

A New Understanding of Friendships in Space: Complex Networks Meet Twitter

Won-Yong Shin, Bikash C. Singh, Jaehee Cho, and André M. Everett

Abstract

Studies on friendships in online social networks involving geographic distance have so far relied on the city location provided in users' profiles. Consequently, most of the research on friendships have provided accuracy at the *city level*, at best, to designate a user's location. This study analyzes a Twitter dataset because it provides the exact geographic distance between corresponding users. We start by introducing a strong definition of "*friend*" on Twitter (i.e., a definition of *bidirectional friendship*), requiring bidirectional communication. Next, we utilize *geo-tagged mentions* delivered by users to determine their locations, where "@username" is contained anywhere in the body of tweets. To provide analysis results, we first introduce a friend counting algorithm. From the fact that Twitter users are likely to post consecutive tweets in the static mode, we also introduce a two-stage distance estimation algorithm. As the first of our main contributions, we verify that the number of friends of a particular Twitter user follows a well-known power-law distribution (i.e., a Zipf's distribution or a Pareto distribution). Our study also provides the following newly-discovered friendship degree related to the issue of space: The number of friends according to distance follows a *double power-law* (i.e., a *double Pareto law*) distribution, indicating that the probability of befriending a particular Twitter user is significantly reduced beyond a certain geographic distance between users, termed the *separation point*. Our analysis provides concrete evidence that Twitter can be a useful platform for assigning a more accurate scalar value to the degree of friendship between two users.

Index Terms

Befriend, bidirectional friendship, complex network, double power-law, geo-tagged mention, separation point, Twitter.

This research was supported by the Basic Science Research Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Education (2014R1A1A2054577).

W.-Y. Shin (corresponding author) and B. C. Singh are with the Department of Computer Science and Engineering, Dankook University, Yongin 448-701, Republic of Korea (E-mail: wyshin@dankook.ac.kr; bikash070@gmail.com).

J. Cho is with the Department of Business Administration, Kwangwoon University, Seoul 139-701, Republic of Korea (E-mail: misl@kw.ac.kr)

A. M. Everett is with the Department of Management, University of Otago, Dunedin 9054, New Zealand (E-mail: andre.everett@otago.ac.nz).

I. INTRODUCTION

In recent years, research in the field of online social networks (OSNs) has grown dramatically with the evolution of technologies while harnessing Big Data. Focusing on the relationships (edges) among users or profiles (vertices), OSN analysis has emerged as one of the most popular and familiar approaches for examining interaction, information sharing, and collaboration among online users [1]. Simultaneously, the field of *complex networks* has emerged as an independent research area, with strong connections to random graph theory from mathematics as well as to social network analysis by physicists, interested in understanding the behaviors of large-scale interacting networks. Based on massive datasets of large-scale real-world OSNs such as Twitter [2], Facebook [3], Flickr [4], and Foursquare [5], extensive studies have validated that the small-world phenomenon (originally introduced by Watts and Strogatz [6]) and scale-free degree distribution,¹ which are the two most representative features of complex networks, nearly hold in OSNs [7]. Twitter is one of the most popular micro-blogs (or social media), allowing users to “tweet” about any topic within the 140-character limit and to “follow” others to receive their tweets. At the start of 2015, Twitter played a vital role in facilitating social contacts, boasting 284 million active users per month, publishing 500 million tweets daily from their web browsers and smart phones.²

A. Related Work

To understand the nature of friendships online with respect to geographic distance, some efforts have focused on users’ online profiles that include their city of residence [8], [9]. In [8], experimental results based on the LiveJournal social network³ demonstrated a close relationship between geographic distance and probability distribution of friendship, where the probability of befriending a particular user on LiveJournal is inversely proportional to the positive power of the number of closer users. Contrary to [8], based on the data collected from Tuenti,⁴ a Spanish social networking service, it was found in [9] that social interactions online are only weakly affected by spatial proximity, with other factors dominating.

However, the effect of distance on online social interactions has not yet been fully understood. In the previous studies, the geographic location points only to the location of users at a *city scale*. For this reason, the friendship degree distribution contains a background probability that is independent of geography due to the city-scale resolution [8], [9]. On the other hand, *geo-located Twitter* can provide high-precision location information down to 10 meters through the Global Positioning System (GPS) interface [10] of users’ smart phones while offering comprehensive metadata with a gigantic sample of the whole population.

For this reason, there is extensive and growing interest among researchers to understand a variety of social behaviors through geo-located Twitter, or, equivalently, geo-tagged tweets [11]–[19]. Even if geo-tagged tweets account for approximately 1% of the total amount [20], thanks to the increasing penetration of smart devices and mobile applications, the volume of geo-located Twitter has grown constantly and now forms an invaluable register for understanding human behavior and modelling the way people interact in space. In [11], along with geo-locations for collected tweets, analysis included how geo-related factors such as physical distance, frequency of air travel, national boundaries, and language differences affect formation of social ties on Twitter. In [12], it was found that the geo-locations of Twitter users across

¹A “small-world” network is a type of mathematical graph in which two arbitrary nodes (people) are connected by a short chain of intermediate links (friends), and a “scale-free” network is a network whose degree distribution follows a power-law.

²<https://about.twitter.com/company>

³<https://www.livejournal.com>

⁴<https://www.tuenti.com>

different countries considerably impact their participation in Twitter, their connectivity with other users, and the information that they exchange with each other. As another application, the use of geo-tagged tweets was evaluated as a complementary source of information for urban planning including i) a technique to determine land uses in a specific urban area based on tweeting patterns and ii) a technique to identify urban points of interest at places with high tweeting activity [13]. New approaches based on geo-tagged tweets were also proposed to find top vacation spots for a particular holiday by applying indexing, spatio-temporal querying, and machine learning techniques [14] and to detect unusual geo-social events by measuring geographical regularities of crowd behaviors [15].

Benefiting from the increasing availability of location information from geo-tagged tweets, there has been a steady push to understand individual human mobility [16]–[19], which is of fundamental importance for many applications to human and electronic virus prediction and traffic and population forecasting. Recent effort has focused on the studies of human mobility using tracking technologies such as mobile phones [21]–[24], GPS receivers [25], WiFi logging [26], Bluetooth [27], and RFID devices [28] as well as location-based social network check-in data [29], but these technologies involve privacy concerns or data access restrictions. In contrast, geo-tagged tweets can capture much richer features of human mobility. For example, in [16], global human mobility patterns were widely revealed, and a comparative study on the mobility characteristics of different countries was conducted. Furthermore, it was found in [17] that the geo-located Twitter data for Australia reveals multiple modes of human mobility from intra-site to metropolitan and inter-city movements. As another point of view, in [18], it was reported that in Australia, the gravity law is applicable for estimating human mobility by showing that mobility between an origin and its destination is proportional to the product of populations of these two places and is inversely proportional to the power-law of distance between them. In [19], the problem of labelling the places of a city based on collected spatio-temporal data was addressed, including i) to infer whether a place belongs to a certain category or not and ii) to choose the category of a place among a set of categories.

