International Journal of Geo-Information

MDPI

Article

Discover Patterns and Mobility of Twitter Users—A Study of Four US College Cities

Yue Li [1], Qinghua Li [2] and Jie Shan [2,*]

[1] Libraries, Purdue University, 504 W State Street, West Lafayette, IN 47907, USA; li1050@purdue.edu
[2] Lyles School of Civil Engineering, Purdue University, 550 Stadium Mall, West Lafayette, IN 47907, USA; li975@purdue.edu
* Correspondence: jshan@purdue.edu

Academic Editor: Wolfgang Kainz
Received: 21 November 2016; Accepted: 29 January 2017; Published: 10 February 2017

Abstract: Geo-tagged tweets provide useful implications for studies in human geography, urban science, location-based services, targeted advertising, and social network. This research aims to discover the patterns and mobility of Twitter users by analyzing the spatial and temporal dynamics in their tweets. Geo-tagged tweets are collected over a period of six months for four US Midwestern college cities: (1) West Lafayette, IN; (2) Bloomington, IN; (3) Ann Arbor, MI; (4) Columbus, OH. Various analytical and statistical methods are used to reveal the spatial and temporal patterns of tweets, and the tweeting behaviors of Twitter users. It is discovered that Twitter users are most active between 9:00 p.m. and 11:00 .pm. In smaller cities, tweets aggregate at campuses and apartment complexes, while tweets in residential areas of bigger cities make up the majority of tweets. We also found that most Twitter users have two to four places of frequent visits. The mean mobility range of frequent Twitter users is linearly correlated to the size of the city, specifically, about 40% of the city radius. The research therefore confirms the feasibility and promising future for using geo-tagged microblogging services such as Twitter to understand human behavior patterns and carry out other geo-social related studies.

Keywords: spatial patterns; temporal patterns; human mobility; human dynamics; Twitter; social media

1. Introduction

Twitter is the most popular micro-blogging service in the world. According to Milstein et al. [1], millions of people use this online social network to connect socially with friends, family members and co-workers. They use it to inform others what they are doing, thinking or what is happening. A status update message is called a "tweet" and one tweet is limited to 140 characters. All (Twitter) users can follow other users and read the tweets they posted. Users being followed by others do not need to follow them back. The number of Twitter users has increased rapidly since Twitter's launch in 2006. As of March 2016, Twitter [2] has more than 310 million monthly active users. Stone [3] stated that 9.1% of the U.S. population has become the pulse of a planet-wide news organism, hosting the dialogue about everything from the Arab Spring to celebrity deaths. Similarly, Lunden [4] and Leetaru et al. [5] reported that in a period of seven years over 170 billion tweets were sent, totaling 133 terabytes, with more than 500 million tweets posted per day. Twitter offers an unprecedented opportunity to study human communication and social networks [6], and has caught the attention of social scientists. The number of published papers has tripled from 27 to 84 between 2009 and 2012 [7]. Furthermore, Twitter provides real-time programmatic access to a massive seven-year archive via APIs, Its ease and availability of use have turned Twitter into one of the favorite data sources of social scientists [5]. According to [7], Twitter data have been used in (1) event detection, including disaster

management, disease management [8] and traffic management; (2) location inference [9]; (3) social network analysis, user characteristics and their social relationship research [10].

One important feature of Twitter is its availability on smartphones, which may have embedded location sensors such as GPS, allowing users to send messages with their geographic coordinates [11]. In addition, since August 2009, Twitter has permitted users to indicate their city or neighborhood location manually [2]. On average, 2% of all tweets include location information [5], which translates to around 10 million tweets per day. Therefore, Twitter is becoming a key source of open and free geospatial data generated by citizens [12]. For example, geo-tagged tweets have been primarily utilized in disaster management [8,13,14]. With accurate location information, tweets are proven to be highly spatiotemporally reliable and useful in such applications [7]. The immense volume and diversified information available in tweets have made them a promising supplementary or alternative to traditional survey data, opening new avenues for discovering geo-social knowledge and in the meantime challenging for novel research approaches.

With their immediacy, global coverage, and people's volunteerism, geo-tagged tweets and other forms of social media data have been used in a variety of applications, such as emergency or crisis management [15,16], event detection [17–19], knowledge discovery combined with topic modeling and semantics analysis [20–22], location prediction [9] and urban network model improvement [23]. Data shared on the Location Based Social Networks of Microsoft Research have made a comprehensive collection of information on human behaviors in space and time available for investigation [24]. Particularly, the spatial-temporal pattern is of great importance for human behavior study, which draws insights on human mobility [24–26]. The potential of geo-tagged tweets and other social media geographic data has been proven in geo-social research, which ultimately aims to enable new intelligent geo-social systems. These aforementioned studies are believed to have major breakthrough in the coming years with the growing interests in spatial computing [27].

