Exploring Multi-scale Spatiotemporal Twitter User Mobility Patterns Using Geo-located Twitter Data

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 have extracted high-resolution mobility data from a collection of 1.3 billion geo-located Twitter messages to study multi-scale spatiotemporal Twitter user mobility patterns in the United States during the year 2014. Regarding the concerns of infringement on individual privacy, such as the mobile phone call data with restricted access, the dataset is collected from publicly accessible Twitter data streams. In this study, we have developed 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 have performed a set of statistical analysis to understand Twitter user mobility patterns across multi-level spatial scales and temporal granularity. In particular, the Twitter user mobility patterns measured by the displacements and radius of gyrations of individuals have revealed different groups of Twitter users with multi-scale or multi-modal mobility patterns. 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.

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 (Zheng et al., 2008), traffic management (Jiang et al., 2009), and even the spatial spread of epidemic diseases (Belik et al., 2011). Earlier research efforts relied on low resolution mobility data to understand human mobility patterns, such as using census records to understand human migration patterns (Greenwood, 1985), or delivering questionnaires and asking volunteers to report the track of bank notes to infer human travel patterns (Brockmann et al., 2006). However, the lack of detailed human movements with fine-grained spatial and temporal granularity limits the findings to capture mobility patterns of individuals (Gonzalez et al., 2008; Jurdak et al., 2015). In addition to the mobility data collected by GPS trackers (Rhee et al., 2011; Zheng et al., 2008) and mobile phone call records (Gonzalez et al., 2008; Kung et al., 2014; Sevtsuk and Ratti, 2010), 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 (Thatcher, 2014) in the form of real-world trails and footprints. One significant advantage of LBSM data streams is the large spatial coverage, for example, researchers have used geo-located Twitter data for studying global mobility patterns (Hawelka et al., 2014), which is otherwise impossible by using other mobility datasets (e.g., GPS traces and mobile phone call records). In addition, the publicly available LBSM data streams offer unique opportunities for conducting replicable scientific findings regarding the concerns of infringement on individual privacy, such as using mobile phone call records (Crampton, 2014; Giannotti and Pedreschi, 2008; Jurdak et al., 2015).

Many 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 (Gonzalez et al., 2008) in the collective user movements to reveal different travel modes (Jurdak et al., 2015), travel demands (Hasan et al., 2013; Wu et al., 2014), and the impact of social connections (Cho et al., 2011). These studies have provided strong support for using LBSM data as proxies for studying mobility patterns of individuals and valuable insights into human mobility dynamics. However, in these studies, the measurements of distances are either fixed in a certain time period or a specific region, where they do not consider the variations of movements in different spatial scales and temporal ranges. To be more specific, these studies lack examinations of the detailed movements across different geographical scales and temporal granularity, for instances, whether there are temporal (e.g., monthly or seasonal) changes within the movements, and how the mobility patterns vary across different spatial scales (e.g., intra- or inter county, city or country level). These insights are critical to advance our understandings of the collective mobility patterns for a variety of applications, such as examining the mobility patterns across different cities (Noulas et al., 2012), the spread patterns of disease (Balcan et al., 2009; Tamerius et al., 2011) and touristic activities (Hawelka et al., 2014). 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 a multi-scale spatiotemporal analysis framework (Tsou, 2015) to deal with the complexities in filtering movements of individuals,

modeling and aggregating user trajectories at multiple spatial and temporal scales. In addition, to enable efficient analysis and visualization of the mobility patterns at different spatiotemporal scales, it is essential for such a framework to provide scalability for addressing the data intensive challenges (Cao et al., 2014).

In this paper, we have explored and studied the Twitter user mobility patterns across multi-level spatial scales and temporal granularity in the United States during the year 2014. The mobility data is extracted from1.3 billion geo-located Twitter messages (i.e., tweets) from 1^{st} January to 31^{st} December, 2014 in the United States with over 6 million Twitter users and over 1 TB in file size. To address the data-intensive challenges embedded in this dataset, 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. 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. With this framework, we have performed a set of statistical analysis to understand the 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 displacements and radius of gyrations of individuals (Gonzalez et al., 2008) have revealed different groups of Twitter users with multi-scale or multi-modal mobility patterns and multiple travel modes (Jurdak et al., 2015). 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.

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, geo-located Twitter data. We focus on research challenges in visual-analytics methods to enable multi-scale spatiotemporal analysis with massive movement datasets, including data management, multi-level spatiotemporal user trajectory modeling and visualization. In particular, we look into high performance distributed computing approach for addressing such challenges. 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 Twitter mobility patterns in the 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 human mobility patterns of individuals, the capability of capturing human movements with fine-grained spatial and temporal granularity is critical. However, the low-resolution mobility data collected from census records (Greenwood, 1985) is estimated and aggregated at census tract level, while tracking of bank notes (Brockmann et al., 2006) is at zip code level and do not necessarily reflecting movements of the same individuals (Gonzalez

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

et al., 2008). 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 (Zheng et al., 2008). However, the data is often limited in spatial scale (e.g. within a specific city or region) with a small group of people, for example, 182 and 226 volunteers participated in collecting such mobility data in (Zheng et al., 2010) and (Rhee et al., 2011) respectively. Other than tracking people directly, the vehicle-based GPS data is often tied to a specific vehicle (e.g. taxi), which may only be accessible for a certain group of people (Kung et al., 2014).

Another approach from the literatures for studying human mobility is using mobile phone call data in the form of Call Detail Records (CDR), where the locations of mobile users are estimated by cell tower triangulation with accuracy in the order of kilometers (Gonzalez et al., 2008; Kung et al., 2014; Sevtsuk and Ratti, 2010). Such a dataset can cover relatively large spatial scale (e.g., country level) (Becker et al., 2013; Sobolevsky et al., 2013) and a large portion of the population in the study region (Kung et al., 2014). However, due to the concerns of infringement on individual privacy, the mobile phone call data is not publicly accessible, which is not ideal for conducting replicable scientific findings, 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 will post, which is derived from either the GPS or Wi-Fi positioning with a high position resolution down to 10 meters (Jurdak et al., 2015). A Big Data scenario emerges when millions social media users constantly posting messages. However, there are some limitations and complexities in directly using the LBSM data for study 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. It has been criticized for the lack of representativeness of the population as not all people use social media or send geo-located messages (Kung et al., 2014). Another research challenge is to identify the users, as a social media account is not equal to a real person in the physical world (Tsou, 2015). Although many studies started to look into the demographic information of LBSM data, in particular Twitter data (Longley et al., 2015; Mitchell et al., 2013), these issues certainly require us to pose more strict criteria in filtering and extracting individual movements.