B. Main Contributions

In our work, we utilize *geo-tagged mentions* on Twitter, sent by users, to identify their exact location information. A ‘mention’ in Twitter consists of inclusion of “@username” anywhere in the body of tweets. From the fact that we tend to interact offline with people living very near to us, we derive as a natural extension the question whether geography and social relationships are inextricably intertwined on Twitter. Our research significantly differs from a variety of studies on human mobility in the literature [16]–[19], [21]–[29] since it is interested in how a pair of users interacts. To the best of our knowledge, such an attempt to analyze one-to-one friendship based on geo-located tweets (or mentions) has not yet been described in the literature.

As people normally spend a substantial amount of time online, data regarding these two dimensions (i.e., geography and online social relationships) are becoming increasingly precise, thus motivating us to build more reliable models to describe social interactions [30]. Previous studies have employed large amounts of data from diverse sources, such as smart devices and web-based applications, to examine how social data resources (e.g., photos on Flickr) are processed with tagging [31], [32]. Both a co-clustering approach [31] and a spatial ranking approach [32] have been introduced to discover meaningful relationships between a set of relevant resources and a set of tags. This paper goes beyond past research to determine how friendship patterns are geographically represented by Twitter, analyzing a single-source dataset (to avoid potential confounds) that contains a huge number of geo-tagged mentions from users

in i) the state of California in the United States (US) and Los Angeles (the most populous city in the state) and ii) the United Kingdom (UK) and London (the most populous city in the UK). These two location sets were selected as demographically comparable, yet distinct and geographically separated, leading adopters of Twitter with sufficient data to enable meaningful comparative analysis for our intentionally exploratory study (which will be specified in Section II). In this dataset, each mention record has a geo-tag (spatial information) and a timestamp (temporal information) indicating from where, when, and by whom the mention was sent. We propose and apply the following new framework, which establishes a more accurate friendship degree on Twitter, and a method to enable analysis based on geographic distance:

- To fully take into account the intensity of communication between users, we start our analysis by introducing a rather strong definition of “*friend*” on Twitter, i.e., a definition of *bidirectional friendship*, instead of naïvely considering the set of followers and followees (unidirectional terms). This definition requires bidirectional communication within a designated time frame to constitute a friendship.
- Using the above definition, we introduce a friend counting algorithm, which computes the distribution of the number of friends for each Twitter user.
- By showing that almost all Twitter users are likely to post consecutive tweets in the static mode, we propose a two-stage distance estimation method, where the geographic distance between two befriended users (denoted by Users u and v) based on our definition of bidirectional friendship is estimated by sequentially measuring the two senders’ locations. More specifically, the location of User u is recorded at the moment when User u sends a mention to User v , while the location of User v can also be recorded when User v sends a replied mention to User u at the next closest time, enabling estimation of the distance between Users u and v .

Note that the above definition is suitable for evaluating one-to-one bidirectional social interactions on Twitter since Twitter users tend to personally interact with only a few of their followers/followees by sending and receiving direct mentions. We would like to synthetically analyze how the geographic distance between Twitter users affects their interaction, based on our new framework. Our main contributions are as follows:

- Based on the definition of bidirectional friendship, we first verify that the number of friends of *one user* on Twitter follows a power-law distribution (i.e., a Zipf’s distribution [33] or a Pareto distribution [34]) even on Twitter, which is known to be asymptotically equivalent to the degree distribution of scale-free networks. This finding is consistent with the earlier results in other OSNs.
- Next, more interestingly, we characterize a newly-discovered probability distribution of the number of friends according to *geographic distance*, which does not follow a homogeneous power-law but, instead, a *double power-law* (i.e., a *double Pareto law* [35]). From this new finding, we identify not only two fundamentally *separate regimes*, termed the intra-city and inter-city regimes, which are characterized by two different power-laws in the distribution, but also the *separation point* between these regimes.

C. Organization

The rest of this paper is organized as follows. Section II describes the dataset, and Section III explains our analysis methodology. In Section IV, experimental results are presented by analyzing the number of friends of a particular user and the number of friends with respect to distance. Finally, we summarize the paper with some concluding remarks in Section V.

II. DATASET

We use a dataset collected from crawling the Twitter network via Twitter Streaming Application Programming Interface (API),⁵ which returns tweets matching a query provided by the Streaming API user. Although the Twitter Streaming API only returns at most a 1% sample of all the tweets produced at a given moment, it constitutes a valid representation of users' activity on Twitter when more specific parameter sets such as different users, geographic bounding boxes, and keywords are created (thereby enabling extraction of more data from the Streaming API) [20], [36]. It was found that the Streaming API returns an almost complete set of *geo-tagged* tweets despite sampling [20]. Thus, there is no doubt that this research is working with an almost complete sample of geo-located Twitter data.

In our work, we examined data from all possible devices (sources) that indicate the user's location information at the time that they access Twitter. The statistics based on our dataset demonstrate that a large majority of the Twitter users in our sample posted geo-tagged tweets through smart phones rather than web browsers on a desktop or laptop computer.⁶ This reveals that our dataset is much more inclined toward geo-tagged tweets (more rigorously, geo-tagged mentions) transmitted through the GPS interface.

The dataset consists of a huge amount of geo-tagged mentions recorded from Twitter users from September 22, 2014 to October 23, 2014 (about one month) in the following four large regions: California, Los Angeles, the UK, and London. Note that this short-term (one month) dataset is sufficient to examine how closely one user has recently interacted with another online (i.e., a personal online relationship between two users). The four regions in our dataset were selected since they are quite comparable at both the macro (state or country) and micro (city) scales in terms of i) area, ii) population density, and iii) Twitter popularity (e.g., the number of Twitter accounts or the number of posted tweets). The comparison between location sets for the aforementioned three representative attributes is summarized in TABLE I, divided according to the types of two geographic scales.⁷

The representative statistics of the collected dataset, such as the total number of mentions and the total number of senders, are also summarized by regional group in TABLE II. In this dataset, each mention record has a geo-tag and a timestamp indicating from where, when, and by whom the mention was sent. Based on this information, we are able to construct a user's location history denoted by a sequence $L = (x_{ki}, y_{ki}, t_i)$, where x_{ki} and y_{ki} are the x - and y -coordinates of User k at time t_i , respectively. The location information provided by the geo-tag is denoted by latitude and longitude, which are measured in degrees, minutes, and seconds.

Each mention on Twitter contains a number of entities that are distinguished by their attributed fields. For data analysis, we adopted the following five essential fields from the

⁵<https://dev.twitter.com/decs/streaming-apis>

⁶We note that smart devices and mobile applications enable us to provide high-precision location information through the built-in GPS interface. On the other hand, with the Geo-location API, web browsers can detect the users' approximate location information inferred from network signals such as IP address, WiFi, Bluetooth, MAC address, and GSM/CDMA cell ID, which are not guaranteed to return the users' actual location. Based on our dataset, it is found that 77.84% and 82.21% of Twitter users tend to post geo-tagged tweets in California and the UK, respectively, via iPhone and Android Phone, which are the smart phone types using the two most popular mobile platforms among all devices. It is also found that 90.52% and 81.14% of posted geo-tagged tweets tend to be recorded in California and the UK, respectively, via iPhone and Android Phone.

⁷<http://en.wikipedia.org/wiki/California>
http://en.wikipedia.org/wiki/United_Kingdom
http://en.wikipedia.org/wiki/Los_Angeles
<http://en.wikipedia.org/wiki/London>

http://semiocast.com/publications/2012_07_30_Twitter_reaches_half_a_billion_accounts_140m_in_the_US

TABLE I
COMPARISON OF THE LOCATION SETS.