On the other hand, studies on modeling human mobility have received attention from various fields including traffic forecasting [28], urban planning [29], recommender systems [30,31], and disease spread [32,33]. Moreover, previous studies at the individual level have been performed on GPS datasets, such as cellphone data [34,35], card transactions [36], and taxi trajectories [37]. More recently, researchers have used data derived from location-based social network such as georeferenced tweets to analyze mobility pattern. It has proven to be successful in using Twitter data in activity pattern analysis [38]. The results complied with previous findings from other data sources, such as mobile phone data [10], or corresponded with existing survey results such as the American Community Survey data [39]. Although human behavior studies used tweets and other forms of social media data, they focused on a certain group of people, such as tourists using photo-sharing services [40–42], on the general public either at a regional scale [11], a country scale [43] or a global scale [5]. Limited work has been done to explore and model human mobility patterns at a city or town scale [44].

The research reported in this paper aims to fill this gap. In addition, due to the great volume and public accessibility of tweets, the focus of this research is to utilize geo-tagged tweets rather than traditional GPS datasets to better depict human mobility patterns. In consideration of the characteristics of Twitter data and its potential in geo-social knowledge discovery, the objectives of this research are as follows. We expect to explore the spatial and temporal patterns of geo-tagged tweets by various geospatial mining methods. In addition, we will infer and understand the tweeting behaviors and mobility patterns of the tweet users. Finally, this study intends to showcase a framework for geo-social media data mining and knowledge discovery, especially in the context of human behavior and their interaction with city settings. It is expected to benefit a wide variety of applications and inspire sociologists, anthropologists, policy makers, and geographers.

The rest of the paper is structured as follows. Section 2 describes the geo-demographic and Twitter data of the four cities used in this study. Spatial and temporal patterns for Twitter users are explored and depicted in Section 3 at full scales: from city to buildings, and from days to hours. In addition to Twitter data, land use data also participate in the analysis so that urban settings can be attributed to

tweet locations. Section 4 addresses certain behaviors of individuals by looking into the distribution of their tweet counts, frequencies and mobility ranges, to the latter of which an experimental model is established in terms of the city size. The findings and significance of the work are summarized in Section 5.

2. Study Areas and Data

2.1. Study Areas

The work is carried out by using geo-tagged tweets over four Midwestern college cities/towns in the U.S., i.e., West Lafayette, IN (Purdue University), Bloomington, IN (Indiana University), Ann Arbor, MI (University of Michigan), and Columbus, OH (The Ohio State University). Table 1 summarizes the population and size of the four cities, while Figure 1 presents their maps. Geographic and demographic facts about these four cities are described below, mostly based on public information, e.g., Wikipedia.

Table 1. Geo-demographic data of the four study cities.

City/Area	Population	Size (km^2)	Lower-Left (Deg.)	Upper-Right (Deg.)
West Lafayette, IN	29,596	19.76	(−86.970, 40.414)	(−86.896, 40.475)
Bloomington, IN	80,405	60.50	(−86.623, 39.102)	(−86.473, 39.196)
Ann Arbor, MI	113,934	74.33	(−83.804, 42.221)	(−83.674, 42.322)
Columbus, OH	822,553	577.85	(−83.195, 39.843)	(−82.773, 40.204)

West Lafayette (Figure 1) is the most densely populated city in Indiana as well as the most culturally diverse city in the Midwest. The median age is 22.8, and 49.4% of the population are between the ages of 18 and 24. The population density is 1499.6/km^2 [45]. The city lies in the center of Tippecanoe County, and overlooks the Wabash River (Figure 1). Purdue University is located in West Lafayette, and has 39,256 students, 30,147 of which were undergraduate students in the fall semester of 2012 [46]. The university has 15 residence halls, in which approximately one-third of the single undergraduate students live [45].

Bloomington (Figure 1) is the county seat of Monroe County in southern Indiana. It is the sixth largest city in Indiana with a population density of around 1340.4/km^2. The median age in the city is 23.3, and 44.5% are between the ages of 18 and 24 [47]. Indiana University Bloomington is located in Bloomington and had 32,532 undergraduates out of a total student body of 42,731 [47]. The campus has 12 residence centers clustered into three neighborhoods [48].

Ann Arbor (Figure 1) is the sixth largest city in Michigan with a population density of 1580.7/km^2. The median age of the population in the city is 28, of which 26.8% are between the ages of 18 and 24, and 31.2% are between 25 and 44 [49]. The city is the home of University of Michigan, which shapes the city, lending a college-town character [49]. The university had 43,246 students as of the fall of 2012, among which 27,979 were undergraduate students. It has four main campuses (North, Central, Medical, and South). The on-campus housing is located on the Central Campus, the Hill Area and the North Campus; and nearly 40% of the undergraduate students live on campus [50]. Besides the large student population, the university also has about 30,000 employees, including about 12,000 in the medical center [49]. Besides University of Michigan, Ann Arbor is also home to Concordia University Ann Arbor, a campus of the University of Phoenix, and Cleary University [49].

Columbus (Figure 1) is the capital of the state of Ohio and its largest city. It is the 15th largest city in the U.S. and the most populous city in Ohio with a population density of 1399.2/km^2. The median age of the population from the 2010 census was 31.2, of which 14% were between the ages of 18 and 24, and 32.3% were between 25 and 44. The city has a diversified economy, including education, insurance, banking, government, energy, health care, retail, technology, food, clothing, logistics, and health care. Five U.S. Fortune 500 corporation headquarters are located in Columbus as well. The Ohio State University, Columbus State Community College, and many private institutions are located in

Columbus [51]. The Ohio State University has 56,867 students in total, of which 42,916 are undergraduate students. There are 31 on-campus residence halls, located on the South, North, and West Campuses [52].