In this study, geo-located Twitter data is chosen as 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 (Cranshaw et al., 2012; Hasan et al., 2013). In addition, it presents some unique advantages that make it suitable for studying human mobility patterns. For example, the high-resolution location information enables to identify multiple travel modes in user mobility patterns (Jurdak et al., 2015); the large spatial coverage enables to study global

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

mobility patterns (Hawelka et al., 2014), 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).

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

Mobility data is essentially a collection of spatiotemporal records of people moving from one location to another in the geographic space. To study mobility patterns of individuals, a space-time trajectory (Hägerstrand et al., 1985) of each individual 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 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 (Hägerstrand et al., 1985), which has provided useful insights to explore movements across different geographical scales and temporal granularity. 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 (Kwan and Lee, 2004). While such an approach enables visualization of detailed trajectories, its capability is limited in dealing with large-volume movement datasets (Andrienko and Andrienko, 2007). Instead of directly visualizing individual user trajectories, a space-time cube approach was proposed in analyzing and visualizing the collective trajectories by providing flexibilities in setting up both spatial and temporal ranges, and therefore to study the mobility patterns across different spatial units (e.g. countries, states, and cities, etc.) and identify the changes over space and time (MacEachren, 2004; MacEachren and Kraak, 2001). In this connection, visual-analytics methods are proposed to help better convey the findings in terms of analyzing and visualizing multi-level spatiotemporal mobility patterns (Andrienko et al., 2007; Andrienko and Andrienko, 2007). 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 (Andrienko and Andrienko, 2007). Also, it is useful to group the subsets of user trajectories based on the shared characteristics for smooth transformations of space and time. For example, a network approach was used to summarize the actual movements in Twitter user trajectories as movement flows among different countries (Hawelka et al., 2014).

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. By conveying the findings from statistical analysis, it provides another visualization form regarding the corresponding mobility patterns, in a sense, to avoid the problems of visual clutters. However, in the context of studying mobility patterns using large volume of geo-located Twit-

ter data, the inherited large data volume poses significant data intensive challenges for these mentioned visual-analytics methods to scale with both the data volume and the computational requirements (e.g., movement extraction and trajectory modeling) (Cao et al., 2014). For example, 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. In particular, developing a multi-scale spatiotemporal analysis framework is identified as one of the research challenges for dealing with social media Big Data (Tsou, 2015). 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.

2.3 Accommodating LBSM Big Data with Hadoop

Apache Hadoop is an open source software framework designed to facilitate data intensive computing on large commodity clusters. It combines a distributed file system, namely Hadoop Distributed File System(HDFS) (Shvachko et al., 2010) with MapReduce programming paradigm, which can be applied to wide range of data-intensive problems. MapReduce is a programming model and an associated software framework designed to process massive data in a distributed fashion (Dean and Ghemawat, 2008). MapReduce breaks the entire computation into small tasks and schedule them among different computing nodes. MapReduce consists of two main stages: map and reduce. In the map stage, input data is converted into series of intermediate $\langle key, value \rangle$ pairs. Customized computation will be performed in the reduce stage based on the same intermediate key. This framework provides parallelization since both map and reduce tasks are considered independent and can be done in parallel. More importantly, it provides scalability in relation to the growth of data size, where Hadoop can scale to more computing nodes in the cluster to maintain the performance.

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. Second, by using Hadoop we can parallelize the computational tasks and make the data processing faster and more efficient. Further, by adopting the MapReduce computing paradigm, we can model each individual user's trajectory by treating user's unique ID as key, and summarize the movement flows among different spatial units by utilizing the ID of the spatial unit as key. The details will be introduced in the following section.

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. In our case, to ensure the complete coverage over the 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: -167.276413, longitude: 5.499550) and upper right (latitude: -52.233040, longitude: 83.162102). The crawler is constantly running with over 2 million geo-located raw tweets (∼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 (Tsou, 2015), to ensure the data quality, the collected raw tweets were further filtered by the following steps: We first removed the duplicated messages[3] in the dataset; and then we attempted to remove non-human users based on unusual relocating speed (Hawelka et al., 2014; Jurdak et al., 2015). In this case, we adopted the speed limit value as 240 m/s used in (Jurdak et al., 2015), where we examined all the consecutive locations of each user and excluded those with relocating speed over the limit. Finally, we used the geographic boundaries of the 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 collection unit: a Twitter data crawler that continuously collects raw geo-located tweets using the Twitter Streaming API. (2) 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 spatiotemproal levels. An overall system architecture of the framework that chains the three units is shown in Figure 1. The details regarding the function and implementation of each unit are presented in the next sections.

