 2, pp. 249–262, 2015.
[41] 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*, vol. 2338, no. 1, pp. 58–66, 2013.
[42] Y. M. Nie and X. Wu, "Shortest path problem considering on-time arrival probability," *Transp. Res. B, Methodol.*, vol. 43, no. 6, pp. 597–613, 2009.
[43] H. J. Miller, "Necessary space—Time conditions for human interaction," *Environ. Planning B, Urban Anal. City Sci.*, vol. 32, no. 3, pp. 381–401, 2005.
[44] 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. C, Emerg. Technol.*, vol. 66, pp. 176–194, May 2016.
[45] L. Tong, X. Zhou, and H. J. Miller, "Transportation network design for maximizing space–time accessibility," *Transp. Res. B, Methodol.*, vol. 81, pp. 555–576, Nov. 2015.
[46] N. J. Thrift, *An Introduction to Time Geography*," Norwich, U.K.: Geo Abstracts, Univ. East Anglia, 1977.
[47] R. G. Golledge, *Spatial Behavior: A Geographic Perspective*. New York, NY, USA: Guilford Press, 1997.
[48] 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*. Dordrecht, The Netherlands: Springer, 2007, pp. 103–118.
[49] H. Yu and S.-L. Shaw, "Exploring potential human activities in physical and virtual spaces: A spatio-temporal GIS approach," *Int. J. Geograph. Inf. Sci.*, vol. 22, no. 4, pp. 409–430, 2008.
[50] 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.
[51] 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. C, Emerg. Technol.*, vol. 8, nos. 1–6, pp. 185–203, 2000.
[52] 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.
[53] H. J. Miller and S. A. Bridwell, "A field-based theory for time geography," *Ann. Assoc. Amer. Geogr.*, vol. 99, no. 1, pp. 49–75, 2009.
[54] 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 *Proc. 18th Int. Conf. Geoinform.*, 2010, pp. 1–6.
[55] 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.
[56] H. J. Miller, "Measuring space-time accessibility benefits within transportation networks: Basic theory and computational procedures," *Geogr. Anal.*, vol. 31, no. 1, pp. 187–212, 1999.
[57] 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. Geograph. Syst.*, vol. 5, no. 1, pp. 71–91, 2003.
[58] 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. Geograph. Syst.*, vol. 6, no. 4, pp. 325–354, 2004.
[59] T. Kobayashi, H. J. Miller, and W. Othman, "Analytical methods for error propagation in planar space–time prisms," *J. Geograph. Syst.*, vol. 13, no. 4, pp. 327–354, 2011.
[60] S. Winter and Z.-C. Yin, "The elements of probabilistic time geography," *GeoInformatica*, vol. 15, no. 3, pp. 417–434, 2011.
[61] S. Winter and Z.-C. Yin, "Directed movements in probabilistic time geography," *Int. J. Geograph. Inf. Sci.*, vol. 24, no. 9, pp. 1349–1365, 2010.
[62] Y. Song and H. J. Miller, "Simulating visit probability distributions within planar space–time prisms," *Int. J. Geograph. Inf. Sci.*, vol. 28, no. 1, pp. 104–125, 2014.
[63] Y. Leung, Z. Zhao, and J.-H. Ma, "Uncertainty analysis of space–time prisms based on the moment-design method," *Int. J. Geograph. Inf. Sci.*, vol. 30, no. 7, pp. 1336–1358, 2016.
[64] P. Laube, S. Imfeld, and R. Weibel, "Discovering relative motion patterns in groups of moving point objects," *Int. J. Geograph. Inf. Sci.*, vol. 19, no. 6, pp. 639–668, 2005.
[65] Z. Patterson and S. Farber, "Potential path areas and activity spaces in application: A review," *Transp. Rev.*, vol. 35, no. 6, pp. 679–700, 2015.
[66] G. Trajcevski, O. Wolfson, K. Hinrichs, and S. Chamberlain, "Managing uncertainty in moving objects databases," *ACM Trans. Database Syst.*, vol. 29, no. 3, pp. 463–507, Sep. 2004.
[67] S. Hoteit, S. Secci, S. Sobolevsky, C. Ratti, and G. Pujolle, "Estimating human trajectories and hotspots through mobile phone data," *Comput. Netw.*, vol. 64, pp. 296–307, May 2014.
[68] S. Hoteit, S. Secci, and M. Premoli, "Crowded spot estimator for urban cellular networks," *Ann. Telecommun.*, vol. 72, nos. 11–12, pp. 743–754, 2017.
[69] G. Chen, S. Hoteit, A. C. Viana, M. Fiore, and C. Sarraute, "Individual trajectory reconstruction from mobile network data," Ph.D. dissertation, INRIA Saclay-Île-de-France, Palaiseau, France, 2018.
[70] Z. Cheng, J. Caverlee, K. Lee, and D. Z. Sui, "Exploring millions of footprints in location sharing services," in *Proc. ICWSM*, 2011, pp. 81–88.
[71] 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.
[72] A. Clauset, C. R. Shalizi, and M. E. J. Newman, "Power-law distributions in empirical data," *SIAM Rev.*, vol. 51, no. 4, pp. 661–703, 2009.