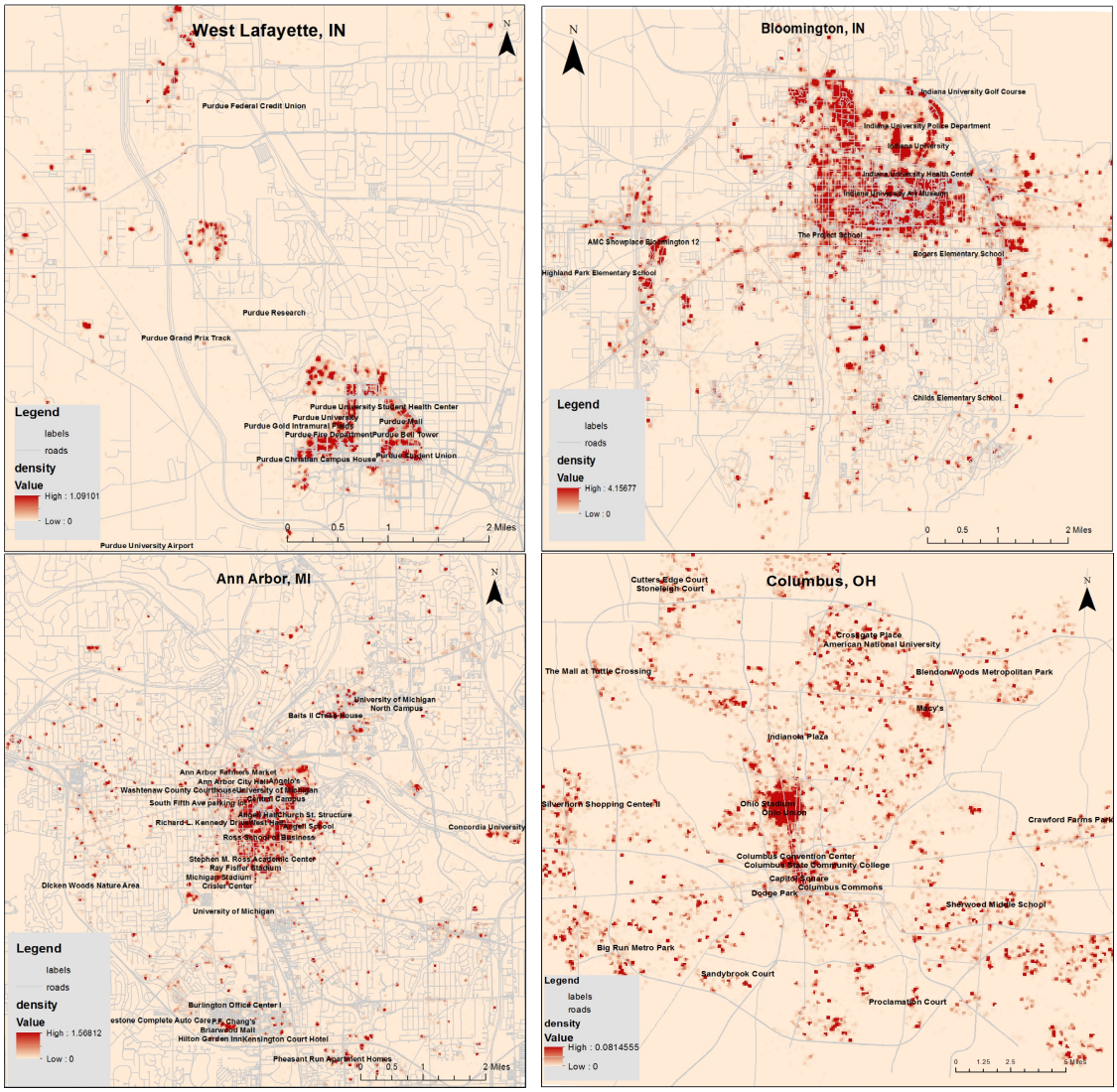

Figure 1. Maps of the fours study areas and their tweet densities (tweet counts/square meters).

All of the four cities are located in the Midwest of the U.S. Economically, the region is balanced between heavy industry and agriculture and have high employment-to-population ratio. All of them have major public universities with tens of thousands of students. Although they share certain common attributes, their size or scale varies. West Lafayette, Bloomington, Ann Arbor are mostly college towns, whereas Columbus not only has multiple colleges but large scale industry, making it a metropolitan area. Such selection allows our study be focused on geographic regions with common attributes but at different scales. In addition, they share the same time zone, which makes the analysis and comparison meaningful. Moreover, one of the authors has visited the four cities multiple times and has first-hand experience on the structure of the cities. It is expected that such in situ knowledge, though limited, would be helpful when interpreting the data and its patterns.

