Article

Exploring Multi-Scale Spatiotemporal Twitter User Mobility Patterns with a Visual-Analytics Approach

Junjun Yin [1] and Zhenhong Du [2,*]

[1] Department of Geography and Geographic Information Science, University of Illinois at Urbana-Champaign, Urbana, IL 61801, USA ; jyn@illinois.edu

[2] Key Laboratory of Geographic Information Science, School of Earth Sciences, Zhejiang University, Hangzhou, 310028, China; duzhenhong@zju.edu.cn

* Correspondence: duzhenhong@zju.edu.cn; Tel.: +8613819192989

Academic Editor: name

Version August 14, 2016 submitted to ISPRS Int. J. Geo-Inf.; Typeset by LaTeX using class file mdpi.cls

Abstract: Understanding human mobility patterns is of great importance for urban planning, traffic management, and even marketing campaign. However, the capability of capturing detailed human movements with fine-grained spatial and temporal granularity is still limited. In this study, we extracted high-resolution mobility data from a collection of over 1.3 billion geo-located Twitter messages. Regarding the concerns of infringement on individual privacy, such as the mobile phone call records with restricted access, the dataset is collected from publicly accessible Twitter data streams. In this paper, we employed a visual-analytics approach to studying multi-scale spatiotemporal Twitter user mobility patterns in the contiguous United States during the year 2014. Our approach included a scalable visual-analytics framework to deliver efficiency and scalability in filtering large volume of geo-located tweets, modeling and extracting Twitter user movements, generating space-time user trajectories, and summarizing multi-scale spatiotemporal user mobility patterns. We performed a set of statistical analysis to understand Twitter user mobility patterns across multi-level spatial scales and temporal ranges. In particular, Twitter user mobility patterns measured by the displacements and radius of gyrations of individuals revealed multi-scale or multi-modal Twitter user mobility patterns. By further studying such mobility patterns in different temporal ranges, we identified both consistency and seasonal fluctuations regarding the distance decay effects in the corresponding mobility patterns. At the same time, our approach provides a geo-visualization unit with an interactive 3D virtual globe web mapping interface for exploratory geo-visual analytics of the multi-level spatiotemporal Twitter user movements.

Keywords: Geo-located tweets, mobility patterns, multi-scale spatiotemporal analysis, scalable visual-analytics framework

1. Introduction

Understanding human mobility patterns is of great importance for a broad range of applications from urban planning [1], traffic management [2], and even the spatial spread of epidemic diseases [3]. Earlier research efforts relied on low-resolution mobility data to understand human mobility patterns, such as using census records to study human migration patterns [4], or delivering questionnaires and asking volunteers to report the track of bank notes to infer human travel patterns [5]. However, such mobility data do not provide detailed human movements with fine-grained spatial and temporal granularity, which are usually aggregated and therefore are limited to capture mobility patterns of individuals [6,7]. In addition to the mobility data collected by GPS trackers [1,8] and mobile phone call records [6,9,10], emerging as a new source for mobility data, today's pervasive Location Based Social

Media (LBSM) platforms (e.g., Twitter and Foursquare) offer continuous spatial Big Data streams with massive amount of detailed and frequently updated user digital footprints in the form of real-world user trails and footprints [11]. A significant advantage of utilizing LBSM data streams as proxies for studying human mobility patterns is the large spatial coverage. For instance, researchers have used geo-located Twitter data for studying global mobility patterns [12], which is otherwise impossible for other mobility datasets (e.g., GPS traces and mobile phone call records). Regarding the concerns of infringement on individual privacy, such as the mobile phone call records with restricted access [7,13,14], the publicly available LBSM data streams offer unique opportunities for conducting reproducible and comparative scientific findings across different geographic regions.

Many recent studies have adopted the LBSM data streams to study human mobility patterns. For example, they modeled and extracted trajectories of individuals and performed statistical analysis focusing on the distance decay effects in the collective user movements [6], which were used to reveal different travel modes [7], travel demands [15,16], and the impact of social connections [17]. These studies have provided strong supports for using LBSM data as proxies for studying mobility patterns of individuals and valuable insights into human mobility dynamics. However, the variations of movements in different spatial scales and temporal ranges are neglected in these studies, where the measurements of distances are either fixed in a certain time range or to a specific geographic region. For instances, the examinations on whether there are temporal (e.g., monthly or seasonal) changes within the movements or how the observed mobility patterns vary across different spatial scales (e.g., intra- or inter city or national level), are lacking. Such insights are critical to advance our understandings of the collective mobility patterns for various applications, such as examining the mobility patterns across different cities [18], the spread patterns of disease [19,20] and touristic activities [12]. On the other hand, while the high-resolution spatiotemporal records from LBSM present unique research opportunities in this direction, the inherited large data volume poses significant data-intensive challenges for developing multi-scale spatiotemporal analysis approaches to dealing with the complexities in filtering movements of individuals, modeling and aggregating user trajectories at multiple spatial and temporal scales [21].

In this paper, we have employed a visual-analytics approach to exploring the Twitter user mobility patterns across multi-level spatial scales and temporal ranges in the continuous United States (i.e., excluding Alaska and Hawaii) during the year 2014. The mobility data is extracted from over 1.3 billion geo-located Twitter messages (i.e., tweets) from 1^{st} January to 31^{st} December, 2014 over North America with over 6 million Twitter users and over 1 TB in file size. To address the data-intensive challenge embedded in this dataset, we have developed a scalable visual-analytics framework tailored to accommodate large volume of geo-located tweets. This framework is implemented based on high-performance distributed computing environment using Apache Hadoop[1], which is an open source software framework to enable distributed processing of large datasets across computing clusters. Enabled by this framework, we have performed a set of statistical analysis to understand multi-scale spatiotemporal Twitter user mobility patterns. We have modeled the frequency of Twitter users visiting different locations to study the collective user visiting behaviors, where we have identified temporal similarities in the distributions. In particular, the Twitter user mobility patterns measured by the user displacements and radius of gyrations of individuals [6] have revealed different groups of Twitter users with multi-scale or multi-modal mobility patterns and multiple travel modes [7]. By further studying such mobility patterns in different temporal ranges, we have identified both consistency and seasonal fluctuations regarding the distance decay effects in the corresponding mobility patterns. In particular, our approach provides an interactive 3D virtual globe web mapping interface to enable exploratory geo-visual analytics for understanding the detailed Twitter user movement flows within a given spatial scale and time window.

[1] http://hadoop.apache.org/

The remainder of this paper is organized as follows. Section 2 describes the related work in the context of studying mobility patterns using LBSM data, in particular, the geo-located Twitter data. We focus on research challenges in using visual-analytics methods to enable multi-scale spatiotemporal analysis with massive movement datasets, including data management, multi-level spatiotemporal user trajectory modeling and visualization. Section 3 details the processes for extracting, aggregating and summarizing multi-level spatiotemporal Twitter user mobility patterns. Section 4 presents the case study of performing visual-analytics for seeking multi-scale spatiotemporal Twitter mobility patterns in the continuous United States of year 2014. Section 5 concludes the paper.

2. Mobility patterns in Location Based Social Media data

2.1. Geo-located Twitter data for studying large-scale user movements

To understand detailed mobility patterns of individuals, the capability of capturing human movements with fine-grained spatial and temporal granularity is critical. In terms of collecting detailed mobility data of individuals, using GPS trackers tends to produce, to date, the most accurate records of individuals' movements regarding the accuracy of recorded user locations and update frequency [1]. However, the data are often limited in spatial scale (e.g. within a specific city or region) with a small group of people, for example, 226 and 182 volunteers participated in collecting such mobility data in [8] and [22] respectively. Other than tracking people directly, the vehicle-based GPS traces are often tied to specific vehicles (e.g. taxi), which are only accessible for a certain group of people [10].

Another approach from the literatures for studying human mobility is using mobile phone call data, such as Call Detail Records (CDR), where the locations of mobile users are estimated by cell tower triangulation with accuracy in the order of kilometers [6,9,10]. Such a dataset can cover relatively large spatial scale [23,24] (e.g., national level) and a large portion of the population in the study region [10]. However, due to the concerns of infringement on individual privacy, mobile phone call data are not publicly accessible at all. Even such data were obtained in the mentioned studies, they came from various service providers covering different groups of users. These issues limit the capability for conducting reproducible scientific findings regarding mobility research, such as validating or extending the existing discoveries.