(a) California versus UK (state scale or country scale)

Attribute	California	UK
Area (km ²)	423,970	243,610
Population density (population/km ²)	95.0	225.6
Global ranking among countries by the number of Twitter accounts	1st (US as whole country)	4th

(b) Los Angeles versus London (city scale)

Attribute	Los Angeles	London
Area (km ²)	1,302	1,572
Population density (population/km ²)	3,198	5,354
Global ranking among cities by the number of posted tweets (June 2012)	8th	3rd

TABLE II
STATISTICS OF THE DATASET: THE NUMBER OF MENTIONS AND UNIQUE USERS IN EACH REGION.

Region	Number of mentions	Number of users (senders)
California	2,349,901	217,439
Los Angeles	918,360	51,625
UK	3,721,716	612,368
London	614,045	58,046

metadata of mentions:⁸

- *user_id_str*: string representation of the sender ID
- *in_reply_to_user_id_str*: string representation of the receiver ID
- *lat*: latitude of the sender
- *lon*: longitude of the sender
- *created_at*: UTC/GMT time when the mention is delivered, i.e., the timestamp

Note that the two location fields, *lat* and *lon*, corresponds to spatial (geo-tagged) information while the last field, *created_at*, represents temporal (time-stamped) information.

III. RESEARCH METHODOLOGY

We start by introducing the following definition of “bidirectional friendship” on Twitter.

Definition 1 (Bidirectional friendship in Twitter): If two users send/receive direct mentions to/from each other (i.e., bidirectional personal communication occurs) within a designated amount of time, then they form a bidirectional friendship with each other.

Note that our definition differs from the conventional definition of “friend” on Twitter, which is referred to as a followee and thus represents a *unidirectional* relation [37], [38].⁹

⁸<https://dev.twitter.com/overview/api/tweets>

⁹Twitter shows a low level of reciprocity; 77.9% of user pairs with any link between a Twitter user and his/her follower are connected one-way, and only 22.1% exhibit a reciprocal relationship between them (i.e., two-way links) [2].

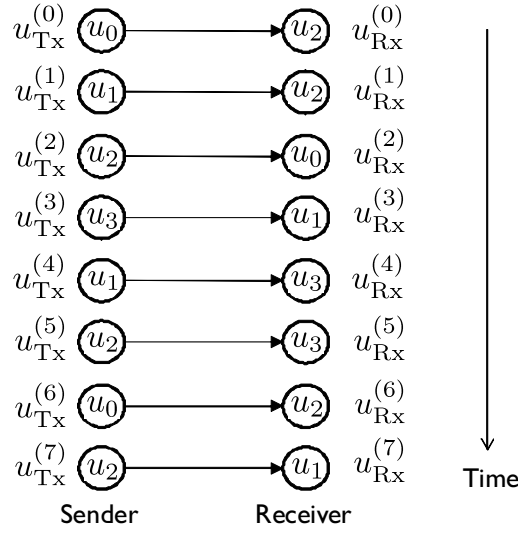


Fig. 1. One example that illustrates how geo-tagged mentions are delivered from senders to receivers according to time sequence, where $u_{Tx}^{(t)}$ and $u_{Rx}^{(t)}$ denote the transmitter and the corresponding receiver at time $t \in \{0, 1, \dots\}$. In this example, three pairs of friends, (u_0, u_2) , (u_1, u_2) , and (u_1, u_3) , are made among four users u_0 , u_1 , u_2 , and u_3 .

Since friendship relations in the offline world and on other OSNs such as Facebook [39] are generally not unidirectional, our intention is to formulate a *bidirectional* friendship that can be directly applicable to offline relationships. This strong definition enables exclusion of *inactive friends* (or passive friends) who have been out of contact online for a long designated amount of time (e.g., about one month in our work) and to count the number of *active friends* who have recently communicated with each other.

A. Counting Number of Friends of a Particular User

In this subsection, we explain how to count the number of friends of each user who sent at least one geo-tagged mention. Suppose that there are four Twitter users, denoted by u_0 , u_1 , u_2 , and u_3 , who sent or received at least one geo-tagged mention according to temporal event sequences, as illustrated in Figure 1. Here, $u_{Tx}^{(t)}$ and $u_{Rx}^{(t)}$ denote the transmitter and the corresponding receiver sequentially at time instance $t \in \{0, 1, \dots\}$. In this example, according to the aforementioned definition, three pairs of friends (u_0, u_2) , (u_1, u_2) , and (u_1, u_3) are found out of the above user set. Moreover, one can find that the number of friends of each user u_0 , u_1 , u_2 , and u_3 is given by 1, 2, 2, and 1, respectively. In our framework, if bidirectional communication between two certain users occurs at least once, then their friendship degree is set to one. Otherwise, it is set to zero, i.e., no friendship between the two users is created. That is, even with more than two bidirectional communications between two users, their friendship degree is maintained at one in this binary or Boolean evaluation. In our sample space, we exclude the user set whose friendship degree is zero since including such users will lead to scaling down the probability distribution of the nonzero number of friends.

The overall procedure of the friend counting algorithm (**Algorithm 1**) is described in TABLE III, where n_u denotes the number of friends of User $u \in \{u_0, u_1, \dots, u_{I-1}\}$ who sent a geo-tagged mention to User $v \in \{v_0, v_1, \dots, v_{J-1}\}$, and I and J are the total number of senders and receivers in a dataset, respectively.

TABLE III
THE OVERALL PROCEDURE OF THE FRIEND COUNTING ALGORITHM.

Algorithm 1 Friend counting algorithm
Input: $u_{Tx}^{(t)}$ and $u_{Rx}^{(t)}$ for $t = 0, 1, \dots, T - 1$, $u \in \{u_0, u_1, \dots, u_{I-1}\}$ and $v \in \{v_0, v_1, \dots, v_{J-1}\}$ Output: n_u for all u Initialization: $c_{uv} \leftarrow 0$ and $n_u \leftarrow 0$ for all u and v 00: for $t \leftarrow 0$ to $T - 1$ do 01: Find the user indices u and v for $u_{Tx}^{(t)}$ and $u_{Rx}^{(t)}$, respectively 02: for $s \leftarrow t + 1$ to $T - 1$ do 03: if $(u_{Tx}^{(s)} == u_{Rx}^{(t)})$ then 04: if $(u_{Rx}^{(s)} == u_{Tx}^{(t)})$ then 05: $c_{uv} \leftarrow 1$ 06: break (go back to line 00) 07: end if 08: end if 09: end for 10: end for 11: for all u and v do 12: $n_u \leftarrow n_u + c_{uv}$ 13: end for

B. Finding Friend Distribution With Respect to Distance

In this subsection, let us turn to characterizing the friendship degree of individuals regarding geography by analyzing their sequences $L = (x_{ui}, y_{ui}, t_i)$ of geo-tagged mentions, where only the senders' location information is recorded. We propose a two-stage method to estimate the geographic distance between Twitter friends. If User u sends a mention to User v , then the location information of User u is recorded (the first stage). In order to find the location of User v , we need to wait for the moment at which User v sends a mention back to User u (the second stage). That is, after bidirectional communication between two Twitter users occurs, the location of each user can be identified.