Mohsen Parsafard received the M.S. degree in civil engineering from the Sharif University of Technology, Iran and the Ph.D. degree in civil engineering from the University of South Florida. He is currently a Research Scientist at Coyote Logistics. His research interests include transportation network design, mobility pattern analysis, supply chain, and logistics. His awards and honors include the Anne Brewer Scholarship from the Intelligent Transportation Society of Florida, Gulf Region Intelligent Transportation Society Scholarships, the Research Week Poster Award from the College of Engineering, University of South Florida, the Civil and Environmental Engineering Departmental Scholarship, and a few travel awards.



Guangqing Chi received the Ph.D. degree in environmental demography from the University of Wisconsin–Madison and the M.S. degree in environmental policy from Michigan Technological University. He is currently an Associate Professor of rural sociology and demography and the Director of the Computational and Spatial Analysis Core, Pennsylvania State University. His research examines the interactions between population health and the built and natural environments. He pursues his research program within interwoven research projects on climate change, land use, and community resilience, with an emphasis on critical infrastructure/transportation and population change within the smart cities framework. He has studied issues of generalizability and reproducibility for use of Twitter data in population health and social science research.

Xiaobo Qu received the bachelor's degree from the National University of Singapore, the master's degree from Tsinghua University, and the Ph.D. degree from Jilin University. He has served on the Faculty of the University of Technology Sydney and Griffith University. He is currently a Professor and the Chair of urban mobility systems with the Department of Architecture and Civil Engineering, Chalmers University of Technology. His research is focused on practically improving transport safety, efficiency, equity, and sustainability through traffic flow and network modeling and optimization.

Xiaopeng Li received the B.S. degree in civil engineering from Tsinghua University, Beijing, China, in 2006, and the M.S. degrees in civil engineering and applied mathematics and the Ph.D. degree in civil engineering from the University of Illinois at Urban–Champaign, Urbana, IL, USA, in 2007, 2010, and 2011, respectively. He is currently an Associate Professor and Susan A. Bracken Faculty Fellow with the Department of Civil and Environmental Engineering, University of South Florida, Tampa, FL, USA. His recent research focuses on modeling and decision-making for emerging transportation technologies and services with trajectory optimization, traffic flow analysis, network modeling, continuum approximation, and machine learning methods. He received the U.S. National Science Foundation CAREER Award. He serves as an Associate Editor for the Department of Transportation Systems Analysis for *IIE Transactions*, and he is on the editorial boards of *Transportation Research Part B* and *Transportation Research Part C*.

Haizhong Wang received the B.S. degree from the Hebei University of Technology, the M.S. degree from the Beijing University of Technology, China, and the M.S. degree in applied mathematics and the Ph.D. degree in civil engineering (transportation) from the University of Massachusetts Amherst, Amherst, MA, USA. He is currently an Associate Professor with the School of Civil and Construction Engineering, Oregon State University, Corvallis, OR, USA. His research goal is to advance the theoretical and practical understanding of how social, natural, and engineered systems interact to enhance human life safety and infrastructure network resilience in normative and disruptive scenarios. His recent research interests focus on interdisciplinary agent-based evacuation modeling for rapid-onset post-disaster emergent mobility and life safety under social-technical network disruptions; stochastic traffic flow modeling in a mixed connected and automated vehicle environment under varying market penetrations; and a network-of-networks approach to critical interdependent lifeline infrastructure network resilience and robustness through a percolation theoretical modeling framework.