2.2. Twitter Data

The Twitter data used in this analysis were downloaded using the Twitter Streaming Application Programming Interface (API), which provides developers low latency access to the global stream of

Tweet data. There are three main streaming endpoints: (1) the public streams, by which the streams of public data flowing through Twitter can be pushed; (2) the user streams, by which a single-user's stream containing almost all of the data corresponding to the user's view can be accessed; (3) the site stream, which is a multi-user version of the user streams [53]. Because this work aims to reveal and understand the patterns of geo-tagged tweets in the four study areas, only the tweets within their boundaries are needed. The public stream method was used with two Python libraries, Tweepy [54] and Twitter-Streamer [55]. The search terms used were the coordinate boundaries of the study areas defined in Table 1, which contains the city municipal boundaries. The only tweets included were those attached with longitude and latitude, which are usually generated from smartphones by users who explicitly opt to publish their present locations. It should be noted that Twitter restricts the public access by allowing only up to 180 tweets per 15 min [56]. As a result, what used in this study were not the complete tweets generated by Twitter users but their random samples. Furthermore, positioning accuracy of smartphones is reported to be 2–3 m under good multipath conditions and can degrade to 10 m or worse under adverse multipath conditions [57]. Such quality however is good enough for us since this study is mostly related to patterns of tweeting activities. Of note, among the four study areas it is found that around 70%–80% of the tweets were sent from iPhone OS platforms, and 10%–20% were from Android platforms; other platforms were less than 10%.

Our tweet collection was from. 18 November 2013 to 1 June 2014 and is summarized in Table 2. As shown in its first two rows, about 3.4 million tweets were downloaded, with about 71 k from West Lafayette, 348 k from Bloomington, 295 k from Ann Arbor, and more than 2.6 million from Columbus. Columbus had the most Twitter users, more than 52,000, yielding the highest (more than 50) average number of tweets per user. Ann Arbor had the lowest, less than 20 tweets per user over a period of more than six months.

Table 2. Number of tweets, users, and frequent users with more than 100 tweets.

Study Area	West Lafayette	Bloomington	Ann Arbor	Columbus
# Tweets	71,658	348,478	295,057	2,671,648
# Users	2884	8336	15,394	52,149
Avg. tweets per user	24.85	41.80	19.17	51.23
# Users with 100+ tweets	153	725	571	2661
% Frequent users	5.3%	8.6%	3.7%	5.1%
# Tweets from frequent users	41,402	248,549	168,138	1,071,941
% Tweets from frequent users	57.7%	71.3%	56.9%	40.1%

2.3. Land Use Data

Local land use data were included to assist interpreting the spatial and temporal patterns of the tweets and to establish an understanding on people's life style. To compare the patterns between different study areas, the land use types in each city were grouped into categories that are more general and as common as possible among all four cities. This is a necessary step since the zoning maps of different cities follow different semantics and ontology. Regrouping the classes makes possible a comparative evaluation across the cities. For West Lafayette, the land use data were digitized based on the zoning map provided by the Tippecanoe County GIS website. The original zoning classes were re-clustered into five groups: institutional, residential, business, development, and others. The Bloomington land use data were downloaded from the City of Bloomington GIS website; and the land use classes were regrouped into five groups: institutional, residential, commercial, planned unit development (PUD), and others. Ann Arbor's land use information was retrieved from the city's website; the classes were reclassified into five groups: institutional, residential, commercial, transportation, and others. The Columbus land use map was acquired from the Columbus city GIS office; and the zoning classes were categorized into five groups: institutional, residential, commercial, downtown district, and manufacturing.

3. Spatial and Temporal Patterns

3.1. Spatial Patterns

Knowing the locations where people usually tweet can be important for a variety of applications. However, due to the point aggregations resulting from the large volume of data, simply displaying all the tweets on a map would not be useful for revealing the patterns of interest in this study. Therefore, the spatial density of tweets, i.e., the number of tweets per square meter in each study area was created and plotted in the maps of Figure 1. When calculating the density, the radius was chosen as 0.25% of the diagonal length of the study area. The cell size was about the same as the radius, which is 20, 30, 25 and 100 m respectively for the four cities.

As shown in Figure 1, in all these four cities the biggest tweet clusters emerge on the university campuses and their surroundings. A few closer views of the hot spots in the four cities are highlighted in Figure 2. They were at apartment complexes, downtown districts and shopping malls (Figure 2). In addition, compared to West Lafayette, Bloomington, and Ann Arbor, where most of the tweet clusters appeared around the campuses (Figure 2), the locations of the clusters in Columbus were scattered all over the city and were more evenly distributed.