It is not possible to evaluate the geographic distance between two Twitter users through a one-shot process due to the fact that the location information of only the sender is recorded at a given instance when a geo-tagged mention is sent. Moreover, because of the users' movements, it is, however, not straightforward to measure the exact distance. In this subsection, we introduce a two-stage distance estimation method, where the geographic distance between two befriended users is estimated by sequentially measuring the two senders' locations.

Before describing the estimation algorithm, let us first focus on the time interval between the following two events for a befriended pair: a mention and its *replied* mention at the next closest time. We count only the events with a time duration between a mention and its replied mention, or inter-mention interval, of *less than one hour* to exclude certain inaccurate location information that may occur due to users' movements.¹⁰ Figure 2 illustrates the instance for which User u , originally placed at (x_{u0}, y_{u0}, t_0) , sent a mention to User v at (x_{v0}, y_{v0}, t_0) , and

¹⁰Note that inter-mention interval of one hour may be shortened, but this will lead to a reduction in the available dataset.

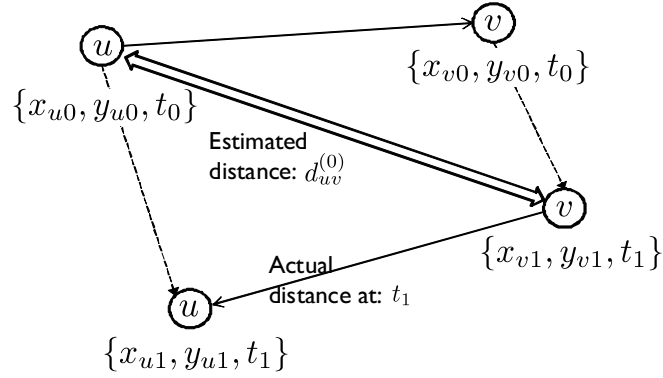


Fig. 2. User movement in which User $k \in \{u, v\}$ changes location from (x_{k0}, y_{k0}, t_0) to (x_{k1}, y_{k1}, t_1) between sending a geo-tagged mention and receiving a corresponding replied mention.

then received a replied mention at the location (x_{u1}, y_{u1}, t_1) from User v placed at (x_{v1}, y_{v1}, t_1) . Here, the single solid arrows indicate the actual distances at time instances t_0 and t_1 while the double solid arrow indicates the estimated distance. The distance that users moved between the two moments in time t_0 and t_1 (i.e., inter-mention interval) is indicated as the dashed arrows in the figure. From these two consecutive mention events, it is possible to estimate the geographic distance based on the two sequences (x_{u0}, y_{u0}, t_0) and (x_{v1}, y_{v1}, t_1) . In our framework, by assuming that the Earth is spherical, we deal with the shortest path between two users' locations measured along the surface of the Earth, instead of the rather naïve straight-line Euclidean distance. Following an approach similar to that employed in [40], [41], the distance between two locations on the Earth's surface can be computed according to the spherical law of cosines.¹¹ Then, when we denote the distance between the two users measured from (x_{u0}, y_{u0}, t_0) and (x_{v1}, y_{v1}, t_1) by $d_{uv}^{(0)}$, we obtain¹²

$$d_{uv}^{(0)} = R \cos^{-1} (\sin x_{u0} \sin x_{v1} + \cos x_{u0} \cos x_{v1} \cos (y_{v1} - y_{u0})), \quad (1)$$

where R [in kilometers (km)] denotes the Earth's radius and is given as 6,371, and the superscript 0 in $d_{uv}^{(0)}$ represents the time slot. Here, for notational convenience, it is assumed that the x - and y -coordinates represent the latitude and longitude, respectively.

While the estimated distance (double solid arrow in Figure 2) may differ from the actual distance (single solid arrow in Figure 2) between Users u and v at time t_1 , it is worth noting that people tend to send/receive multiple consecutive tweets from the same location to convey a series of ideas [17], [18]. To validate this user mobility argument, we turn our attention to analyze the distribution of the number of tweets (i.e., the tweet frequency) with respect to user velocity.

In our experiments, we use the same dataset collected from the Twitter users as shown in Section II, but focus on the two populous metropolitan areas, Los Angeles and London. To exclude certain inaccurate location information that may exist due to users' movements, we take into account only the case only where two consecutive geo-tagged tweet events occur *within*

¹¹When Sinnott published the haversine formula [42], computational precision was limited. Nowadays, JavaScript (and most modern computers and languages) uses IEEE 754 64-bit floating-point numbers, which provide 15 significant digits of precision. With this precision, the simple spherical law of cosines formula gives well-conditioned results down to distances as small as around 1 meter. In view of this, it is probably worth, in most situations, using the simpler law of cosines in preference to the haversine formula.

¹²<http://mathworld.wolfram.com/SphericalTrigonometry.html>

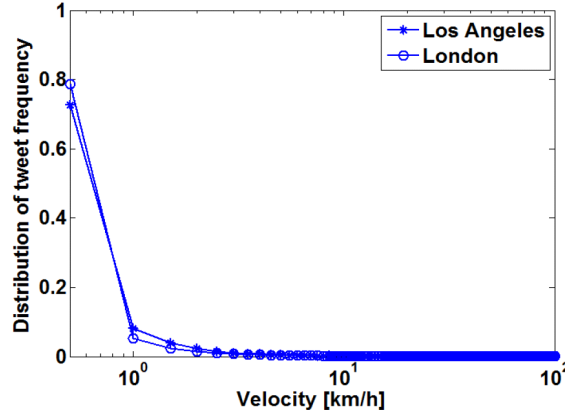


Fig. 3. Probability distribution of the tweet frequency with respect to user velocity (log-linear plot).

one hour. When the location history for two consecutive geo-tagged tweets of User k at time slots t_i and t_{i+1} is expressed as sequences (x_{ki}, y_{ki}, t_i) and $(x_{k(i+1)}, y_{k(i+1)}, t_{i+1})$, respectively, the average velocity $v_k^{(i)}$ of the user within this time interval is given by $v_k^{(i)} = d_k^{(i)} / (t_{i+1} - t_i)$, where $d_k^{(i)}$ is the distance that User k moved during the interval $[t_i, t_{i+1}]$ and thus is given by $d_k^{(i)} = R \cos^{-1} (\sin x_{ki} \sin x_{k(i+1)} + \cos x_{ki} \cos x_{k(i+1)} \cos (y_{k(i+1)} - y_{ki}))$ (refer to equation (1) for more details). From the set of average velocities $\{v_k^{(0)}, v_k^{(1)}, \dots, v_k^{(T-1)}\}$ obtained from all users in the dataset, the tweet frequency can be categorized according to the user velocity.

Figure 3 shows the log-linear plot of the distribution of the number of tweets (i.e., the tweet frequency) versus the user velocity [km/h], which is obtained from empirical data. As illustrated in Figure 3, most of the Twitter users (approximately 90%) in the two metropolitan areas are likely to post consecutive tweets in the *static* mode whose average velocity ranges from 0 to 2 km/h. Our experiments also demonstrate that Twitter users in large scale geographic areas (e.g., state scale (California) or country scale (the UK)) are more likely to post consecutive tweets in the static mode than city-scale users, even if the results are not presented in Figure 3. Although the inter-tweet interval may show a different pattern from that of the inter-mention interval (i.e., the time duration between a mention and its replied mention from another user), we believe that the above results are sufficient to support our analysis methodology.