In this connection, it becomes increasingly popular for researchers to exploit the publicly accessible mobility data captured from today's pervasive Location Based Social Media (LBSM) platforms (e.g., Foursquare and Twitter). LBSM enables users to attach their current location as a geo-tag to the message they post, which is derived from either the GPS or Wi-Fi positioning with a high position resolution down to 10 meters [7]. A Big Data scenario emerges when millions social media users constantly posting messages. In this study, geo-located Twitter data are chosen as a source for studying detailed mobility patterns. Compared to other LBSM platforms, Twitter is one of the most popular platforms and is been actively used in many countries. It provides a publicly accessible streaming API[2] for easy access to its data, in fact, many other LBSM data can be collected from the data streams, such as Foursquare check-in data [16,25].

However, it is worth noting that there are some limitations and complexities in directly using LBSM data for studying human mobility patterns. For example, comparing to GPS traces, the update frequency of an individual's location varies depending on when a user is posting a new geo-located message or check-in at a new place. There is a potential mismatch regarding the representativeness of the overall population as not all people use social media or send geo-located messages [10] and the demographic information of the Twitter users cannot be easily identified, the derived mobility patterns may lead to an over or under-representation of the real-world mobility

[2] https://dev.twitter.com/streaming/overview

patterns. Many studies started to look into the demographic information of LBSM data, in particular Twitter data [26,27]. Although the used methods are still arguable, these issues certainly require us to pose stricter criteria in understanding human mobility patterns using geo-located Twitter data. On the other hand, geo-located Twitter dataset presents some unique advantages that make it a valuable proxy for studying human mobility patterns. For example, the high-resolution location information enables to identify multiple travel modes in user mobility patterns [7]; the large spatial coverage enables to study global mobility patterns [12], which is almost impossible for other mobility datasets. More importantly, by continuously monitoring the geo-located Twitter data streams with large volumes of detailed and frequently updated spatiotemporal records of Twitter users, it offers a great deal of potential for studying mobility patterns of large groups of individuals at different spatial scales (e.g., movements across cities, states or even countries) and temporal gratuity (e.g., weekly, monthly, and seasonal movements), which is one of the motivations for this study.

2.2. Data-intensive challenges for multi-scale geo-visual analytics

Mobility data are essentially a collection of spatiotemporal records of people moving from one location to another across the geographic space. To study mobility patterns of individuals, a space-time trajectory of each individual user should be modeled and constructed to quantify the collective movements over space and time. Based on the extracted space-time trajectories, aforementioned studies are able to perform analysis, such as the measurements of user displacements and radius of gyrations of individuals, to study the mobility dynamics. At the same time, space-time trajectory is one of the core concepts in Hägerstrand's time geography to understand the embedded spatiotemporal dynamics [28], which has provided useful insights to explore movements across different geographical scales and temporal ranges. For example, a geo-visualization approach was used to study human activity patterns, where user trajectories are mapped in a 3D space ordered by timestamps in the third dimension [29]. While such an approach enables visualization of individual trajectories, its capability is limited in dealing with large-volume movement datasets [30]. Instead of directly visualizing individual user trajectories, a space-time cube approach was proposed to analyzing and visualizing the collective trajectories. It provides flexibilities in setting up both spatial scales and temporal ranges, and therefore is used to study mobility patterns across different spatial units (e.g. countries, states, and cities, etc.) and identify the changes over space and time [31,32]. In this regard, visual-analytics methods are proposed to help better convey the findings in terms of analyzing and visualizing multi-level spatiotemporal mobility patterns [30,33]. Visual-analytics methods focus on the synergy of computational and analytical methods to reduce the visual clutter, where aggregation methods are suggested to perform grouping/dividing individual's moving trajectories at different spatial and temporal granularity, e.g., utilizing the space-time cube approach [30]. Employing visual-analytics methods dealing with massive movement datasets is not only beneficial for optimizing visualizations but also provides a great deal of flexibilities for performing statistical analysis in seeking mobility patterns with different level of spatiotemporal details.

However, in the context of studying mobility patterns using large volume of geo-located Twitter data, the inherited large data volume poses significant data intensive challenges for visual-analytics methods to scale with both the data volume and the computational requirements (e.g., movement extraction and trajectory modeling) [34]. In particular, in our study, 1.3 billion geo-located tweets were collected with over 1 TB in file size. To construct a space-time trajectory of an individual, it is necessary to go through the massive dataset to sort and update the trajectory whenever a new location is found. Such a task is already computationally demanding, let along breaking the trajectories to construct space-time cube with multiple spatial scale and temporal ranges. Indeed, developing a multi-scale spatiotemporal analysis approach is identified as one of the research challenges for dealing with social media Big Data [21]. To address the data-intensive challenges, there is a need to develop a scalable visual-analytics framework tailored to accommodate large volume of geo-located tweets for studying multi-level spatiotemporal Twitter user mobility patterns.

3. Materials and Methods

3.1. Geo-located Twitter data

Geo-located tweets are tweets appended with an additional geo-tag in the form of a pair of geographical coordinates, which represents the location a tweet was sent at. In this study, the geo-located tweets were downloaded using the Twitter Streaming API, where we specified a geographical bounding box as an area-of-interest to retrieve all the geo-located tweets that fall within it. To ensure complete coverage over the continuous United States, we implemented a crawler that selects North America as the initial area-of-interest, where the geographical boundary is specified with lower left (latitude: 5.4, longitude: -167.3) and upper right (latitude: 83.2, longitude: -52.2). The crawler is constantly running with over 2 million geo-located raw tweets ($\sim$2GB in size) collected per day. We have collected more than 1.3 billion geo-located tweets from 1^{st} January to 31^{st} December, 2014 with 6,147,430 Twitter users and 1 TB in file size.

As a social media account is not equal to a real person in the physical world [21], to ensure the data quality, the collected raw tweets were further filtered by the following steps: We first removed duplicated messages[3] in the dataset; and then we removed non-human users based on the heuristic of unusual relocating speed discussed in [7,12]. In this case, we adopted the speed limit value as 240 m/s used in [7], where we examined all the consecutive locations of each user and excluded those with relocating speed over the limit. Note that the original location information embedded in each geo-located tweet is given in units of latitude and longitude, the distance is calculated by the great-circle distance between two points on a sphere with the haversine formula. Finally, we used the geographic boundaries of the continuous United States[4] (excluding Alaska and Hawaii) to further restrict the remaining tweets, where the technical details is presented in the following section. Based on these reinforcements, the dataset contains 1,052,861,000 tweets and 4,559,205 unique users.

3.2. A scalable visual-analytics framework

To address the data-intensive challenges, we have developed a scalable visual-analytics framework tailored to accommodate large volume of geo-located tweets for studying multi-level spatiotemporal Twitter user mobility patterns. The scalable visual-analytics framework consists of two main units: (1) Data processing unit: a distributed computing environment using Apache Hadoop for modeling and extracting Twitter user movements, generating space-time user trajectories, and summarizing the movements at multiple spatiotemporal scales. (2) Geo-visualization unit: an interactive 3D virtual globe web mapping interface for exploratory geo-visual analytics for understanding the detailed Twitter user movement flows across different spatial scales and temporal ranges.

Apache Hadoop combines a distributed file system, namely Hadoop Distributed File System(HDFS) [35] with MapReduce programming paradigm [36], which can be applied to a wide range of data-intensive problems. Our framework benefits from using Hadoop in both data management and processing. First, since the input data is large it is desirable to store it on multiple machines. This provides scalability in relation to the growth of data size, where Hadoop can scale to more computing nodes in a cluster to maintain the performance. Second, by using Hadoop we can parallelize the computational tasks, where MapReduce breaks the entire computation into small tasks and schedule them among different computing nodes, to make the data processing faster and more

[3] https://support.twitter.com/articles/18311-the-twitter-rules
[4] https://www.census.gov/geo/maps-data/data/

efficient. An overall system architecture of the framework is shown in Figure 1. The details regarding the function and implementation of each unit are presented in the next sections.