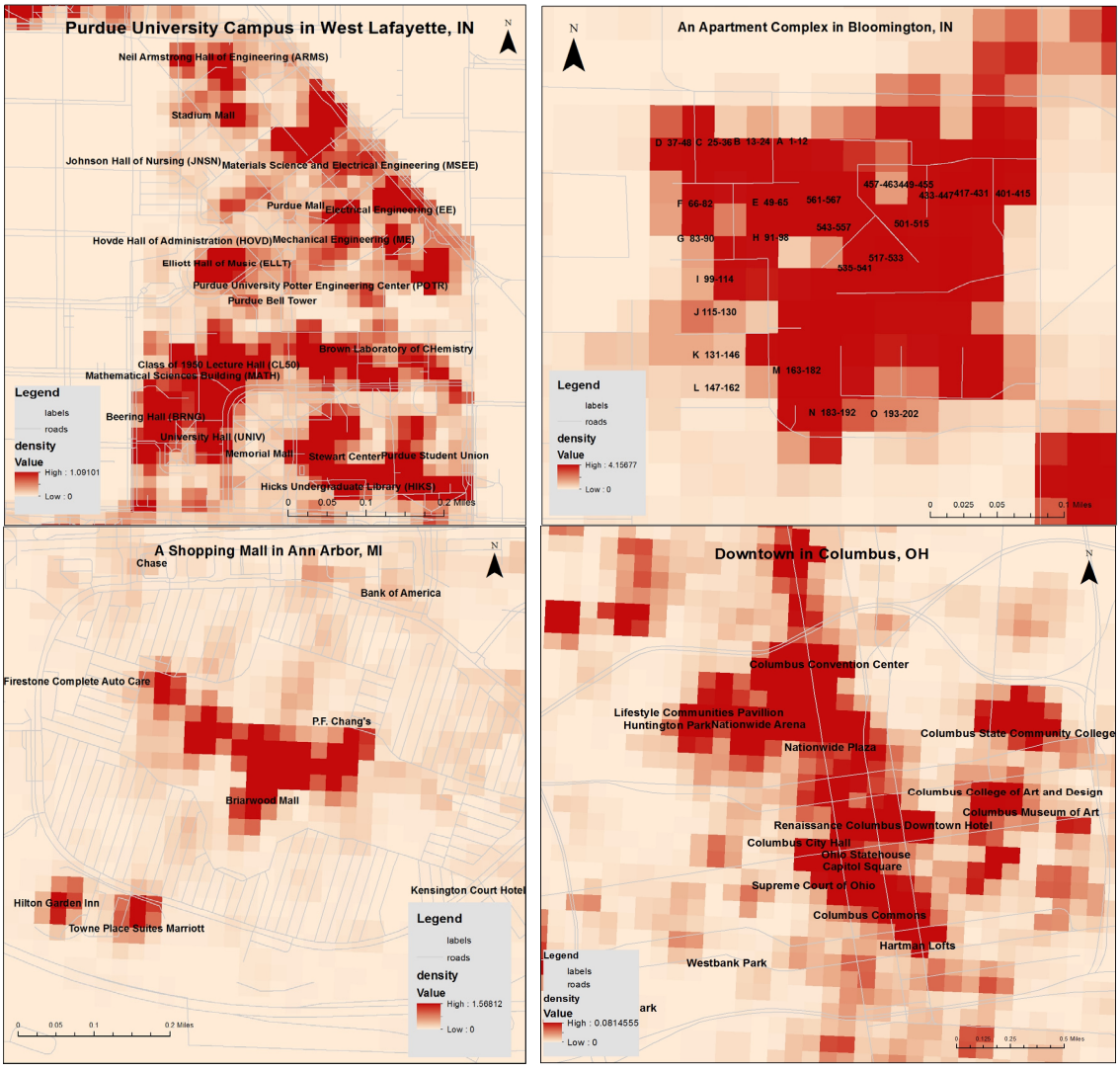

Figure 2. Closer views of selected dense tweet locations in the four cities.

3.2. Temporal Patterns

Tweets in all four study areas appeared to have similar hourly patterns (Figure 3). The number of tweets, as well as the number of users increased around 6:00 a.m. when people were awakening and getting ready for school or work. The tweets continued to grow in all four cities until 12:00 p.m. For West Lafayette, the increase continued until 1:00 p.m. when it hit at a peak and then began to decline until 4:00 p.m. In the meantime, the number of tweets in the other three cities remained quite stable. After 4:00 p.m., the tweets began to rise again until around 9:00 p.m. where they reached a peak. This evening period was likely when people returned from work or study, taking care of the household or relaxing. For West Lafayette, Bloomington, and Ann Arbor, the total tweets around 9:00 p.m., the peak time, comprised about 6% of all the tweets. However, for Columbus, the tweets at the peak time were almost 9% of the total tweets, indicating that they may have had more variations in their daily routines compared to the others. It is also noticed that compared to Columbus the number of users in West Lafayette started to decline at night. As contrary, the number of users in Bloomington remained still, implying that people in Columbus (bigger city) were more active at night than those in other cities, which was possibly due to the size of Columbus and the variety of activities available there. After 9:00 p.m., the tweet counts declined again until 12:00 a.m. when most people were probably getting ready to go to sleep. The number of tweets continued to decrease until around 4:00~5:00 a.m., where it reached a valley. From above statistics, it is concluded that there were two peak tweeting time during the days in these four cities, one at noon and the other around 9:00 p.m. at night. The number of tweets during daytime is at the lowest around 5:00 a.m. and 5:00 p.m. Finally, it should be noted that the number of tweets and the number of Twitter users follow a very similar hourly pattern during the day.