Now, we are ready to present our distance estimation algorithm (**Algorithm 2**). The overall procedure of the proposed algorithm is described in TABLE IV, where d_{uv} denotes the estimated geographic distance between user pair $u \in \{u_0, u_1, \dots, u_{I-1}\}$ and $v \in \{v_0, v_1, \dots, v_{J-1}\}$, and I and J are the total number of senders and receivers in a dataset, respectively. Note that as shown in lines 14–18 of the table, the estimated distance for one pair is obtained by taking the average of all distance values computed over the available inter-mention intervals, each of which is less than one hour.

IV. ANALYSIS RESULTS

In this section, we first verify whether a Zipf's power-law holds for the Twitter network along with the definition of bidirectional friendship. Next, we show a newly-discovered distribution of the number of friends with respect to the geographic distance and then identify the two fundamentally separated regimes in the distribution.

TABLE IV
THE OVERALL PROCEDURE OF DISTANCE ESTIMATION ALGORITHM.

Algorithm 2 Distance estimation algorithm	
Input: $u_{Tx}^{(t)}$ and $u_{Rx}^{(t)}$ for $t = 0, 1, \dots, T-1$, $u \in \{u_0, u_1, \dots, u_{I-1}\}$ and $v \in \{v_0, v_1, \dots, v_{J-1}\}$	
Output: d_{uv} for all u and v	
Initialization: $c_{uv}^{(t)} \leftarrow 0$ and $d_{uv} \leftarrow 0$ for all u and v	
00: for $t \leftarrow 0$ to $T-1$ do	
01:	Find the user indices u and v for $u_{Tx}^{(t)}$ and $u_{Rx}^{(t)}$, respectively
02:	for $s \leftarrow t+1$ to $T-1$ do
03:	if $(u_{Tx}^{(s)} == u_{Rx}^{(t)})$ then
04:	if $(u_{Rx}^{(s)} == u_{Tx}^{(t)})$ then
05:	if (time interval between t and $s < 1$ hour) then
06:	Compute $d_{uv}^{(c_{uv}^{(t)})}$ in equation (1)
07:	$c_{uv}^{(t)} \leftarrow c_{uv}^{(t)} + 1$
08:	break (go back to line 00)
09:	end if
10:	end if
11:	end if
12:	end for
13:	end for
14:	for all u and v do
15:	for $l \leftarrow 0$ to $c_{uv}^{(t)}$ do
16:	$d_{uv} \leftarrow d_{uv} + d_{uv}^{(l)} / c_{uv}^{(t)}$
17:	end for
18:	end for

A. Number of Friends of a Particular User

We first find that the probability distribution $P_N(N = n)$ of the number of friends for an individual, denoted by n , on Twitter fits into a single power-law function $P_N(N = n) \sim n^{-\alpha}$ for $\alpha > 0$. Figure 4 shows the log-log plot of the distribution $P_N(N = n)$ obtained from empirical data, logarithmically binned data, and fitting function, where the fitting is applied to the binned data. As depicted in the figure, statistical noise exists in the tail where the number of friends is very large. Such noise can be eliminated by applying logarithmic binning, which averages out the data that fall in specific bins [43].¹³ We use the traditional least squares estimation to obtain the fitting function. In TABLE V, the value of the exponent of $P_N(N = n)$, α , is summarized for each region. From Figure 4 and TABLE V, the following interesting comparisons are performed according to types of regions:

- **Comparison between the city-scale and state-scale/country-scale results:** Figures 4(a) and 4(b) illustrate that the exponent α is 3.48 and 2.29 in California and Los Angeles, respectively, which implies that Twitter users in populous metropolitan areas are more

¹³It is also verified that this binning procedure does not fundamentally change the underlying power-law exponent of the distribution $P_N(N = n)$.

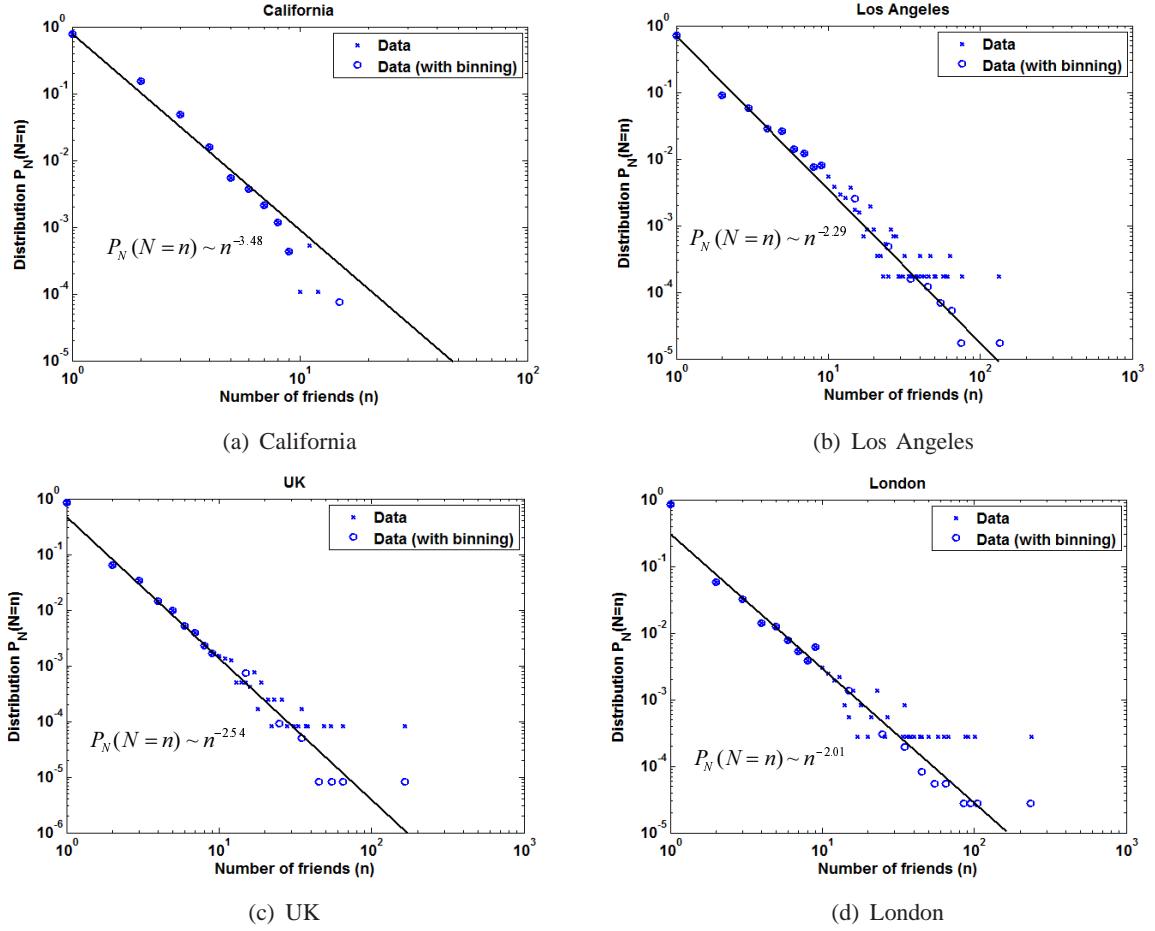


Fig. 4. Probability distribution $P_N(N = n)$ of the number of friends of a particular user (log-log plot).