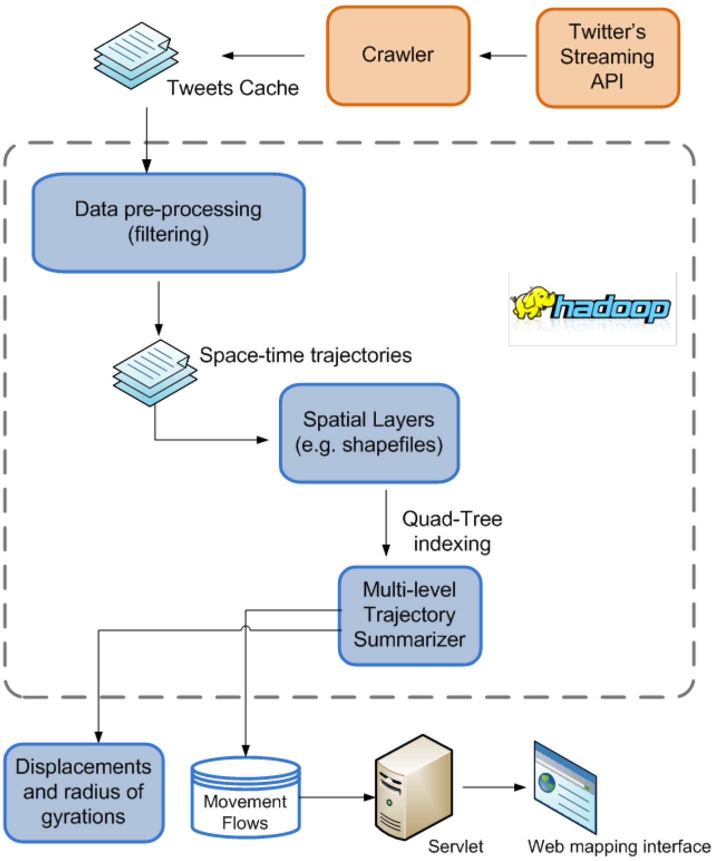

Figure 1. The overall system architecture of the framework

3.3. Space-time Twitter user trajectories

To derive meaningful mobility patterns of individuals, a space-time trajectory of each individual user should be constructed [28]. Each raw geo-located tweet contains multiple fields of information, such as the created time, country, language code, and location, etc. To construct a space-time trajectory from the data collection, we are interested in the following fields: *User ID*, *location*, *timestamp*, which can be represented by a tuple $\langle id, loc, t\rangle$, where *id* is a unique string representing a Twitter user's id; *loc* is the recorded location of the message represented as a pair of projected coordinates $\langle x, y\rangle$; and t is the timestamp of when the message was posted; A Twitter user's space-time trajectory is defined as follows.

Definition 1. Space-time Twitter user trajectory: The space-time trajectory of a Twitter user is defined as a collection of recorded geo-locations in the chronological order (i.e., based on the attached timestamp):

$$Trajectory_{user_{id}} \equiv \{\langle id, loc_1, t_1\rangle, \langle id, loc_2, t_2\rangle, \langle id, loc_i, t_i\rangle, ...\langle id, loc_n, t_n\rangle\}, i = 1, 2, 3...n$$

To remove non-human users based on unusual relocating speed, a user will be removed if the speed between any two consecutive locations in the user's trajectory with *speed* $(loc_i - loc_{i-1}) > 240m/s$. Based on this definition for modeling the space-time Twitter user trajectories, we converted the process of extracting trajectories from the raw geo-located Twitter data as a MapReduce task. Specifically, each mapper utilizes the unique user id as a key to prepares the records that belong to the same user and send them to reducer. Once the reducers receive the $\langle key, value \rangle$ pairs, a Twitter user's space-time trajectory is formed by the sorting the locations in chronological order while considering the speed limit.

Definition 2. visitation behavior, displacement and radius of gyration: As each space-time Twitter user trajectory records all the locations a user has visited, the visitation behavior refers the frequency of a user visiting different locations within a specific time frame. This metric provides an overall assessment regarding the diversity and similarity in the collective mobility pattern [37].

In particular, the measurements of displacements and radius of gyrations of individuals are two popular metrics to investigate and quantify the distance decay effects in the collective mobility patterns [6]. The displacement refers to an individual's re-allocation across the geographic space measured in distance, i.e., $distance(loc_i - loc_{i-1})$. It is not equivalent to a "trip" took by an individual, for example, even the time interval between two recorded locations is one month, it will still count as a displacement. By studying the collective displacements from a group people, it helps to identify the distance bounds associated with different travel modes [7] and to quantitatively differentiate the mobility patterns from random walks [5]. On the other hand, radius of gyration is a metric to distinguish mobility patterns of individuals [6], which is defined as fellows.

$$r_g = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(p_i - p_{centroid})^2}, \text{ where } p_{centroid} = \frac{1}{n}\sum_{i=1}^{n} p_i$$

It measures the accumulated distances of deviation from the center of mass of an individual user's trajectory, and therefore indicates the individual's spatial coverage, where p_i is one of the user's locations and $p_{centroid}$ is the center of mass of the user's trajectory. When applying the measurement to the study population, it identifies different groups of people in terms of spatial coverage from their corresponding mobility patterns. Note that both displacements and radius of gyrations are measured by "crow's fly distance" in this study (i.e., the direct great-circle distance between two recorded locations). Since these metrics are based on the generated trajectories, by breaking and aggregating the trajectories in multiple spatial scales and temporal ranges, it enables performing multi-scale spatiotemporal analysis on these measurements and studying the corresponding mobility patterns.

3.4. Multi-level spatiotemporal trajectory aggregation

An important strategy for visual-analytics methods dealing with massive movement datasets is performing spatial aggregations to provide different levels-of-detail [30,33]. It is similar to the map generation approach that when a user is interacting with a map interface, the details of visualization should be adaptive to a user's area-of-interest [38]. To enable aggregating Twitter trajectories into multiple spatial scales, we have extended the hierarchical space-time cube model developed in [34], where we partitioned the geographic space of the continuous United States into 10 hierarchical spatial layers. To be specific, the state boundaries of the continuous United States are treated as the base layer (i.e. level 0) for aggregating state-level Twitter user movements, Alaska and Hawaii are excluded for the consideration of better visualization effects in the mapping interface of the framework. We then created an hierarchical fishnet by diving the study region into regular cells, where the finest level (level 10) consists 1 km × 1 km cells. Such a cell size is consistent with the spatial resolution in landscan[5] product for measuring the global population density. In our case, the cell size for level i-1 is twice of the size in level i. Figure 2 illustrates an hierarchical fishnet spatial units for mapping multi-level Twitter user movements. Note that any predefined geographic boundaries

[5] http://web.ornl.gov/sci/landscan/

can be used and appended in this framework to show different level-of-detailed movements (e.g., national-level and census-tract level), in our case, we replaced the level 8 fishnet layer with the US county boundaries.

To perform a multi-level spatial aggregation of the Twitter user trajectories using hierarchical spatial layers, each location in a user's trajectory is redistributed to the corresponding spatial units. A MapReduce algorithm for the spatial aggregation is implemented, where the ID of unit in each spatial layer (e.g., polygon in state and country layer and cell in the rest) is treated as key in the at the map stage. It performs a "point-in-polygon" geospatial operation to determine which polygon the point belongs to. If the location does not belong to any polygon, it will be dropped, which is how we used the geographic boundaries of the continuous United States to filer the raw tweet collection that initially covered the North America and kept the "domestic" ones. To optimize the "point-in-polygon" determination without comparing the location with every polygon in the spatial layer, we also created a Quad-Tree [39] for each spatial layer to speed up the process. Finally, the reducers generate two data outputs: (1) reconstructed space-time Twitter user trajectories at each spatial level (2) movement flows in the form of in and out movement flux between the spatial units. The movement flows are stored in the database for interactive explorations in the 3D web mapping interface, whereas the re-constructed trajectories can be further processed to produce distance measures at different spatial scales, which is illustrated as fellows:

$Trajectory_{user_{id}} \equiv \{\langle id, loc_1, t_1, unit_1\rangle, \langle id, loc_2, t_2, unit_2\rangle, \langle id, loc_i, t_i, unit_i\rangle, ...\langle id, loc_n, t_n, unit_n\rangle\}$ where $i = 1, 2, 3...n$

Figure 2. Hierarchical spatial layers for aggregating movements in different level-of-details

4. Multi-scale spatiotemporal Twitter user mobility patterns

4.1. Spatiotemporal Twitter user mobility patterns

The situation of using geo-located tweets as proxies to infer people's movements is complex as users' tweeting behavior can be significantly different from one to another, in particular, the frequency and time-interval between two consecutive tweets. For example, some people may tweet once a day while others do more; some people may tweet regularly while others do not. These tweeting behaviors are expected as such human dynamics are also seen in the mobile phone call data [6]. Many studies have carried out data collection within a certain time period (e.g., a year in our case). However, as the geo-located tweets were collected in a continuous fashion, it is necessary to examine the sensitivities regarding these behaviors to make sure we are not just capturing a random snapshot from the whole data streams.