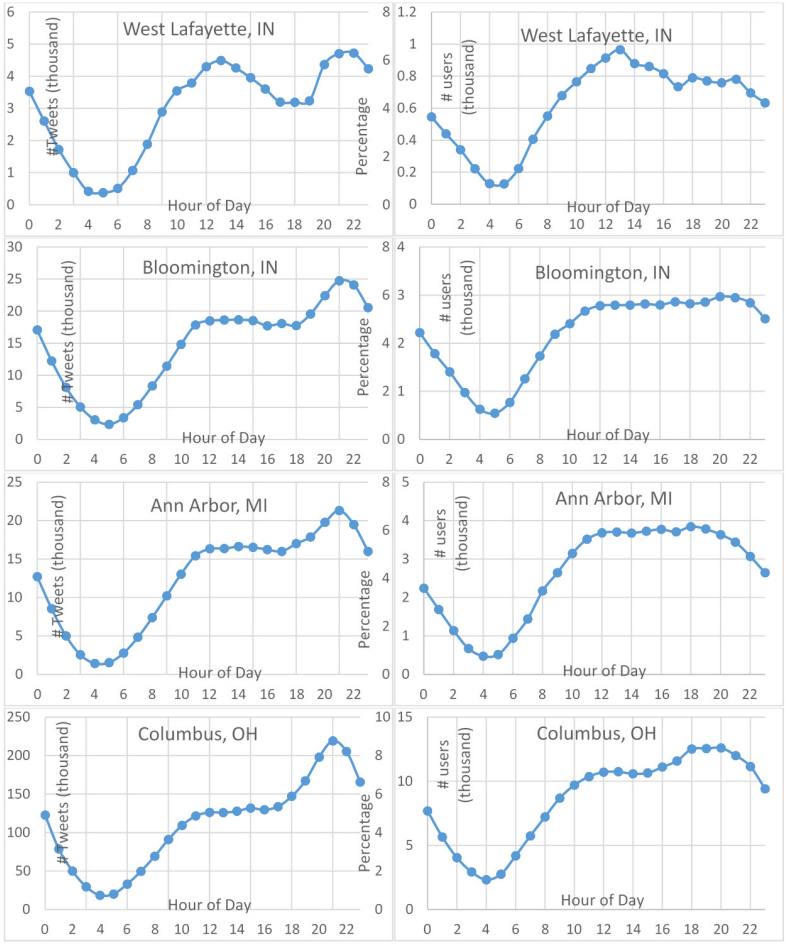

Figure 3. Distribution of number of hourly tweets (**left**) and users (**right**) for the four cities.

There were more users during weekends than weekdays. The average number of tweets per weekday was less than the average for weekend. In addition, the 'weekday' user group and the 'weekend' user group had only about 15%–20% overlap (Table 3), implying that most people only tweeted either on weekdays or on weekends, but not on both. The reason behind this might be the tweeting preference of users, or users leaving or coming to town on weekends, an indication of people's mobility and life styles.

Table 3. Number of distinct users on weekdays and weekends.

# Users	West Lafayette	Bloomington	Ann Arbor	Columbus
# users on both weekdays and weekends (d)	391	1456	1613	7401
# users on weekdays (A)	2137	6437	11,302	41,276
d/A (%)	18.3	22.6	14.3	17.9
# users on weekends (E)	1841	5766	9219	35,184
d/E (%)	21.2	25.2	17.5	21.0

3.3. Spatial-Temporal Patterns

Table 4 summarizes the tweets in terms of land uses. Tweets in institutional areas made up the majority of tweets in West Lafayette and Bloomington, while in Ann Arbor and Columbus tweets in residential areas accounted for the most among various land uses (Table 4). Less than 20% of tweets in West Lafayette were from residential areas, whereas more than 72% tweets were from school or office, a fact clearly showing the typical characteristics of a "college town", where most populations are related to Purdue in one way or the other. Over 10% of the tweets were from commercial land uses in Bloomington and Columbus, where there are no clear city-campus boundaries, suggesting a mixed city structure for these two city areas. On the other hand, West Lafayette and Ann Arbor had much smaller percentage of tweets (1.4%–6.3%) from commercial land uses, suggesting more business and commercial activities be desired to boom the local economy in these two areas.

Table 4. Percentage of Tweets in various land uses.

% in Total	West Lafayette	Bloomington	Ann Arbor	Columbus
Institutional	72.60	45.61	17.67	10.39
Residential	18.52	29.39	44.75	68.48
Commercial	1.40	15.64	6.30	11.78

Tweets in land uses vary with time during the day, as depicted in Figure 4. In West Lafayette, most of the tweets were posted from institutional areas, which implied that most of the Twitter users were likely college students. Slightly different from the overall temporal pattern of all the tweets in the city, the peak for institutional areas was around 12:00 p.m.–1:00 p.m. After that, the tweet count began to decrease until around 7:00 p.m., when it rose to a peak at 10:00 p.m. The land use with the second most tweets was residential areas, where the number of tweets drastically increased at 7:00 p.m. until 10:00 p.m., which corresponds to the period of time when people leave from work or school and return home. Very few tweets were found in other land use types such as industrial and business (Figure 4).

Similar to West Lafayette, the land use type with the most tweets was institutional areas in Bloomington. The temporal pattern was also nearly identical with a peak at 12:00 p.m. followed by a decrease until 6:00 p.m. and then an increase until a peak at ~9:00 p.m.–10:00 p.m., inferring that many users are college students. In addition, the land use with the second most tweets, similar to West Lafayette, was residential areas. Tweets in residential areas also began to rise from 6:00 p.m. to 9:00 p.m. However, different from West Lafayette, where very few tweets posted from other land use areas, commercial area of Bloomington had as up to 1% of the total, indicating users' were active in these areas.