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

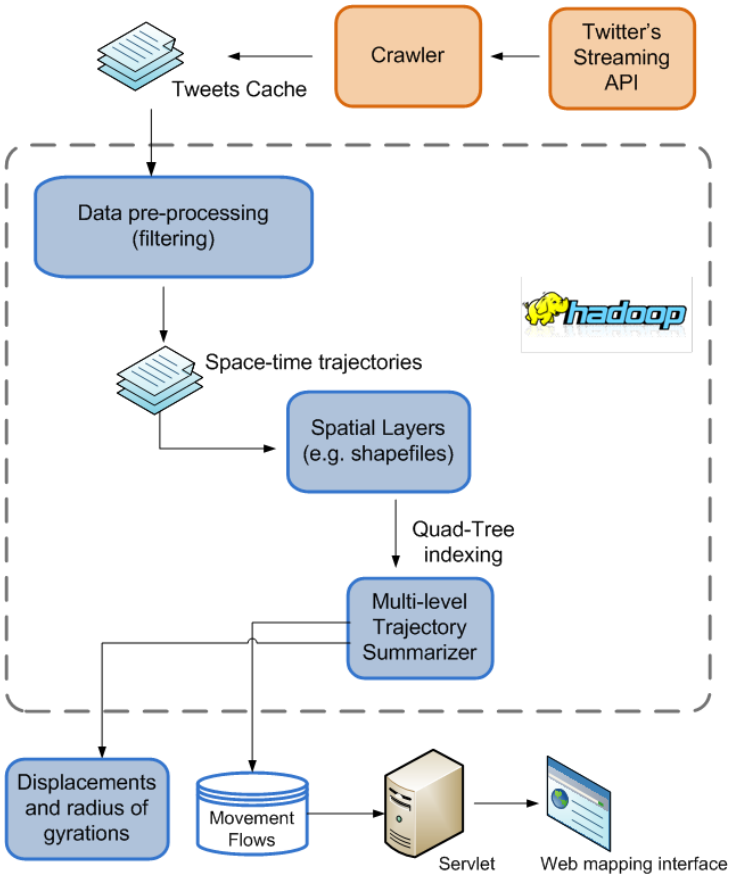

Figure 1: The overall system architecture of the framework

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 (Hägerstrand et al., 1985). 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 coordinates $\langle latitude, longitude \rangle$; 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 $speed\ (loc_i - loc_{i-1}) > 240m/s$.

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 can provide an overall assessments regarding the diversity and similarity in the collective mobility pattern (Gao et al., 2012).

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 (Gonzalez et al., 2008). The displacement refers to an individual's re-allocation in the geographical 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 all the displacements from a group people, it helps to identify the distance bounds associated with different travel modes (Jurdak et al., 2015) or to determine whether movements are random walks (Brockmann et al., 2006). Radius of gyration is a metric to distinguish mobility patterns of individuals (Gonzalez et al., 2008), 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, 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 measurements to all the users, it helps to identify different groups of people and understand the 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 distance between two recorded locations). Since all these metrics are based on the generated trajectories, by breaking and aggregating the trajectories in multiple spatial scales and temporal granularity, it would enable performing multi-level spatiotemporal analysis of 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 (Andrienko et al., 2007; Andrienko and Andrienko, 2007). 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 (Buttenfield and McMaster, 1991). To enable aggregating Twitter trajectories into multiple spatial levels, we have extended the hierarchical space-time cube model developed in (Cao et al., 2014), where we partitioned the geographic space of the United States into 10 hierarchical spatial layers. To be specific, the state boundaries of the 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 1km $\times$ 1km 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 fine-grained geographic boundaries can be used and appended in this framework to show different level-of-detailed movements (e.g., county-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" process 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 United States to filer the raw tweet collection and keep 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 (Samet, 1984) 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 volumes among the units. The movement flows are stored in the database for interactive explorations in the 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 $i = 1, 2, 3...n$

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

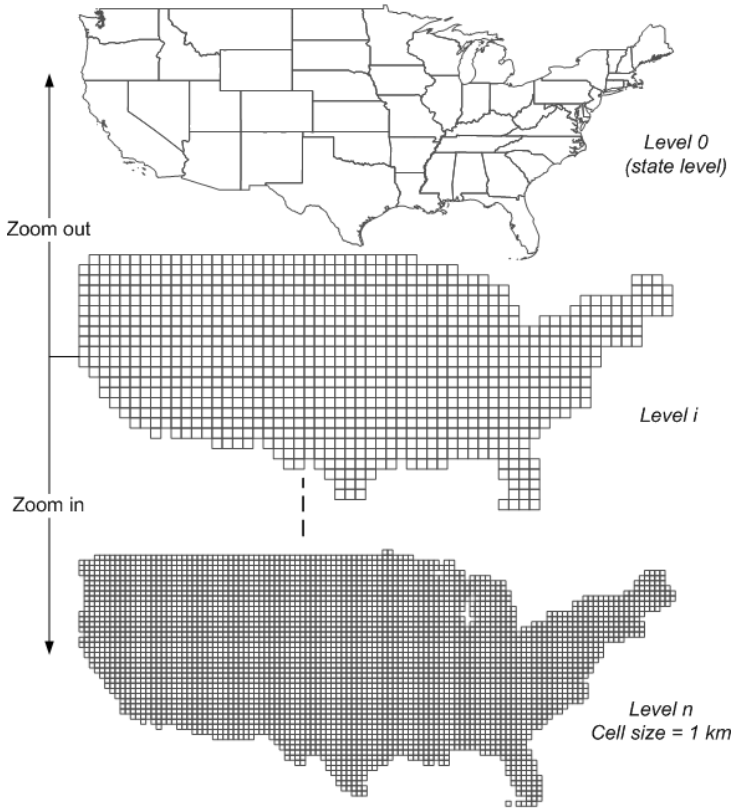

Figure 2: Hierarchical spatial layers for aggregating movements in different level-of-details

4 Multi-scale spatiotemporal Twitter user mobility patterns

4.1 An overview of the Twitter user population

To investigate the overall geographical distribution of the Twitter population in our dataset, we have carried out analysis at both county and state level to test their correlations and stability regarding the census population estimates. For this study, we have derived the census information of 48 states plus Washington, D.C., and 3108 counties of the United States (i.e., excluding Alaska and Hawaii) of year 2014 from the US census bureau[6]. For every state and county, the framework summarizes the number of unique users in each unit for the year 2014 (as well as every month). The results of the correlations between the number of unique Twitter user and the census population estimates in each state is show in Figure 3.