TABLE V
THE VALUE OF α FOR EACH REGION.

Region	α
California	3.48
Los Angeles	2.29
UK	2.54
London	2.01

likely to contact a higher number of friends within a given period (e.g., one month). From Figures 4(c) and 4(d), the same trend is also observed by comparing the results for the UK and London, with α values of 2.54 and 2.01, respectively. That is, urban people are likely to bilaterally interact with more friends by sending and receiving directed geo-tagged mentions, compared on average to people in larger regions that include local small towns.

- **Comparison between the results in the two cities (Los Angeles and London):** From Figures 4(b) and 4(d), one can see that the exponent α is 2.29 and 2.01 in Los Angeles and London, respectively. This reveals that Twitter users in London tend to contact a slightly higher number of friends within a given period, compared to users in Los Angeles. There

may be many explanations for this phenomenon, including that i) London is one of the world's most famous tourist destinations, which would attract relatively more visitors to use Twitter to send/receive direct mentions to/from their friends in the city and ii) London has a relatively higher population density than that of Los Angeles (refer to TABLE I for more details).

B. Number of Friends With Respect to Distance

The most interesting characteristic in friendship degrees is how friends of a user are distributed with respect to the geographic distance between the Twitter user and his/her friend. In this subsection, similarly as in [8], [9], we also verify whether Twitter users establish more relationships with friends who are living in geographic proximity to each other. As mentioned before, in our experiments, we use geo-tagged mentions to identify the location information of a user when he/she sent a mention to his/her friend. To detect his/her friend's location, we then observe *replied* geo-tagged mentions that were sent at the next closest time. Using these bidirectional mentions, we characterize the probability distribution $P_D(D = d)$ of the number of friends according to the distance d , where d [km] is the geographic distance between a user and his/her friend.

Unlike the earlier work in [8], the heterogeneous shape of $P_D(D = d)$ for the entire interval cannot be captured by a single commonly-used statistical function such as a homogeneous power-law using the approach of parametric fitting. Interestingly, as our main result, we observe that for the distance $d \in [d_{\min}, d_{\max}]$, $P_D(D = d)$ can be described as a *double power-law* distribution, which is given as:

$$P_D(D = d) \sim \begin{cases} d^{-\gamma_1} & \text{if } d_{\min} \leq d < d_s \text{ (intra-city regime)} \\ d^{-\gamma_2} & \text{if } d_s \leq d \leq d_{\max} \text{ (inter-city regime)}, \end{cases} \quad (2)$$

where γ_1 and γ_2 denote the exponents for each individual power-law and d_s is the *separation point*. This finding indicates that the friendship degree can be composed of two *separate regimes* characterized by two different power-laws, termed the *intra-city* and *inter-city* regimes. Figure 5 shows the log-log plot of the distribution $P_D(D = d)$ from empirical data, logarithmically binned data, and fitting function, where the fitting is applied to the binned data. As in Section IV-A, we also use the traditional least squares estimation to obtain the fitting function.¹⁴ In TABLE VI, the value of the exponents of $P_N(N = n)$, γ_1 and γ_2 , is summarized for each region.

Unlike the earlier studies in [8], [9] that do not capture the friendship patterns in the intra-city regime, our analysis exhibits two distinguishable features with respect to distance. More specifically, in each regime, the following interesting observations are made:

- In the intra-city regime, the distribution $P_D(D = d)$ decays slowly with distance d , which means that geographic proximity weakly affects the number of intra-city friends with which one user interacts. That is, in this regime, the geographic distance is less relevant for determining the number of friends. This finding reveals that more active Twitter users tend to preferentially interact over *short-distance* connections.
- In the inter-city regime, $P_D(D = d)$ depends strongly on the geographic distance, where there exists a sharp transition in the distribution $P_D(D = d)$ beyond the separation point d_s . Thus, *long-distance* communication is made occasionally.

¹⁴Using maximum likelihood estimation to fit a mixture function (e.g., a double power-law function) is not easy to implement and the performance of mixture functions has not been well understood.

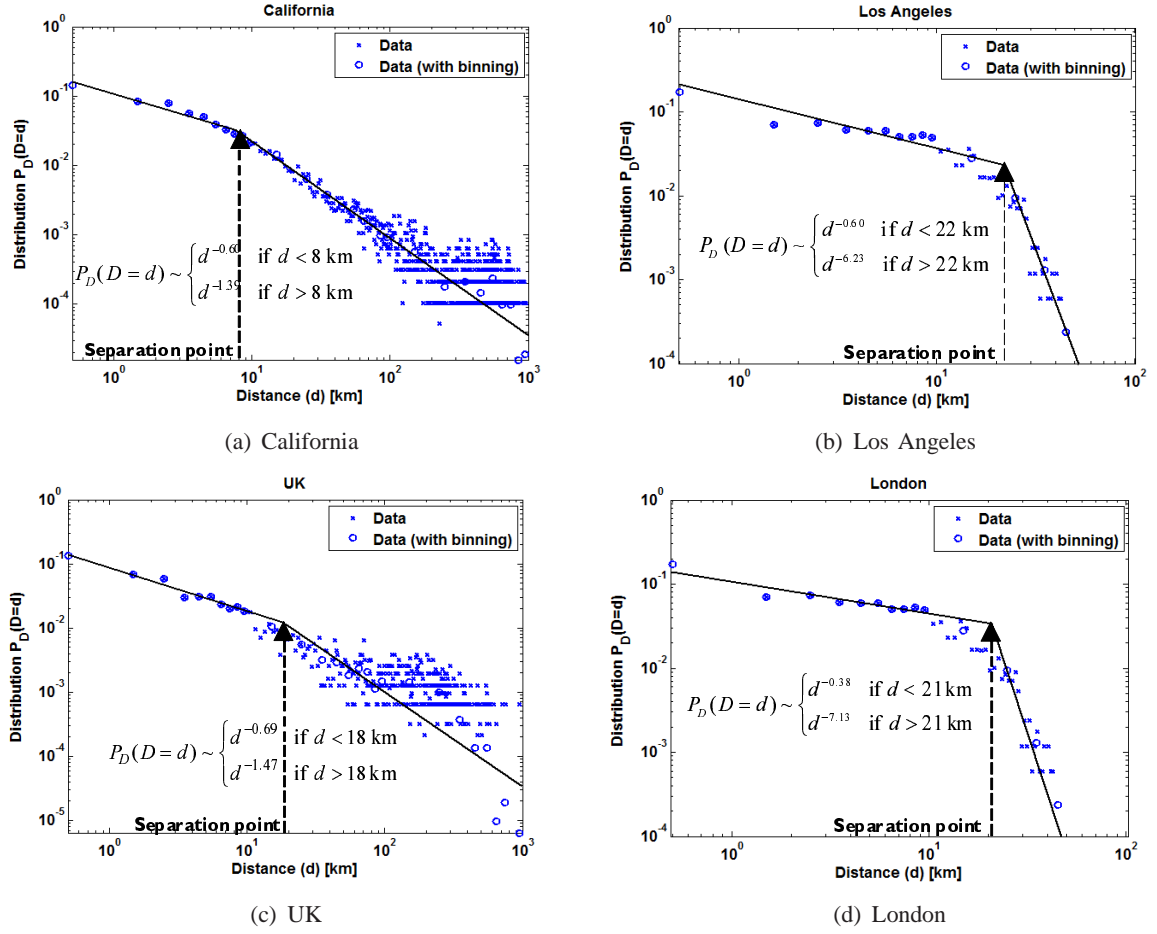


Fig. 5. Probability distribution $P_D(D = d)$ of the number of friends with respect to distance (log-log plot).