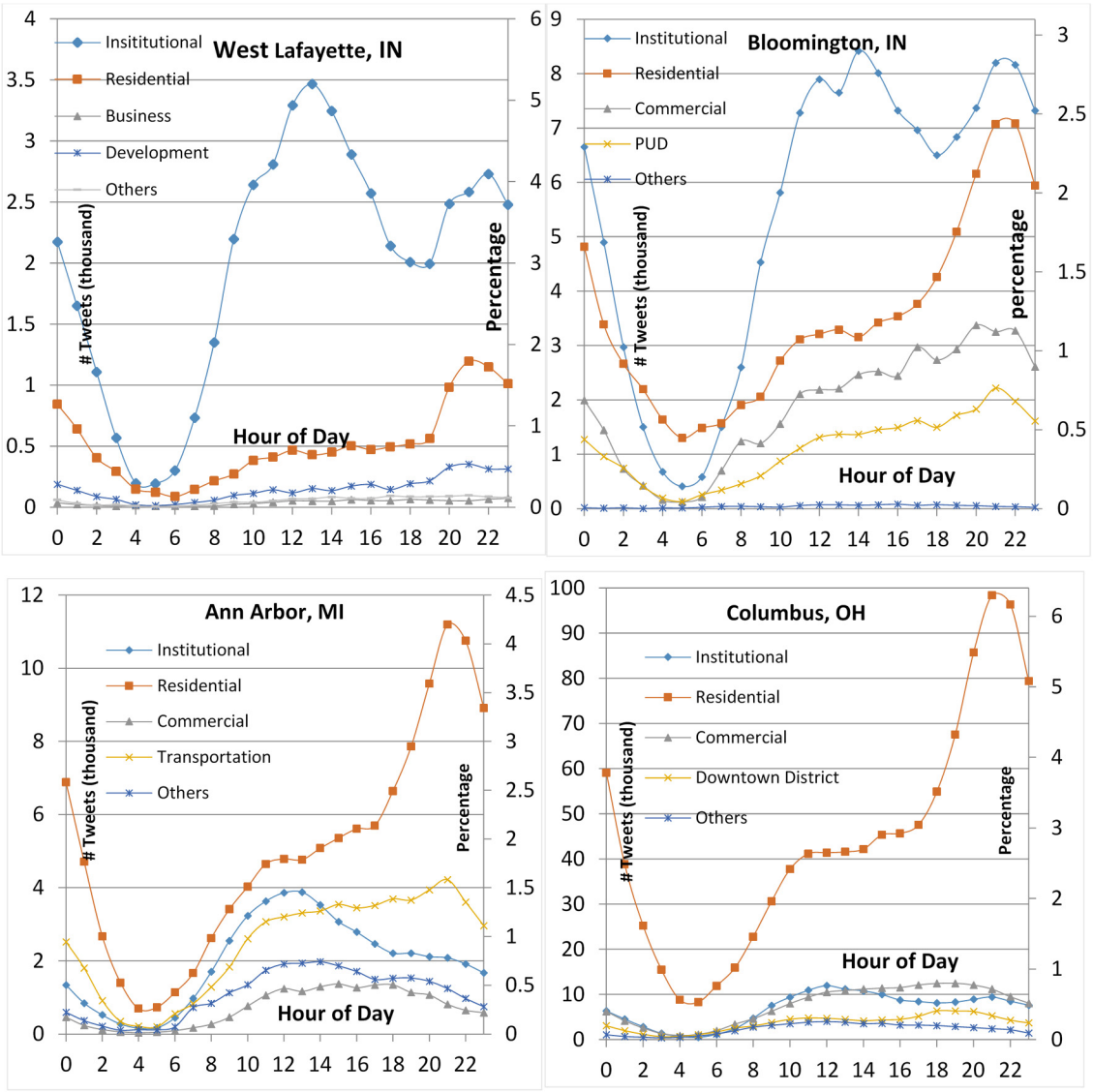

Figure 4. Distribution of number (in count and percentage) of hourly tweets in terms of land use.

Differing from West Lafayette and Bloomington, the land use type with the most tweets in Ann Arbor was residential areas, where the tweet counts began to increase from 6:00 a.m. until 10:00 a.m., remained stable until 6:00 p.m., and then continued to increase until 9:00 p.m. The land use types with the second most tweets were institutional areas and transportation. In institutional areas, the number of tweets began to decrease at 2:00 p.m. and did not rise again until evening, which is different from West Lafayette and Bloomington. This implies that fewer students were on campus in Ann Arbor than in West Lafayette and Bloomington. The land use types in Ann Arbor included transportation, which mainly consisted of roads and highways, and it was surprising to discover that a large number of users were tweeting on or close to the roads. When there was a decrease in the tweet counts in the institution areas and an increase in the transportation and residential areas around 4:00 p.m. to 5:00 p.m., a population flow from the institution areas to the transportation and residential areas was inferred. Finally, knowing that tweet counts in commercial and recreation areas comprised 0.2%–0.5% of the total tweets and that relatively more tweets took place in the daytime, it was concluded that the users were usually active during the daytime in those areas.