[6]http://www.census.gov/

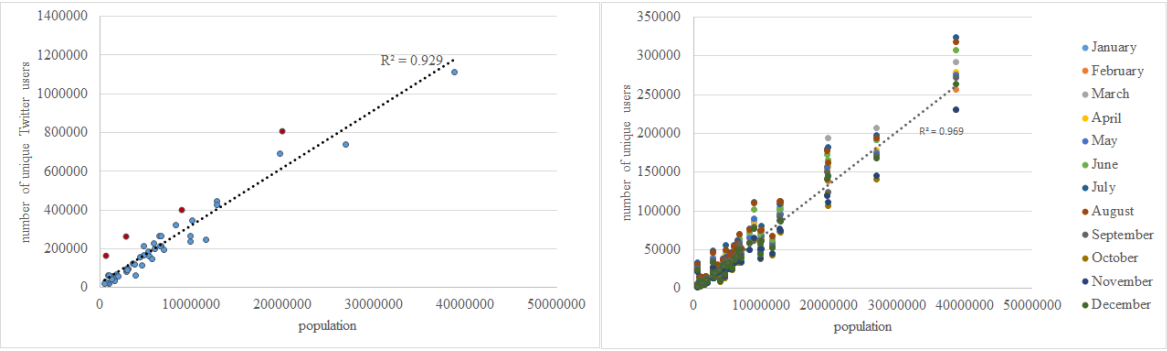

Figure 3: The correlation between the number of unique Twitter user and the census population estimates in each state (a) based on year 2014 and (b) based on each month of year 2014

The results show that the number of unique Twitter users is well ($R^2 = 0.92$ and p-value < 0.01) correlated with the state level census population estimates in year 2014 (Figure 3(a)) and it is consistent (with all $R^2 > 0.95$ and p-value < 0.01) in each month (Figure 3(b)). However, such correlations are not significant at the county level. Interestingly, zero Twitter population were found in 4 (out of 3108) counties (i.e., Manassas Park city, Covington city, Cumberland county, Lipscomb county) throughout the year, which means no geo-located tweets were captured from those areas. It is not clear whether people in those 4 counties do not send geo-located tweets at all or the tweets were lost due to 1 % policy mentioned in (Hawelka et al., 2014) as both cases are unlikely. Nevertheless, this does show the limitation of directly using Twitter population as proxies of real population. In addition, the dramatic difference between the correlations at state level and the county level could be useful for other researchers when studying population dynamics at a specific spatial scale.

4.2 Spatiotemporal Twitter user mobility patterns

The situation of using geo-located tweets to track 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 (Gonzalez et al., 2008). 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, we need 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 (Clauset et al., 2009). The frequency is summarized based on the trajectories of individuals extracted in each time range. 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 4), 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 $\pm$ 0.05 (standard deviation) and the mean λ value as 0.00178 $\pm$ 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 (Reynolds, 2012; Rhee et al., 2011), which has also been identified in many human mobility studies using different mobility data (Zhao et al., 2015). 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 studies collected geo-located tweets in one month can potentially reveal similar findings as the ones collected in multiple months. Of course, different spatial scales can effect such findings as we learned that there were no geo-located tweets collected from 4 counties. 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 rest (2) within each group, the probability of Twitter user visiting more locations decreases significantly with a power function.

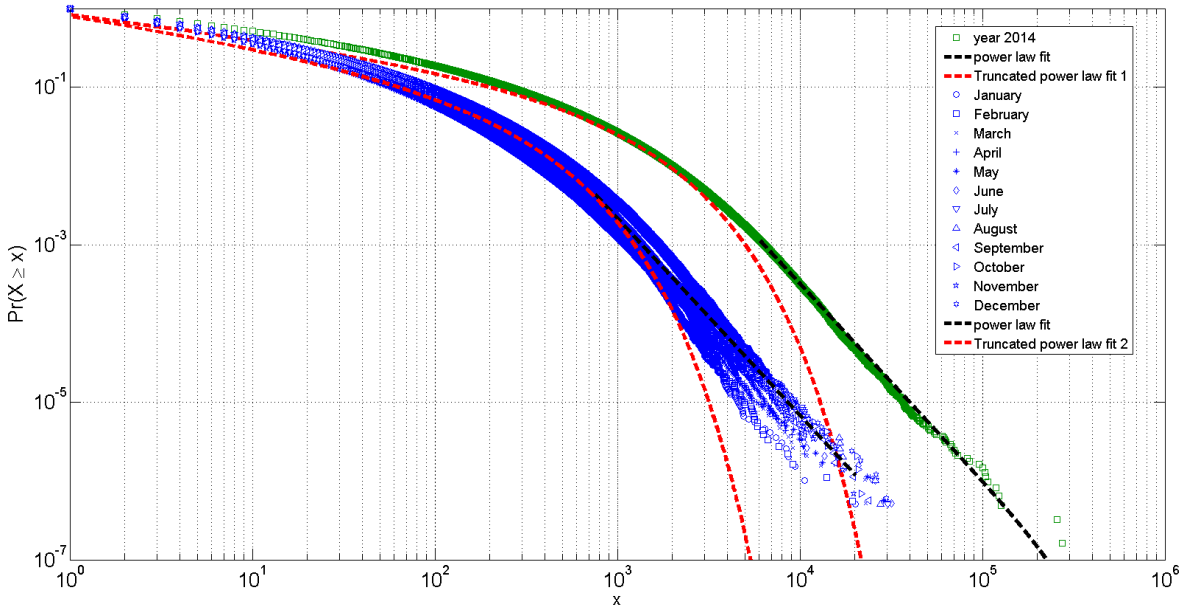

Figure 4: Two-tier power law distribution of the collective Twitter user visitation behaviors

As it is mentioned earlier, 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 (Gonzalez et al., 2008). In this case, we first gathered the displacements from all the collected Twitter users in the United States (excluding Hawaii and Alaska) in year 2014, where those Twitter users with only one geo-located tweet were filtered out. In addition, to investigate the mobility patterns of individuals, we 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 measured by "crow's fly distance", which is the direct 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 $P(d)$ (Brockmann et al., 2006). The probability distribution of the user displacements (as well as accumulated displacements) is shown in Figure 5, whereas the probability distribution of radius of gyrations is shown in Figure 6. In this study, we used the fitting methods developed by (Jurdak et al., 2015). 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 5 (a), the probability distributions of overall displacements is approximated by $P(d) \sim \lambda_1 e^{-\lambda_1(d-d_{min})}, d_{min} = 10m$ from [10 m, 70m] (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 [100m, 80km] (accounting for 93 % of the population), and $P(d) \sim d^{-\alpha}$ [$>$ 80km] (accounting for 5 % of the population). In additon, the displacement in the distance bound from 100m and 80km in Figure 5 (b) can be further approximately by two power-law distributions with a cutting point at 5km (53% distances are less than 5km and 40% distances between 5km and 80km), which indicates two different travel modes, such as inter- or intra-city movements. Overall, the fitting functions with different distance bound suggest the existence of multi-scale or multi-modal mobility patterns (Jurdak et al., 2015) of the Twitter users in the United States, for example the displacements larger than 80km could be related to inter-state travels or travel by flight.