In this study, we have analyzed the cumulative distribution of the frequency Twitter users visiting different locations in year 2014 (and every month), which uses the methods developed in [40]. The frequency is summarized based on the trajectories of individuals extracted from a monthly time span. Note that different groups of Twitter user may exist in each month. It appears that the distribution of the collective Twitter user visitation behaviors in year 2014 follows a two-tiered power law distributions (shown in Figure 3, where the majority (the front part) of the distribution follows a truncated power-law distribution $p(x) \sim x^{-\alpha}e^{-\lambda x}$ and the α value is 1.32, and the tail part (less than 2% of the whole population) follows a power-law distribution $p(x) \sim x^{-\alpha}$ with α value is 3.5. This finding is consistent across all 12 months, with the mean α value as 1.34 ± 0.05 (standard deviation) and the mean λ value as 0.00178 ± 0.0002 (standard deviation).

The two-tier power law distribution indicates that the collective behaviors of Twitter user visiting different locations can be well approximated with a (truncated) Lévy Walk model [8,41], which has also been identified in many human mobility studies using different mobility data [42]. The similarities among the cumulative distributions suggest that the mobility data collected from geo-located tweets are temporally stable, at least at the monthly interval, which indicates the collected geo-located tweets in one month can potentially reveal similar findings as the ones collected in multiple months. In addition, the two-tier power law also reveals the diversity in the Twitter user visiting behaviors: (1) a small group Twitter users visited significantly more locations than the others (2) within each group, the probability of Twitter user visiting more locations decreases significantly with a power function.

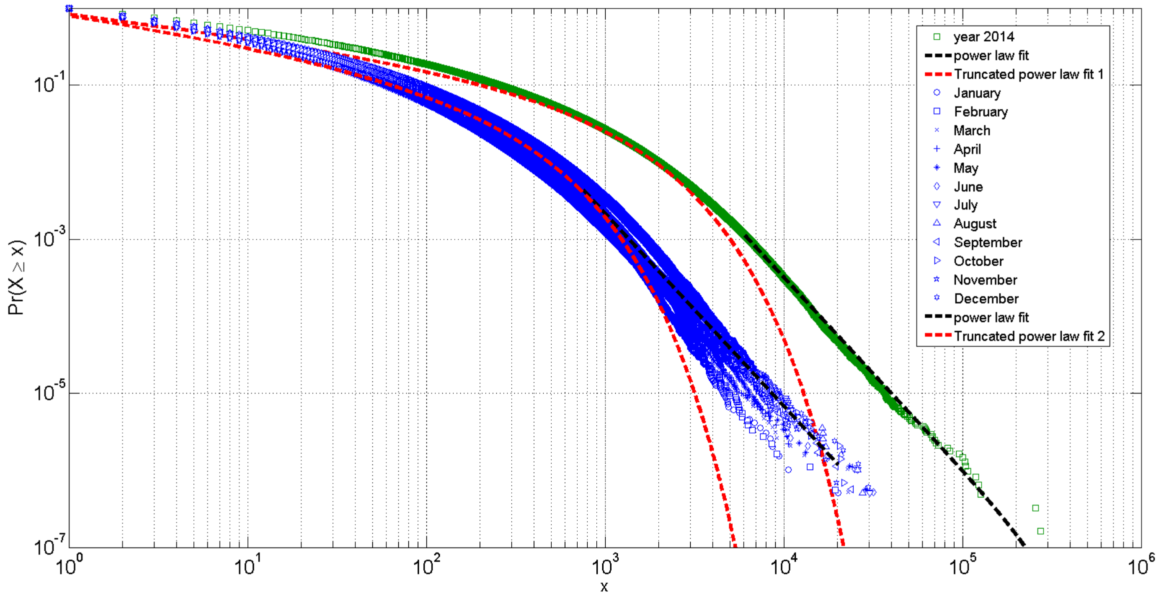

Figure 3. Two-tier power law distribution of the collective Twitter user visitation behaviors

As it is mentioned above, the measurements of displacements and radius of gyrations of individuals are two popular metrics to investigate and quantify the distance decay effects in the collective mobility patterns [6]. In this case, we first gathered the displacements from all the collected Twitter users in the continuous United States in year 2014, where those Twitter users with only one geo-located tweet were filtered out. To investigate the mobility patterns of individuals, we also derived the accumulated displacements and the radius of gyrations of each individual Twitter user based on the corresponding space-time trajectories over the one year period. Note that both displacements and radius of gyrations were calculated by the direct great-circle distance (d) between two consecutively recorded locations in a user's trajectory.

To seek mobility patterns from these measurements, we performed statistical analysis regarding the probability distributions of displacements and radius of gyrations, which is also known as the spatial dispersal kernel $P(d)$ [5]. The probability distribution of the user displacements (as well as accumulated displacements) is shown in Figure 4, whereas the probability distribution of radius of gyrations is shown in Figure 5. In this study, we used the fitting methods developed by [7]. The probability distributions of overall displacements, and the accumulated displacements and radius of gyrations of individuals, can all be approximated by a combination of three functions: an exponential function, a stretched-exponential function and a power-law function.

In particular, as it is shown in Figure 4 (a), the probability distribution of the overall displacements is approximated by $P(d) \sim \lambda_1 e^{-\lambda_1(d-d_{min})}, d_{min} = 10m$ from [10 m, 70 m] (accounting for 2 % of the population), $P(d) \sim \beta\lambda_1 d^{\beta-1} e^{-\lambda^1(d^\beta-d^\beta_{min})}, d_{min} = 100m$ from [100 m, 80 km] (accounting for 93 % of the population), and $P(d) \sim d^{-\alpha}$ [> 80 km] (accounting for 5 % of the population). In addition, the displacement in the distance bound from 100 m and 80 km in Figure 4 (b) can be further approximately by two power-law distributions with a cutting point at 5 km (53% distances are less than 5 km and 40% distances between 5 km and 80 km), which indicates two different travel modes, such as inter- or intra-city movements. Overall, the fitting functions with different distance bounds suggest the existence of multi-scale or multi-modal mobility patterns [7] of the Twitter users in the continuous United States, for example the displacements larger than 80 km could be related to inter-state travels or travel by flight.

The probability distribution of radius of gyrations of individuals at the national level (Figure 5 (a)) is approximated by $P(r_g) \sim \lambda_2 e^{-\lambda_2(r_g-r_{g_{min}})}, r_{g_{min}} = 10m$ from [10 m, 50 m], $P(r_g) \sim$

$\lambda_2 e^{-\lambda_2 (r_g - r_{g_{min}})}$ from [50 m, 30 km], and $P(r_g) \sim {r_g}^{-\alpha}$ [> 30 km]. In particular, the radius of gyration between 50 m and 30 km can be further approximately by two power law distributions with a cutting point at 6 km (Figure 5 (b)), which suggest two main types of spatial coverage of from the collected Twitter users in the continuous United States. The distribution shows that around 10% the tweet population has a radius of gyration less than 50 meters, which indicates those twitter users mostly tweet at a particular place, such as home or office; around 60% of the population has a radius of gyration less than 30 km, which indicates that most of the collected Twitter user movements are "short" distances, e.g., within a city locale. Note that the accuracy of these values for defining the distance bound depends on the accuracy of the location information of each geo-located tweet.