TABLE VI
THE VALUE OF γ_1 AND γ_2 FOR EACH REGION.

Region	γ_1	γ_2
California	0.60	1.39
Los Angeles	0.60	6.23
UK	0.69	1.47
London	0.38	7.13

The above argument stems from the fact that the separation point d_s is closely related to the length and width of the city in which a user resides. From these observations, we may conclude that within a given period, the individual is much more likely to contact online mostly friends who are in location-based communities that range from the local neighborhood, suburb, village, or town up to the city level. In addition, the following interesting comparisons are performed according to types of regions:

- **Comparison between the city-scale and state-scale/country-scale results:** We observe that the separation point d_s in populous metropolitan areas is much greater than that in larger regions that include local small towns (such as at the state or country level). For example, from Figures 5(a) and 5(b), we see that d_s is approximately 8 km and

22 km in California and Los Angeles, respectively. From Figures 5(c) and 5(d), the same trend is observed by comparing the results for the UK and London (18 km and 21 km, respectively). This finding reveals that Twitter users in populous metropolitan areas (e.g., Los Angeles and London) have a stronger tendency to contact friends on Twitter who are geographically away from their location (i.e., interacting over long-distance connections). This is because the average size (referred to as the land area) of the considered metropolitan cities is relatively bigger than that of cities in larger regions including small towns. Furthermore, it is seen that the exponent in the inter-city regimes (i.e., γ_2) in metropolitan areas is significantly higher than that in larger regions. Unlike the state-scale/country-scale results, this finding implies that the distribution $P_D(D = d)$ sharply drops off beyond d_s in huge metropolitan areas.

- **Comparison between the results in the two cities (Los Angeles and London):** From Figures 5(b) and 5(d), one can see that γ_1 is 0.60 and 0.38 and γ_2 is 6.23 and 7.13 in Los Angeles and London, respectively. Thus, in the intra-city regime, the geographic distance is less relevant in London for determining the number of friends. However, in the inter-city regime, the distribution $P_D(D = d)$ in London shows a bit steeper decline.

Our geo-tagged Twitter data provides position resolution at up to 10 meters, compared to the typical city-scale resolution in previous studies on friendship [8], [9], thus allowing much more fine-grained validation of these heterogeneous behaviors in terms of distance.

V. CONCLUDING REMARKS

The present work has developed a novel framework for analyzing the degree of bidirectional online friendship via Twitter, while not only utilizing geo-tagged mentions but also introducing a definition of bidirectional friendship. To provide analysis results, we first introduced two new algorithms, the first for counting friends and the second for a two-stage distance estimation algorithm. We verified that the homogeneous power-law model, also known as Zipf's law, holds on Twitter in terms of the number of friends of one user. More interestingly, we comprehensively demonstrated that the number of friends according to geographic distance follows a double power-law distribution, or equivalently, a double Pareto law distribution, where there exists a strict separation point in distance that distinguishes the intra-city regime from the inter-city regime. Our analysis sheds light on a new understanding of social interaction/relationships online with regard to small-scale space as well as large-scale space.

Characterization of the degree of friendship in space along with a greater variety of city/state/country-scale data on Twitter remains for future work. Suggestions for further research in this area also include analyzing a new friendship in the temporal domain (time) by utilizing geo-located Twitter.

REFERENCES

- [1] Wilson C, Boe B, Sala A, Puttaswamy KPN, and Zhao BY. User interactions in social networks and their implications. In: *Proceedings of the 4th ACM European Conference on Computer Systems (EuroSys'09)*, Nuremberg, Germany, March/April 2009, pp. 205–218.
- [2] Kwak H, Lee C, Park H, and Moon S. What is Twitter, a social network or a news media?. In: *Proceedings of the 19th International World Wide Web Conference (WWW2010)*, Raleigh, NC USA, April 2010, pp. 591–600.
- [3] Viswanath B, Mislove A, Cha M, and Gummadi KP. On the evolution of user interaction in Facebook. In: *Proceedings of the 2nd ACM Workshop on Online Social Networks (WOSN2009)*, Barcelona, Spain, August 2009, pp. 37–42.
- [4] Mislove A, Koppula HS, Gummadi KP, Druschel P, and Bhattacharjee B. Growth of the Flickr social network. In: *Proceedings of the 1st ACM Workshop on Online Social Networks (WOSN2008)*, Seattle, WA USA, August 2008, pp. 25–30.
- [5] Chen Y, Zhuang C, Cao Q, and Hui P. Understanding cross-site linking in online social networks. In: *Proceedings of the 8th ACM Workshop on Social Network Mining and Analysis (SNAKDD2014)*, New York City, NY USA, August 2014.