The probability distribution of radius of gyrations of individuals (Figure 6 (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, 50m], $P(r_g) \sim \lambda_2 e^{-\lambda_2(r_g - r_{g_{min}})}$ from [50m, 30km], and $P(r_g) \sim {r_g}^{-\alpha}$ [$>$ 30km]. In particular, the radius of gyration between 50m and 30km can be further approximately by two power law distributions with a cutting point at 6km, which suggest two main types of spatial coverage of from the collected Twitter users in the 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.

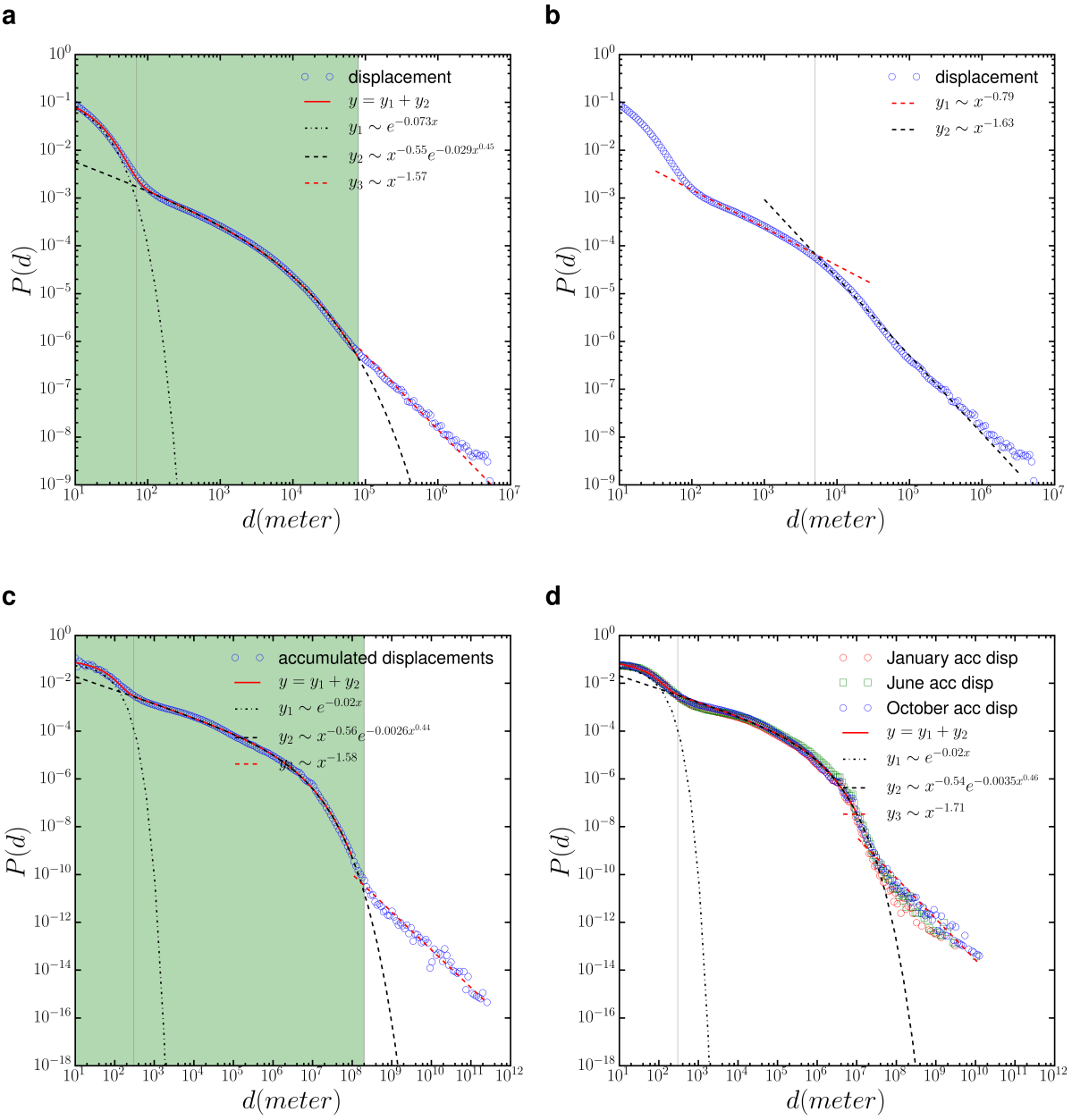

Figure 5: (a) The probability distribution of the collective Twitter user displacements P(d) (b) the distance between [100m, 80km] is approximated by a double power-law functions (c) The probability distribution of the accumulated displacements of individual Twitter users P(d) (d) The probability distribution of the accumulated displacements of individual Twitter users in 3 different months

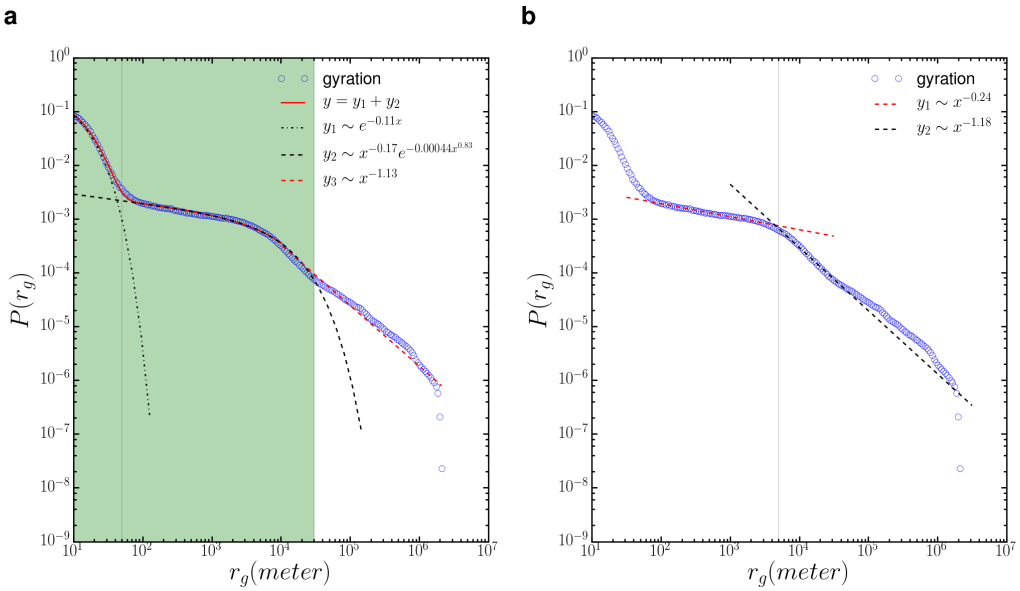

Figure 6: (a) The probability distribution of radius of gyration of individual Twitter users P(r_g) (b) the distance between [50m, 30 km] is approximated by a double power-law functions