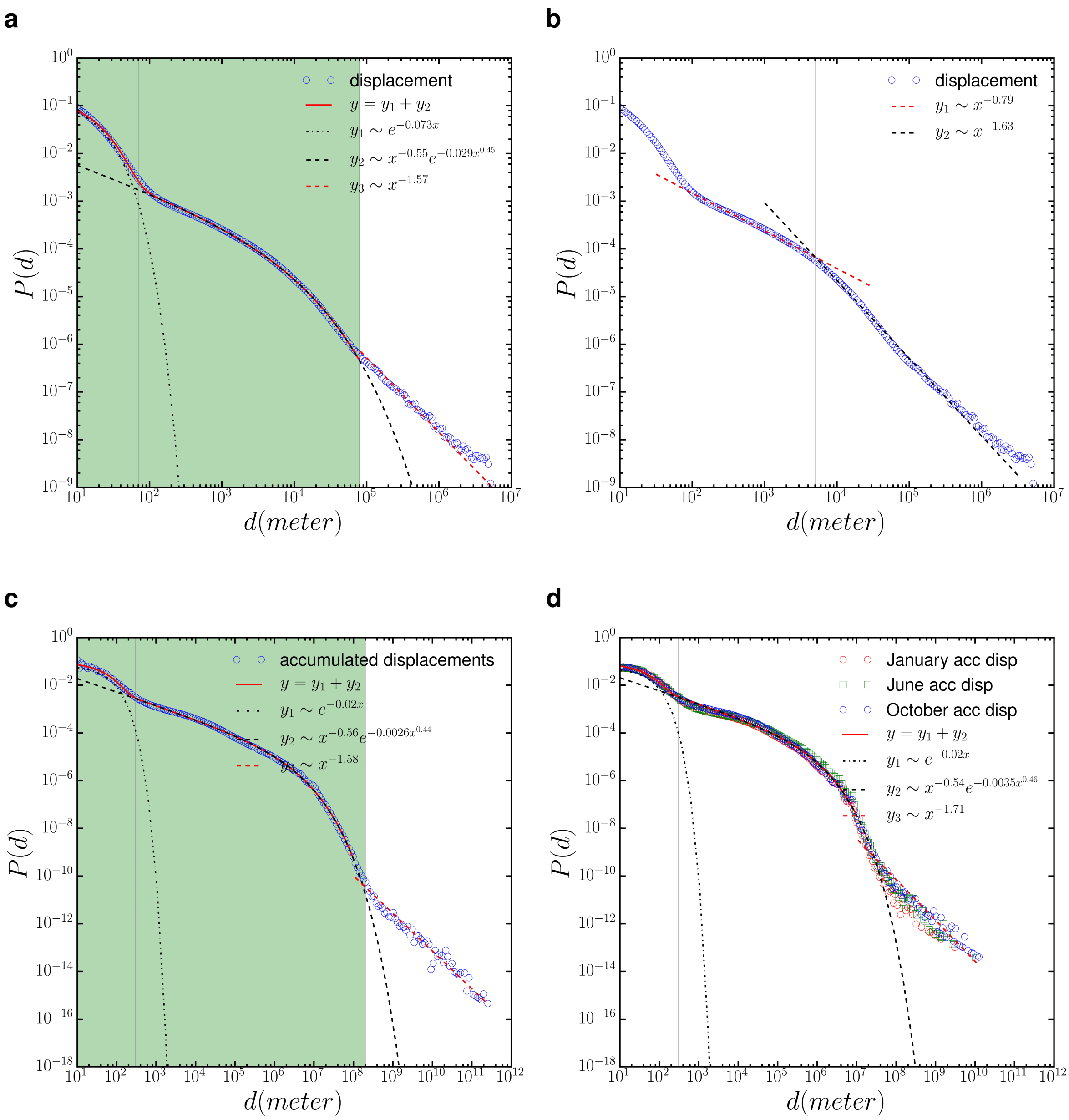

Figure 4. (a) The probability distribution of the collective Twitter user displacements P(d) (b) the distance between [100 m, 80 km] is approximated by a double power-law functions (c) The probability distribution of the accumulated displacements of individual Twitter users P(d) (d) The probability distribution of the accumulated displacements of individual Twitter users in 3 different months

We also measured the distribution of the radius of gyration of Twitter users at different spatial scales, specifically, the state level and city level. In this study, we selected the state Illinois and California for comparisons at the state level (Figure 5 (c)), whereas we chose Chicago city as an example (Figure 5 (d)) at the city level. Interestingly but not surprisingly, the P(r_g) at the state level can also be approximated by a combination of three functions: an exponential function, a stretched-exponential function and a power-law function. We noticed that distance bound of the radius of gyration at the state level is at 10 km instead of 30 km at the national level. The distance decay effects in larger spatial coverage [> 30 km] slightly differ, in this case, the P(r_g) decreases faster in smaller size state (i.e., Illinois) than the large size state (i.e., California). In particular, the P(r_g) over Chicago city can be fitted by similar functions. However, as it reflects intra-city level mobility patterns, there is no distinct distance range to indicate large spatial coverage.

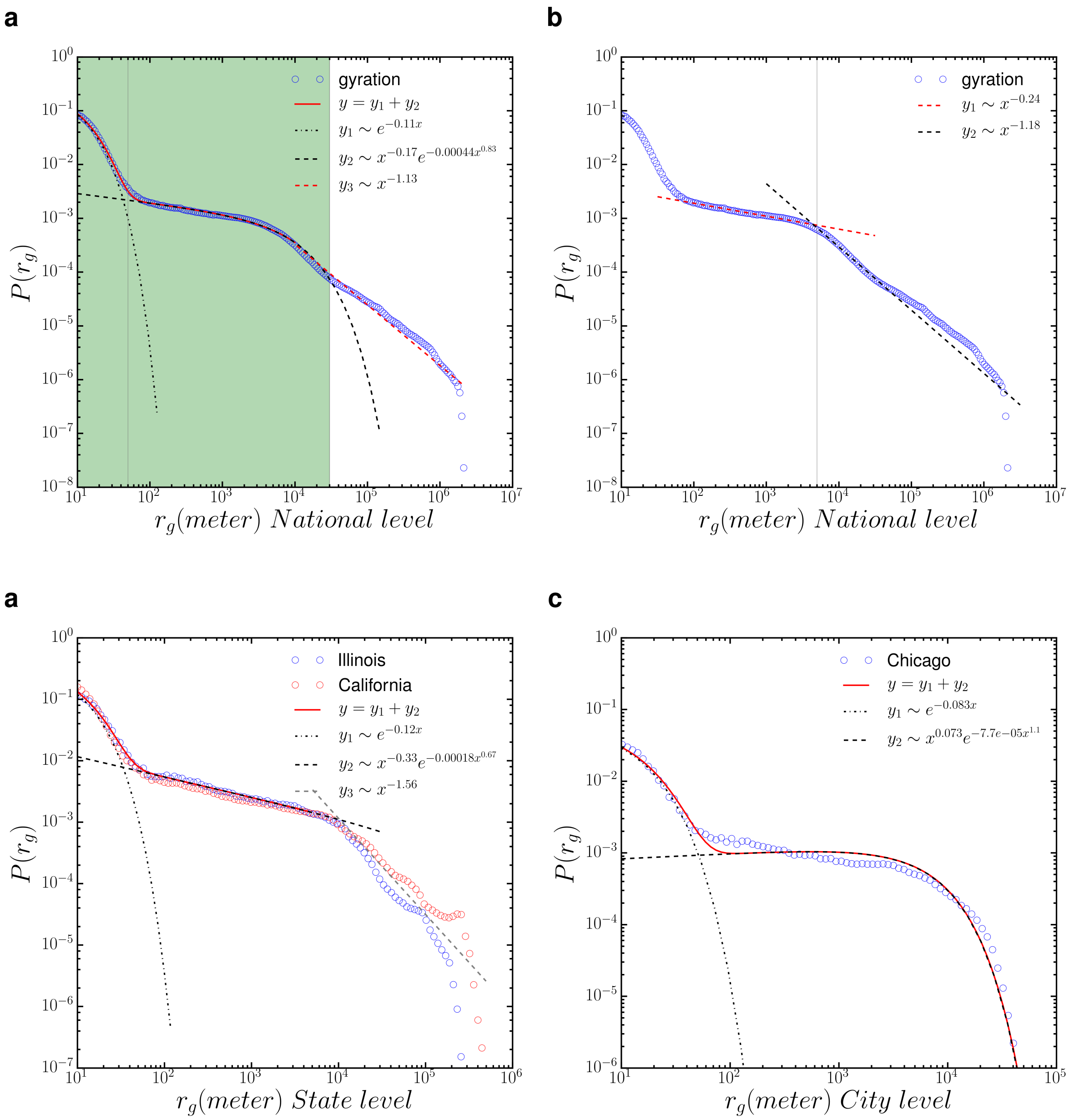

Figure 5. (a) The probability distribution of radius of gyration of individual Twitter users P(r_g) at the national level (b) the distance between [50 m, 30 km] is approximated by a double power-law functions (c) P(r_g) at the state level (Illinois and California) (d) P(r_g) for Chicago city

On the other hand, as our framework can aggregate Twitter user trajectories within different temporal ranges, we further analyzed the probability distributions of accumulated displacements took places in January, June, and October (Figure 4 (d)) and radius of gyrations within 4 quarters in year 2014 (Figure 6, in order to examine whether there are temporal changes in the mobility patterns. While the probability distributions of accumulated displacements are almost identical in those selected three months, we do find changes in the probability distributions of radius of gyrations in different quarters of the year. The fluctuations in the tails of the distributions indicate that long distance radius of gyrations (i.e., above 30 km) will experience more seasonal changes in the Twitter user mobility pattern, which means the increase or decrease of long distance movement activities in the corresponding time period. However, it is worthy noting that the overall trends in the Twitter user mobility patterns revealed by radius of gyrations are still consistent.