- [6] Watts DJ and Strogatz SH. Collective dynamics of ‘small-world’ networks. *Nature* 1998; 393:440–442.
- [7] Svenson P. Complex networks and social network analysis in information fusion. In: *Proceedings of the 9th International Conference on Information Fusion (Fusion2006)*, Florence, Italy, July 2006, pp. 1–7.
- [8] Liben-Nowell D, Novak J, Kumar R, Raghavan P, and Tomkins A. Geographic routing in social networks. *Proceedings of the National Academy of Sciences of the United States of America (PNAS)* 2005; 102: 11623–11628.
- [9] Kaltenbrunner A, Scellato S, Volkovich Y, Laniado D, Currie D, Jutemar EJ, and Mascolo C. Far from the eyes, close on the web: Impact of geographic distance on online social interactions. In: *Proceedings of the 5th ACM Workshop on Online Social Networks (WOSN’12)*, Helsinki, Finland, August 2012, pp. 19–24.
- [10] Jurdak R, Corke P, Cotillon A, Dharman D, Crossman C, and Salagnac G. Energy-efficient localization: GPS duty cycling with radio ranging. *ACM Transactions on Sensor Networks (TOSN)* 2013; 9: A:1–A:32.
- [11] Takhteyev Y, Gruzd A, and Wellman B. Geography of Twitter networks. *Social Networks* 2012; 34: 73–81.
- [12] Kulshrestha J, Kooti F, Nikraves A, and Gummadi KP. Geographic dissection of the Twitter network. In: *Proceedings of the 6th International AAAI Conference on Weblogs and Social Media (ICWSM-12)*, Dublin, Ireland, June 2012, pp. 202–209.
- [13] Frias-Martinez V, Soto V, Hohwald H, and Frias-Martinez E. Characterizing urban landscapes using geolocated tweets. In: *Proceedings of the 4th ASE/IEEE International Conference on Social Computing (SocialCom2012) and the 4th ASE/IEEE International Conference on Privacy, Security, Risk and Trust (PASSAT2012)*, Amsterdam, The Netherlands, September 2012, pp. 239–248.
- [14] Alowibdi JS, Ghani S, and Mokbel MF. VacationFinder: A tool for collecting, analyzing, and visualizing geotagged Twitter data to find top vacation spots. In: *Proceedings of the 6th ACM SIGSPATIAL International Workshop on Location-Based Social Networks (LBSN2014)*, Dallas, TX USA, November 2014.
- [15] Lee R and Sumiya K. Measuring geographical regularities of crowd behaviors for Twitter-based geo-social event detection. In: *Proceedings of the 2nd ACM SIGSPATIAL International Workshop on Location-Based Social Networks (LBSN2010)*, San Jose, CA USA, November 2010, pp. 1–10.
- [16] Hawelka B, Sitko I, Beinart E, Sobolevsky S, Kazakopoulos P, and Ratti C. Geo-located Twitter as proxy for global mobility patterns. *Cartography and Geographic Information Science (CaGIS)* 2014; 41: 260–271.
- [17] Jurdak R, Zhao K, Liu J, AbouJaoude M, Cameron M, and Newth D. Understanding human mobility from Twitter. Preprint, [Online]. Available: <http://arxiv.org/abs/1412.2154>.
- [18] Liu J, Zhao K, Khan S, Cameron M, and Jurdak R. Multi-scale population and mobility estimation with geo-tagged tweets. Preprint, [Online]. Available: <http://arxiv.org/abs/1412.0327>.
- [19] Falcone D, Mascolo C, Comito C, Talia D, and Crowcroft J. What is this place? Inferring place categories through user patterns identification in geo-tagged tweets. In: *Proceedings of the 6th International Conference on Mobile Computing, Applications and Services (MobiCASE2014)*, Austin, TX USA, November 2014.
- [20] Morstatter F, Pfeffer J, Liu H, and Carley KM. Is the sample good enough? Comparing data from Twitters’ Streaming API with Twitter’s Firehose. In: *Proceedings of the 7th International AAAI Conference on Weblogs and Social Media (ICWSM-13)*, Boston, MA USA, July 2013, pp. 400–408.
- [21] Gonzalez MC, Hidalgo CA, and Barabasi AL. Understanding individual human mobility patterns. *Nature* 2008; 453: 779–782.
- [22] Song C, Koren T, Wang P, and Barabasi AL. Modelling the scaling properties of human mobility. *Nature Physics* 2010; 6: 818–823.
- [23] Jiang S, Fiore GA, Yang Y, Ferreira, Jr. J, Frazzoli E, and Gonzalez MC. A review of urban computing for mobile phone traces: Current methods, challenges and opportunities. In: *Proceedings of the 2nd ACM SIGKDD International Workshop on Urban Computing (UrbComp2013)*, Chicago, IL USA, August 2013.
- [24] Wang D, Pedreschi D, Song C, Giannotti F, and Barabasi A-L. Human mobility, social ties, and link prediction. In: *Proceedings of the 17th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD2011)*, San Diego, CA USA, August 2011, pp. 1100–1108.
- [25] Rhee I, Shin M, Hong S, Lee K, and Chong S. In the Levy-walk nature of human mobility. *IEEE/ACM Transactions on Networking (TON)* 2011; 19: 630–643.
- [26] Chaintreau A, Hui P, Crowcroft J, Diot C, Gass R, and Scott J. Impact of human mobility on opportunistic forwarding algorithms. *IEEE Transactions on Mobile Computing (TMC)* 2007; 6: 606–620.
- [27] Hui P and Crowcroft J. Human mobility models and opportunistic communication system design. *Philosophical Transactions of The Royal Society A: Mathematical, Physical and Engineering Sciences* 2008; 366: 2005–2016.
- [28] Cattuto C, Van den Broeck W, Barrat A, Colizza V, Pinton J-F, and Vespignani A. Dynamics of person-to-person interactions from distributed RFID sensor networks. *PLOS ONE* 2010; 5: e11596.
- [29] Cho E, Myers SA, and Leskovec J. 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 (KDD2011)*, San Diego, CA USA, August 2011, pp. 1082–1090.
- [30] Backstrom L, Sun E, and Marlow C. Find me if you can: Improving geographical prediction with social and spatial proximity. In: *Proceedings of the 19th International World Wide Web Conference (WWW2010)*, Raleigh, NC USA, April 2010, pp. 61–70.
- [31] Giannakidou E, Koutsonikola V, and Vakali A. Co-clustering tags and social data sources. In: *Proceedings of the 9th International Conference on Web-Age Information Management (WAIM2008)*, Zhangjiajie, China, July 2008, pp. 317–324.

- [32] Nguyen TT and Jung JJ. Exploiting geotagged resources to spatial ranking by extending HITS algorithm. *Computer Science and Information Systems (ComSIS)* 2015; 12: 185–201.
- [33] Manning C and Schutze H. *Foundations of statistical natural language processing*. Cambridge, MA: MIT Press, 1999.
- [34] Newman MEJ. Power laws, Pareto distributions and Zipf's law. *Contemporary Physics* 2005; 46: 323–351.
- [35] Reed WJ. The Pareto law of income—an explanation and an extension. *Physica A* 2003; 319: 469–486.
- [36] Morstatter F, Pfeffer J, and Liu H. When is it biased? Assessing the representativeness of Twitter's Streaming API. In: *Proceedings of the 23rd International World Wide Web Conference (WWW2013)*, Seoul, Korea, April 2014, pp. 555–556.
- [37] Hodas NO, Kooti F, and Lerman K. Friendship paradox redux: Your friends are more interesting than you. In: *Proceedings of the 7th International AAAI Conference on Weblogs and Social Media (ICWSM-13)*, Boston, MA USA, July 2013, pp. 1–8.
- [38] Bastos MT, Travitzki R, and Puschmann C. What sticks with whom? Twitter follower-followee networks and news classification. In: *Proceedings of the 6th International AAAI Conference on Weblogs and Social Media (ICWSM-12) Workshop on the Potential of Social Media Tools and Data for Journalists in the News Media Industry*, Dublin, Ireland, June 2012, pp. 6–13.
- [39] Ugander J, Karrer B, Backstrom L, and Marlow C. The anatomy of the Facebook social graph. Preprint, [Online]. Available: <http://arxiv.org/abs/1111.4503>.
- [40] Huang Y, Shen C, and Contractor NS. Distance matters: Exploring proximity and homophily in virtual world networks. *Decision Support Systems* 2013; 55: 969–977.
- [41] Ennis A, Chen L, Nugent C, Ioannidis G, and Stan A. High level geospatial information discovery and fusion for geocoded multimedia. *International Journal of Pervasive Computing and Communications (IJPCC)* 2013; 9: 367–382.
- [42] Sinnott RW. Virtues of the haversine. *Sky and Telescope* 1984; 68: 158.
- [43] Milojevic S. Power-law distributions in information science: Making the case for logarithmic binning. *Journal of the American Society for Information Science and Technology (JASIST)* 2010; 61: 2417–2425.