More importantly, 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 5 (d)) and radius of gyrations within 4 quarters in year 2014 (Figure 7), 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 with the same measurements in (Jurdak et al., 2015), (slightly) different distance bounds were identified for describing the Twitter user mobility patterns in the USA and Australia, 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 for conducting replicable human mobility studies.

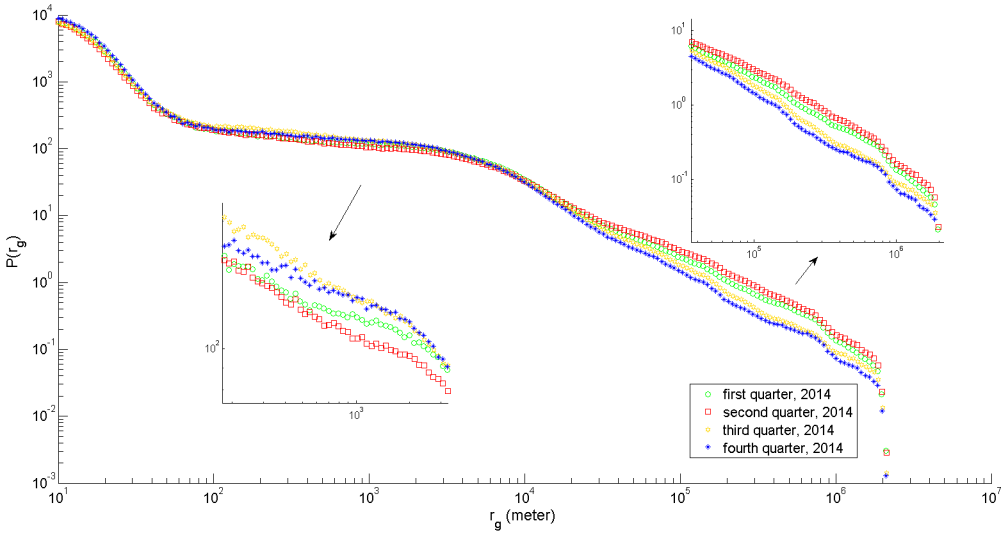

Figure 7: The probability distribution of radius of gyration of individual Twitter users in different quarters of year 2014

The above analysis of Twitter user mobility patterns mainly focus on the temporal aspects where the spatial scale is fixed to the country level. 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 among these spatial regions. In this case, we demonstrate the inter-state mobility patterns by using the framework to capture the movement flows among the states. Note that the movement flows can be summarized across all the 10 spatial layers in the framework. In particular, the movement flows in the state level are summarized and visualized with chord diagrams (illustrated in Figure 10) as part of the interactive mapping interface. 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 also indicates that although the Twitter population seems to be linearly correlated with the population at state level, it is not proportional to account the movement flux among the states, which may provide some insights for other researchers in studying social-economical aspects of the migration dynamics.

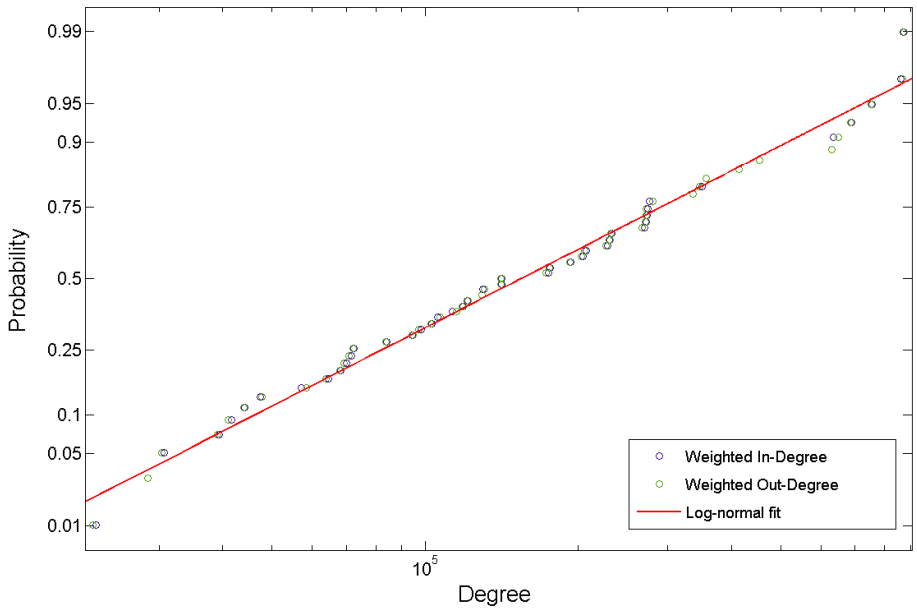

Figure 8: The distribution of Twitter user movement flows among different states in year 2014 measured in weighted in- and out-degrees

For state level Twitter user movements, we provide an interactive chord diagram to explicitly illustrate the movement flows among different states (Figure 9). Figure 9 demonstrates the Twitter movement flows among different states in January, 2014. The size of each color patch is proportional to the volume of total incoming movements from other states, which allow us to visually identify those states that dominate the incoming movement flows and which states the flows are generated from. By interacting with the state names, it will highlight the details of incoming movement flows from other states, in this case California is selected, which has the largest incoming movement flows in January.