In summary, by comparing these results from different spatial scales and temporal ranges, different distance bounds were identified for describing the spatiotemporal Twitter user mobility patterns. However the overall similarity and consistence found in using a combination of three functions to approximate the probability distribution functions of displacements and radius of gyrations, clearly provide supports for using geo-located tweets as useful proxies for understanding human mobility patterns and conducting reproducible findings at multiple spatial scales and temporal ranges.

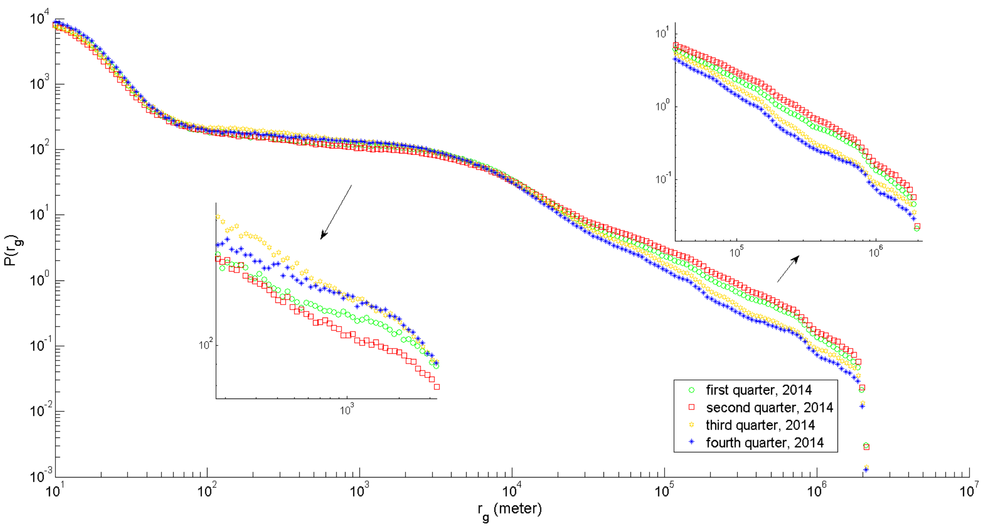

Figure 6. The probability distribution of radius of gyration of individual Twitter users in different quarters of year 2014

The above analysis of Twitter user mobility patterns mainly focus on the spatiotemporal aspects. Our framework provides the flexibility to aggregate and extract Twitter user trajectories in a specific spatial scale and re-produce the analysis. In particular, as it is evident from the above analysis that there are multi-scale or multi-modal Twitter user mobility patterns, this framework can help further look into the mobility pattern regarding how Twitter users move across different spatial scales and temporal ranges, which is measured by the movement flows between these spatial units. In this case, we demonstrate the inter-state mobility patterns by using the framework to capture the movement flows between the states. Note that the movement flows can be summarized across all the 10 spatial layers in the framework. We tested the overall distribution of the movement flows (in the form of weighted in-degree and out-degree of a graph, where each state is treated as a node) among different states in year 2014. We found that the probability distribution of Twitter user movement flows of visiting different states follows a log-normal distribution: $p(x) \sim$

$\frac{1}{x}exp[-\frac{(lnx-\mu)^2}{2\sigma^2}]$, which suggests the flux of Twitter user movements among the states are highly skewed and dominated by a few states. It indicates that the Twitter population is not proportional to account the movement flux between the states, which may provide some insights for other researchers in studying social-economical aspects of the migration dynamics.

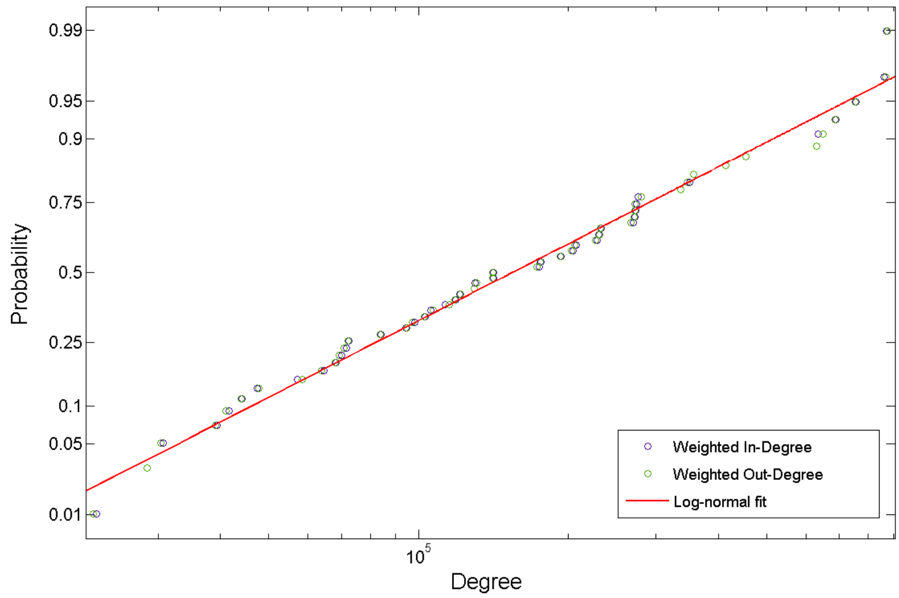

Figure 7. The distribution of Twitter user movement flows among different states in year 2014 measured in weighted in- and out-degrees

4.2. The interactive 3D virtual globe web mapping interface

In addition to providing supports for understanding Twitter user mobility patterns with statistical analysis, the framework integrates a 3D virtual globe interface to enable users to perform exploratory geo-visual analytics of the multi-level spatiotemporal Twitter user movements. The 3D virtual globe is developed and extended from the Cesium library[6], which is an open-source WebGL virtual globe and map engine. We customized the map engine to adapt different spatial scales, which correspond to the hierarchical spatial layers, for aggregating movements in different level-of-details. The map interface interprets user's interactions, such as area-of-interest, time window, and zoom levels, etc. as parameters and send to the dedicated visualization servlet on the CyberGIS Gateway[7], which is the leading online cyberGIS environment for a large number of users to perform computing- and data-intensive, and collaborative geospatial problem-solving enabled by advanced cyberinfrastructure [43]. In return, the map interface visualizes the corresponding movement flows on the virtual globe.

An overview of the 3D web mapping interface is shown in Figure 8. In terms of performing exploratory visual-analytics of Twitter user movement patterns, users can specify the time window to enable the query. When the results are shown, users can hover the mouse over each individual lines on the map to see the value of movement flows for both in and out directions. If the selected criteria keep unchanged, whenever the user zooms in/out, tilt or rotate the globe, the 3D virtual globe mapping interface will automatically provide the corresponding level-of-details on the fly.

[6] http://cesiumjs.org/
[7] http://sandbox.cigi.illinois.edu/home/

For example, Figure 9 and Figure 10 demonstrated the movement flows in different level-of-details around the Chicago city, and between O'Hare International Airport and the city center of Chicago city, where top 20 % movement flows were shown. The URL to access this 3D web mapping interfaces as well as the source code of the visual-analytics framework are available upon request.