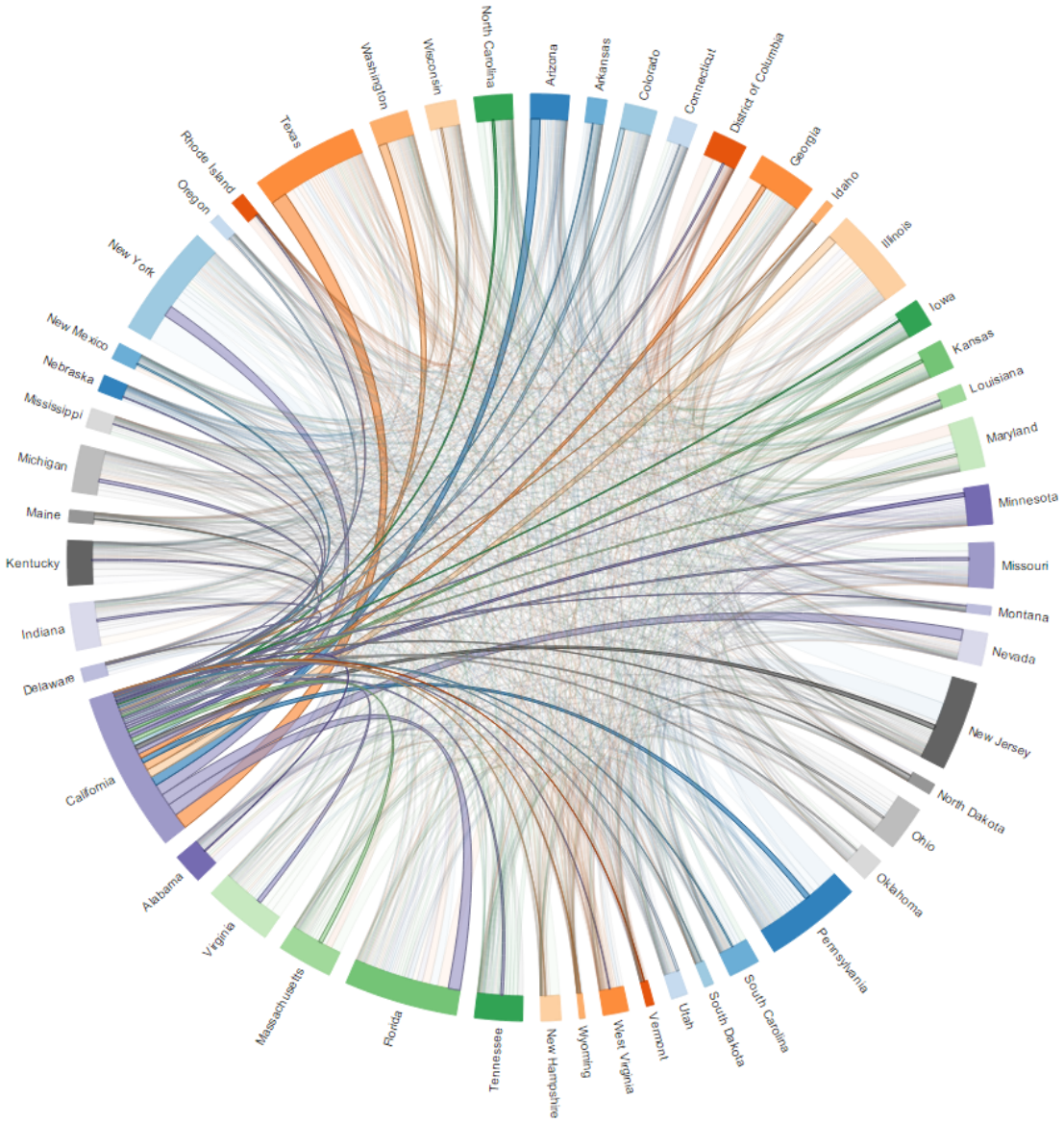

Figure 9: Chord diagrams for the movement flows among different states in January, 2014

5 Conclusions and Discussions

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 granularity in the 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-level 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 statically and visually. We first examined the Twitter population by exploring the correlations

between the number of unique users and the census population estimates at both state and county level (excluding Hawaii and Alaska). Although these two correlate well at the state level, there is no significance at the county level. Interestingly, no geo-located Twitter messages were found in 4 (out of 3108) counties (i.e., Manassas Park city, Covington city, Cumberland county, Lipscomb county) throughout the year. We then 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 in the United States, whereas the distribution of radius of gyration reveals different groups of Tweet users with different types of spatial coverages. More importantly, we further studied these mobility patterns in different temporal scales 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. In particular, by comparing the mobility patterns revealed by displacements and radius of gyrations of Twitter users in the United States and Australia, there are slightly difference regarding the exact value of distance bounds between different fitting functions. However, the fact that these measurements are of people (i.e., Twitter users) in different countries (in this case, continent) and their probability distributions can all be approximated by a combination of three functions, has clearly show the values in using geo-located tweets for conducting replicable human mobility studies.

References

Andrienko, G., Andrienko, N., and Wrobel, S. (2007). Visual analytics tools for analysis of movement data. *ACM SIGKDD Explorations Newsletter*, 9(2):38–46.

Andrienko, N. and Andrienko, G. (2007). Designing visual analytics methods for massive collections of movement data. *Cartographica: The International Journal for Geographic Information and Geovisualization*, 42(2):117–138.

Balcan, D., Colizza, V., Gonçalves, B., Hu, H., Ramasco, J. J., and Vespignani, A. (2009). Multiscale mobility networks and the spatial spreading of infectious diseases. *Proceedings of the National Academy of Sciences*, 106(51):21484–21489.

Becker, R., Cáceres, R., Hanson, K., Isaacman, S., Loh, J. M., Martonosi, M., Rowland, J., Urbanek, S., Varshavsky, A., and Volinsky, C. (2013). Human mobility characterization from cellular network data. *Communications of the ACM*, 56(1):74–82.

Belik, V., Geisel, T., and Brockmann, D. (2011). Natural human mobility patterns and spatial spread of infectious diseases. *Physical Review X*, 1(1):011001.

Brockmann, D., Hufnagel, L., and Geisel, T. (2006). The scaling laws of human travel. *Nature*, 439(7075):462–465.

Buttenfield, B. P. and McMaster, R. B. (1991). *Map Generalization: Making rules for knowledge representation*. Longman Scientific & Technical New York.

Cao, G., Wang, S., Hwang, M., Padmanabhan, A., Zhang, Z., and Soltani, K. (2014). A scalable framework for spatiotemporal analysis of location-based social media data. *arXiv preprint arXiv:1409.2826*.

Cho, E., Myers, S. A., and Leskovec, J. (2011). Friendship and mobility: user movement in location-based social networks. In *Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 1082–1090. ACM.

Clauset, A., Shalizi, C. R., and Newman, M. E. (2009). Power-law distributions in empirical data. *SIAM review*, 51(4):661–703.

Crampton, J. W. (2014). Collect it all: national security, big data and governance. *GeoJournal*, pages 1–13.

Cranshaw, J., Schwartz, R., Hong, J. I., and Sadeh, N. M. (2012). The livehoods project: Utilizing social media to understand the dynamics of a city. In *ICWSM*.

Dean, J. and Ghemawat, S. (2008). Mapreduce: simplified data processing on large clusters. *Communications of the ACM*, 51(1):107–113.

Gao, H., Tang, J., and Liu, H. (2012). Exploring social-historical ties on location-based social networks. In *ICWSM*.

Giannotti, F. and Pedreschi, D. (2008). *Mobility, data mining and privacy: Geographic knowledge discovery*. Springer Science & Business Media.

Gonzalez, M. C., Hidalgo, C. A., and Barabasi, A.-L. (2008). Understanding individual human mobility patterns. *Nature*, 453(7196):779–782.

Greenwood, M. J. (1985). Human migration: Theory, models, and empirical studies. *Journal of regional Science*, 25(4):521–544.

Hägerstrand, T. et al. (1985). Time-geography: focus on the corporeality of man, society, and environment. *The science and praxis of complexity*, pages 193–216.

Hasan, S., Zhan, X., and Ukkusuri, S. V. (2013). Understanding urban human activity and mobility patterns using large-scale location-based data from online social media. In *Proceedings of the 2nd ACM SIGKDD international workshop on urban computing*, page 6. ACM.

Hawelka, B., Sitko, I., Beinat, E., Sobolevsky, S., Kazakopoulos, P., and Ratti, C. (2014). Geolocated twitter as proxy for global mobility patterns. *Cartography and Geographic Information Science*, 41(3):260–271.

Jiang, B., Yin, J., and Zhao, S. (2009). Characterizing the human mobility pattern in a large street network. *Physical Review E*, 80(2):021136.

Jurdak, R., Zhao, K., Liu, J., AbouJaoude, M., Cameron, M., and Newth, D. (2015). Understanding human mobility from twitter. *PLoS ONE*, 10(7):e0131469.

Kung, K. S., Greco, K., Sobolevsky, S., and Ratti, C. (2014). Exploring universal patterns in human home-work commuting from mobile phone data. *PloS one*, 9(6):e96180.

Kwan, M.-P. and Lee, J. (2004). Geovisualization of human activity patterns using 3d gis: a time-geographic approach. *Spatially integrated social science*, 27.

Longley, P. A., Adnan, M., Lansley, G., et al. (2015). The geotemporal demographics of twitter usage. *Environment and Planning A*, 47(2):465–484.

MacEachren, A. M. (2004). *How maps work: representation, visualization, and design*. Guilford Press.

MacEachren, A. M. and Kraak, M.-J. (2001). Research challenges in geovisualization. *Cartography and Geographic Information Science*, 28(1):3–12.

Mitchell, L., Frank, M. R., Harris, K. D., Dodds, P. S., and Danforth, C. M. (2013). The geography of happiness: Connecting twitter sentiment and expression, demographics, and objective characteristics of place.

Noulas, A., Scellato, S., Lambiotte, R., Pontil, M., and Mascolo, C. (2012). A tale of many cities: universal patterns in human urban mobility. *PloS one*, 7(5):e37027.

Reynolds, A. (2012). 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*, 9(68):528–534.

Rhee, I., Shin, M., Hong, S., Lee, K., Kim, S. J., and Chong, S. (2011). On the levy-walk nature of human mobility. *IEEE/ACM transactions on networking (TON)*, 19(3):630–643.

Samet, H. (1984). The quadtree and related hierarchical data structures. *ACM Computing Surveys (CSUR)*, 16(2):187–260.

Sevtsuk, A. and Ratti, C. (2010). Does urban mobility have a daily routine? learning from the aggregate data of mobile networks. *Journal of Urban Technology*, 17(1):41–60.

Shvachko, K., Kuang, H., Radia, S., and Chansler, R. (2010). The hadoop distributed file system. In *Mass Storage Systems and Technologies (MSST), 2010 IEEE 26th Symposium on*, pages 1–10. IEEE.

Sobolevsky, S., Szell, M., Campari, R., Couronné, T., Smoreda, Z., and Ratti, C. (2013). Delineating geographical regions with networks of human interactions in an extensive set of countries. *PloS one*, 8(12):e81707.

Tamerius, J., Nelson, M. I., Zhou, S. Z., Viboud, C., Miller, M. A., and Alonso, W. J. (2011). Global influenza seasonality: reconciling patterns across temperate and tropical regions. *Environmental health perspectives*, 119(4):439.

Thatcher, J. (2014). Living on fumes: Digital footprints, data fumes, and the limitations of spatial big data. *International Journal of Communication*, 8:1765–1783.

Tsou, M.-H. (2015). Research challenges and opportunities in mapping social media and big data. *Cartography and Geographic Information Science*, 42(sup1):70–74.

Wu, L., Zhi, Y., Sui, Z., and Liu, Y. (2014). Intra-urban human mobility and activity transition: Evidence from social media check-in data. *PloS one*, 9(5):e97010.

Zhao, K., Musolesi, M., Hui, P., Rao, W., and Tarkoma, S. (2015). Explaining the power-law distribution of human mobility through transportation modality decomposition. *Scientific reports*, 5.

Zheng, Y., Li, Q., Chen, Y., Xie, X., and Ma, W.-Y. (2008). Understanding mobility based on gps data. In *Proceedings of the 10th international conference on Ubiquitous computing*, pages 312–321. ACM.

Zheng, Y., Xie, X., and Ma, W.-Y. (2010). Geolife: A collaborative social networking service among user, location and trajectory.