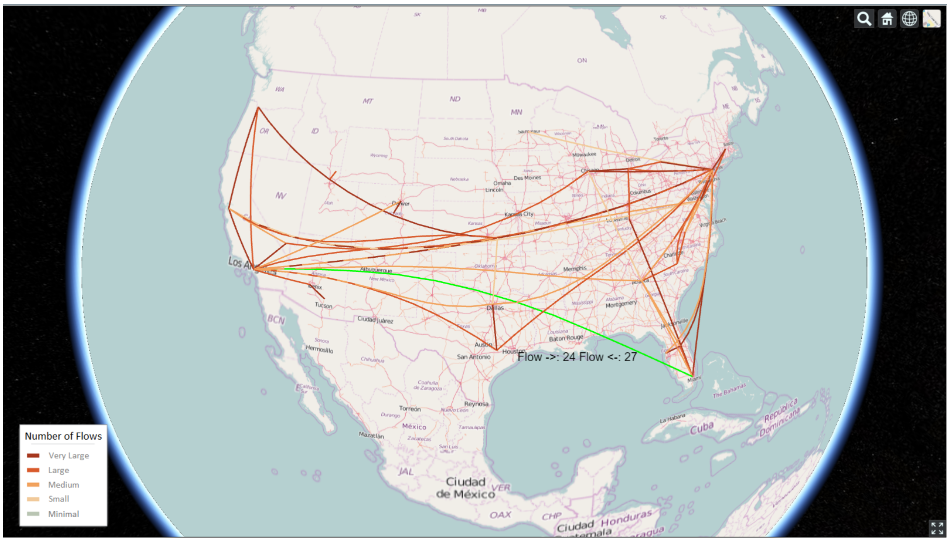

Figure 8. An overview of the 3D interactive web mapping interface

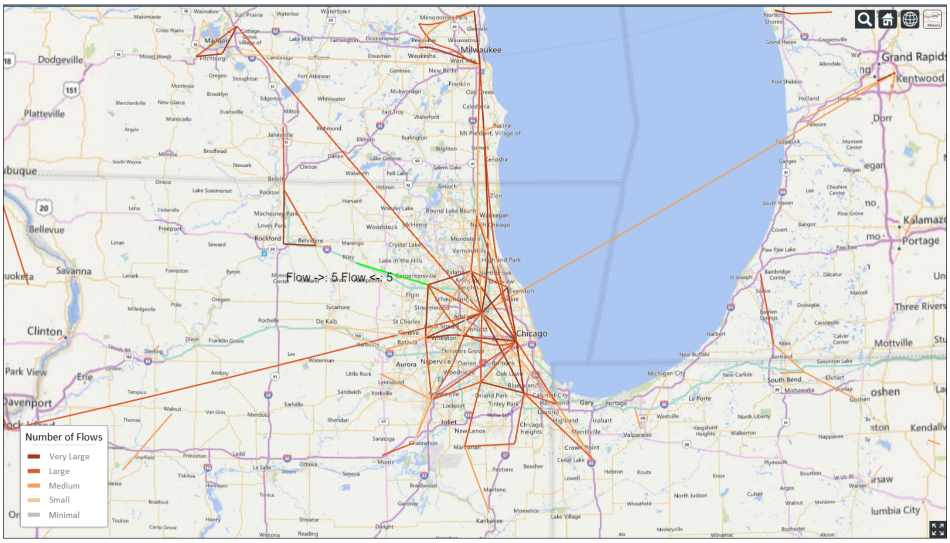

Figure 9. The top 20 % movement flows around Chicago city

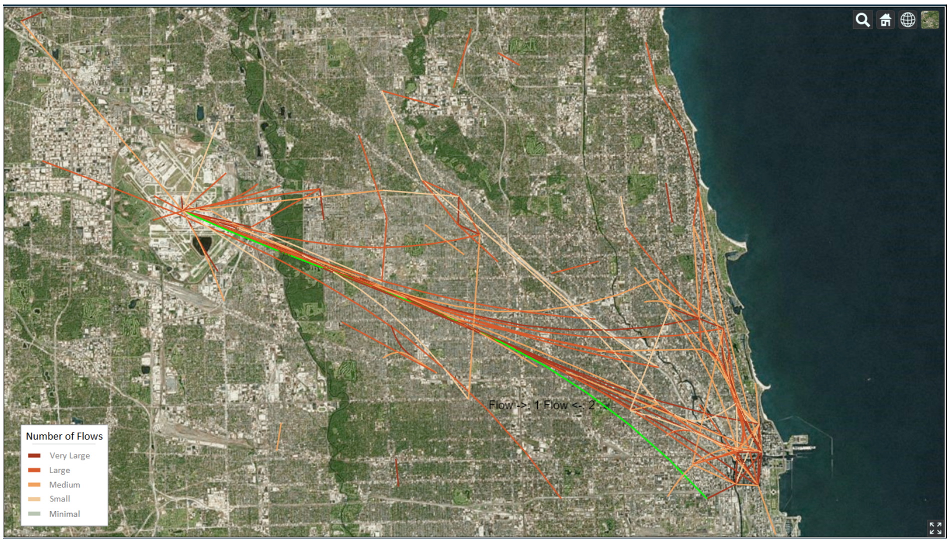

Figure 10. The top 20 % movement flows between O'Hare International Airport and the city center of Chicago city

5. Conclusions

In this study, we have used large volume of geo-located tweets to study Twitter user mobility patterns across multi-level spatial scales and temporal ranges in the continuous United States during the year 2014. To address the data-intensive challenges, we have developed a scalable visual-analytics framework tailored to accommodate large volume of geo-located tweets for studying multi-scale spatiotemporal Twitter user mobility patterns. This framework is implemented based on high-performance distributed computing environment using Apache Hadoop. It delivers scalability in filtering large volume of geo-located tweets, modeling and extracting Twitter user movements, generating space-time user trajectories, and summarizing multi-level spatiotemporal user mobility patterns.

With this framework, we have found some interesting Twitter user mobility patterns, both statistically and visually. We studied the collective Twitter user visiting behavior regarding the frequency of Twitter users visiting different locations, which was fitted by a two-tier power-law distribution function. The two-tier power law distribution indicates that the collective behaviors of Twitter user visiting different locations can be well approximated with a (truncated) Lévy Walk model, which has also been identified in many human mobility studies using different mobility data. The similarities among the cumulative distributions suggest that the mobility data collected from geo-located tweets are temporally stable, at least at the monthly interval, which provides supports that we are not just capturing a random snapshot of the whole data stream.

We studied the distance decay effects in the collective Twitter user movements measured by the probability distributions of the displacements and radius of gyrations of individuals. These distributions can all be approximated by a combination of three functions: an exponential function, a stretched-exponential function and a power-law function. In particular, distance bounds between different fitting functions in displacement distribution reveals the existence of multi-scale or multi-modal mobility patterns of the Twitter users, whereas the distribution of radius of gyration reveals different groups of Tweet users with different types of spatial coverages at multiple spatial scales. We further studied these mobility patterns in different temporal ranges to investigate the temporal changes in the mobility patterns. We found that the accumulated displacements are almost identical in different months, while the long distance radius of gyration (i.e., above 30 km) will experience more seasonal changes in the Twitter user mobility pattern.

Finally, it is worth noting that the geo-located Twitter data is not able to generalize to the entire population. As the demographic information of the Twitter users cannot be easily identified, the results of delineated urban boundaries may not reflect a complete real-world image from human movements, which should be carefully considered in future studies. Nevertheless, as we have discussed in this paper that geo-located Twitter data show the advantages regarding the easy data accessibility, the large spatial coverage and massive sample size, our approach showed that such data can be a valuable proxy for understanding human mobility patterns across multiple spatial scales and temporal ranges. Also, our approach can be applied to the setting of other countries, which can be used to carry out comparative studies regarding spatiotemporal Twitter user mobility patterns.

Acknowledgments: J.Y. would like to thank the funding support from the U.S. National Science Foundation under grant numbers: ACI-1047916, BCS-0846655, and IIS-1354329. D.Z. would like to thank the funding support from the Fundamental Research Funds for the Central Universities under grant number:2016XZZX004-02. The work also used the ROGER supercomputer, which is supported by NSF under grant number: 1429699.

Author Contributions: J.Y. conceived and designed the experiments; J.Y. and D.Z. wrote the paper.

Conflicts of Interest: The authors declare no conflict of interest. The founding sponsors had no role in the design of the study; in the collection, analyses, or interpretation of data; in the writing of the manuscript, and in the decision to publish the results.

Bibliography

1. Zheng, Y.; Li, Q.; Chen, Y.; Xie, X.; Ma, W.Y. Understanding mobility based on GPS data. Proceedings of the 10th international conference on Ubiquitous computing. ACM, 2008, pp. 312–321.
2. Jiang, B.; Yin, J.; Zhao, S. Characterizing the human mobility pattern in a large street network. *Physical Review E* **2009**, *80*, 021136.
3. Belik, V.; Geisel, T.; Brockmann, D. Natural human mobility patterns and spatial spread of infectious diseases. *Physical Review X* **2011**, *1*, 011001.
4. Greenwood, M.J. Human migration: Theory, models, and empirical studies. *Journal of regional Science* **1985**, *25*, 521–544.
5. Brockmann, D.; Hufnagel, L.; Geisel, T. The scaling laws of human travel. *Nature* **2006**, *439*, 462–465.
6. Gonzalez, M.C.; Hidalgo, C.A.; Barabasi, A.L. Understanding individual human mobility patterns. *Nature* **2008**, *453*, 779–782.
7. Jurdak, R.; Zhao, K.; Liu, J.; AbouJaoude, M.; Cameron, M.; Newth, D. Understanding Human Mobility from Twitter. *PLoS ONE* **2015**, *10*, e0131469.
8. Rhee, I.; Shin, M.; Hong, S.; Lee, K.; Kim, S.J.; Chong, S. On the levy-walk nature of human mobility. *IEEE/ACM transactions on networking (TON)* **2011**, *19*, 630–643.
9. Sevtsuk, A.; Ratti, C. Does urban mobility have a daily routine? Learning from the aggregate data of mobile networks. *Journal of Urban Technology* **2010**, *17*, 41–60.
10. Kung, K.S.; Greco, K.; Sobolevsky, S.; Ratti, C. Exploring universal patterns in human home-work commuting from mobile phone data. *PloS one* **2014**, *9*, e96180.
11. Thatcher, J. Living on fumes: Digital footprints, data fumes, and the limitations of spatial big data. *International Journal of Communication* **2014**, *8*, 1765–1783.
12. Hawelka, B.; Sitko, I.; Beinat, E.; Sobolevsky, S.; Kazakopoulos, P.; Ratti, C. Geo-located Twitter as proxy for global mobility patterns. *Cartography and Geographic Information Science* **2014**, *41*, 260–271.
13. Giannotti, F.; Pedreschi, D. *Mobility, data mining and privacy: Geographic knowledge discovery*; Springer Science & Business Media, 2008.
14. Crampton, J.W. Collect it all: national security, Big Data and governance. *GeoJournal* **2014**, pp. 1–13.
15. Wu, L.; Zhi, Y.; Sui, Z.; Liu, Y. Intra-urban human mobility and activity transition: Evidence from social media check-in data. *PloS one* **2014**, *9*, e97010.
16. Hasan, S.; Zhan, X.; Ukkusuri, S.V. Understanding urban human activity and mobility patterns using large-scale location-based data from online social media. Proceedings of the 2nd ACM SIGKDD international workshop on urban computing. ACM, 2013, p. 6.

17. Cho, E.; Myers, S.A.; Leskovec, J. Friendship and mobility: user movement in location-based social networks. Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining. ACM, 2011, pp. 1082–1090.
18. Noulas, A.; Scellato, S.; Lambiotte, R.; Pontil, M.; Mascolo, C. A tale of many cities: universal patterns in human urban mobility. *PloS one* **2012**, *7*, e37027.
19. Balcan, D.; Colizza, V.; Gonçalves, B.; Hu, H.; Ramasco, J.J.; Vespignani, A. Multiscale mobility networks and the spatial spreading of infectious diseases. *Proceedings of the National Academy of Sciences* **2009**, *106*, 21484–21489.
20. Tamerius, J.; Nelson, M.I.; Zhou, S.Z.; Viboud, C.; Miller, M.A.; Alonso, W.J. Global influenza seasonality: reconciling patterns across temperate and tropical regions. *Environmental health perspectives* **2011**, *119*, 439.
21. Tsou, M.H. Research challenges and opportunities in mapping social media and Big Data. *Cartography and Geographic Information Science* **2015**, *42*, 70–74.
22. Zheng, Y.; Xie, X.; Ma, W.Y. GeoLife: A Collaborative Social Networking Service among User, Location and Trajectory. **2010**.
23. Becker, R.; Cáceres, R.; Hanson, K.; Isaacman, S.; Loh, J.M.; Martonosi, M.; Rowland, J.; Urbanek, S.; Varshavsky, A.; Volinsky, C. Human mobility characterization from cellular network data. *Communications of the ACM* **2013**, *56*, 74–82.
24. Sobolevsky, S.; Szell, M.; Campari, R.; Couronné, T.; Smoreda, Z.; Ratti, C. Delineating geographical regions with networks of human interactions in an extensive set of countries. *PLoS One* **2013**, *8*, e81707.
25. Cranshaw, J.; Schwartz, R.; Hong, J.I.; Sadeh, N.M. The Livehoods Project: Utilizing Social Media to Understand the Dynamics of a City. ICWSM, 2012.
26. Mitchell, L.; Frank, M.R.; Harris, K.D.; Dodds, P.S.; Danforth, C.M. The geography of happiness: Connecting twitter sentiment and expression, demographics, and objective characteristics of place **2013**.
27. Longley, P.A.; Adnan, M.; Lansley, G.; others. The geotemporal demographics of Twitter usage. *Environment and Planning A* **2015**, *47*, 465–484.
28. Hägerstrand, T.; others. Time-geography: focus on the corporeality of man, society, and environment. *The science and praxis of complexity* **1985**, pp. 193–216.
29. Kwan, M.P.; Lee, J. Geovisualization of human activity patterns using 3D GIS: a time-geographic approach. *Spatially integrated social science* **2004**, *27*.
30. Andrienko, N.; Andrienko, G. Designing visual analytics methods for massive collections of movement data. *Cartographica: The International Journal for Geographic Information and Geovisualization* **2007**, *42*, 117–138.
31. MacEachren, A.M.; Kraak, M.J. Research challenges in geovisualization. *Cartography and Geographic Information Science* **2001**, *28*, 3–12.
32. MacEachren, A.M. *How maps work: representation, visualization, and design*; Guilford Press, 2004.
33. Andrienko, G.; Andrienko, N.; Wrobel, S. Visual analytics tools for analysis of movement data. *ACM SIGKDD Explorations Newsletter* **2007**, *9*, 38–46.
34. Cao, G.; Wang, S.; Hwang, M.; Padmanabhan, A.; Zhang, Z.; Soltani, K. A Scalable Framework for Spatiotemporal Analysis of Location-based Social Media Data. *arXiv preprint arXiv:1409.2826* **2014**.
35. Shvachko, K.; Kuang, H.; Radia, S.; Chansler, R. The hadoop distributed file system. Mass Storage Systems and Technologies (MSST), 2010 IEEE 26th Symposium on. IEEE, 2010, pp. 1–10.
36. Dean, J.; Ghemawat, S. MapReduce: simplified data processing on large clusters. *Communications of the ACM* **2008**, *51*, 107–113.
37. Gao, H.; Tang, J.; Liu, H. Exploring Social-Historical Ties on Location-Based Social Networks. ICWSM, 2012.
38. Buttenfield, B.P.; McMaster, R.B. *Map Generalization: Making rules for knowledge representation*; Longman Scientific & Technical New York, 1991.
39. Samet, H. The quadtree and related hierarchical data structures. *ACM Computing Surveys (CSUR)* **1984**, *16*, 187–260.
40. Clauset, A.; Shalizi, C.R.; Newman, M.E. Power-law distributions in empirical data. *SIAM review* **2009**, *51*, 661–703.
41. Reynolds, A. Truncated Lévy walks are expected beyond the scale of data collection when correlated random walks embody observed movement patterns. *Journal of The Royal Society Interface* **2012**, *9*, 528–534.

42. Zhao, K.; Musolesi, M.; Hui, P.; Rao, W.; Tarkoma, S. Explaining the power-law distribution of human mobility through transportation modality decomposition. *Scientific reports* **2015**, *5*.
43. Liu, Y.; Padmanabhan, A.; Wang, S. CyberGIS Gateway for enabling data-rich geospatial research and education. *Concurrency and Computation: Practice and Experience* **2014**.