In Columbus, a vast majority of tweets occurred in residential areas followed by commercial and institutional areas. The tweets in institutional areas had patterns similar to West Lafayette and Bloomington, with a peak around 12:00 p.m. and a small rise around 8:00 p.m. to 9:00 p.m. In addition, the tweet counts in commercial areas, with a peak reaching almost 0.8% of all the tweets, were nearly as many as those in the institutional areas. Since the downtown area, which has several commercial businesses, malls, and restaurants, belongs to a separate land use type, i.e., the Downtown District, the tweet counts from the commercial area should be larger than shown here. This percentage was the highest among all the four cities. It can be concluded that many Twitter users posted tweets from their homes and were more active in commercial areas than those in other cities, indicating that Twitter potentially can be utilized for business applications such as market analysis and advertising.

4. User Tweeting Behavior and Mobility

4.1. User Tweet Counts

This section examines the properties of how many tweets a user posted. Figure 5 shows the number of tweets (in logarithm) of all tweet users in a descending order. A long tail phenomenon is revealed for all four cities in the distribution of the number of tweets versus the number of users. It shows that a very large number of users actually contributed only a very small percentage of tweets (the long tail), whereas a relatively small number of users actually posted most tweets (the short head part). To be specific, as shown in Table 2, 5.3%, 8.6%, 3.7% and 5.1% of the total users that posted more than 100 tweets could even generate 57.7%, 71.3%, 56.9% and 40.1% of the total tweets, respectively for the four study areas. As these small percentage users posted significantly more tweets than other users, they are regarded as "significant or frequent users". Therefore, it is reasonable to believe a great deal of information can be potentially discovered by analyzing the tweets of these frequent users. In addition, due to the large number of tweets posted from these users, determining their mobility patterns and the frequent places they visited might become possible. On the other hand, the non-significant users, often more than 95% of the total users, residing in the tail part of the curves, contribute as a group only slightly more than 30% of the total tweets. Inferring reliable patterns about this group of non-frequent users would be a challenge, however, the finding could be more informative and valuable.

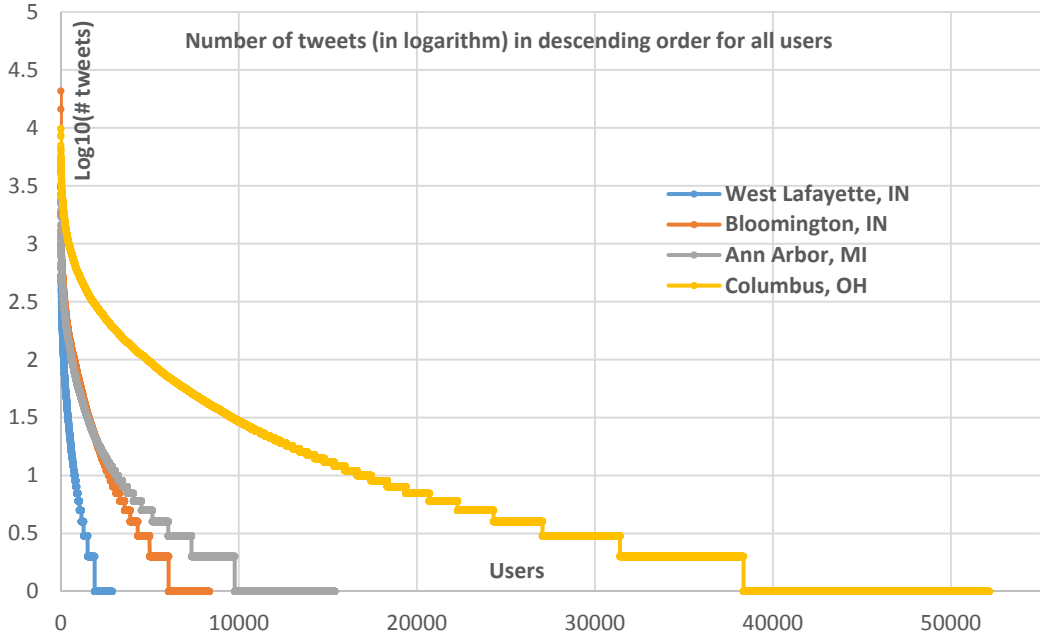

Figure 5. Number of tweets (in logarithm) in descending order for all users.

4.2. User Tweeting Frequency

This section examines the properties of how often a user tweeted. To this end, we calculated the time interval, i.e., the frequency between two successive tweets for each individual user. The corresponding number of users is then summarized against their average tweeting frequency (in logarithm hours) in Figure 6. The distributions of the user average tweeting frequency are all skewed to the right with a long tail (considering the scale effect of logarithm when the frequency is more than 1 h) and follow a similar pattern for all four cities. Again, only a small percentage of the users tweeted more often at a frequency of a few hours (0.2 in x-axis corresponding to 1.6 h), whereas majority of the users tweeted at a frequency as long as tens of hours (2 in x-axis corresponding to 100 h). The peak frequency at which the highest percentage of users (about 4%–5%) tweeted is less than 100 h, i.e., 4 days. Figure 6 also illustrates that only a small percentage of users tweeted at a frequency as low as more than 1000 h, i.e., more than one month. It is interesting to note that no significant difference in tweeting frequency is observed among the Twitter users across four cities; the mode frequency in all cities is about 2–4 days.

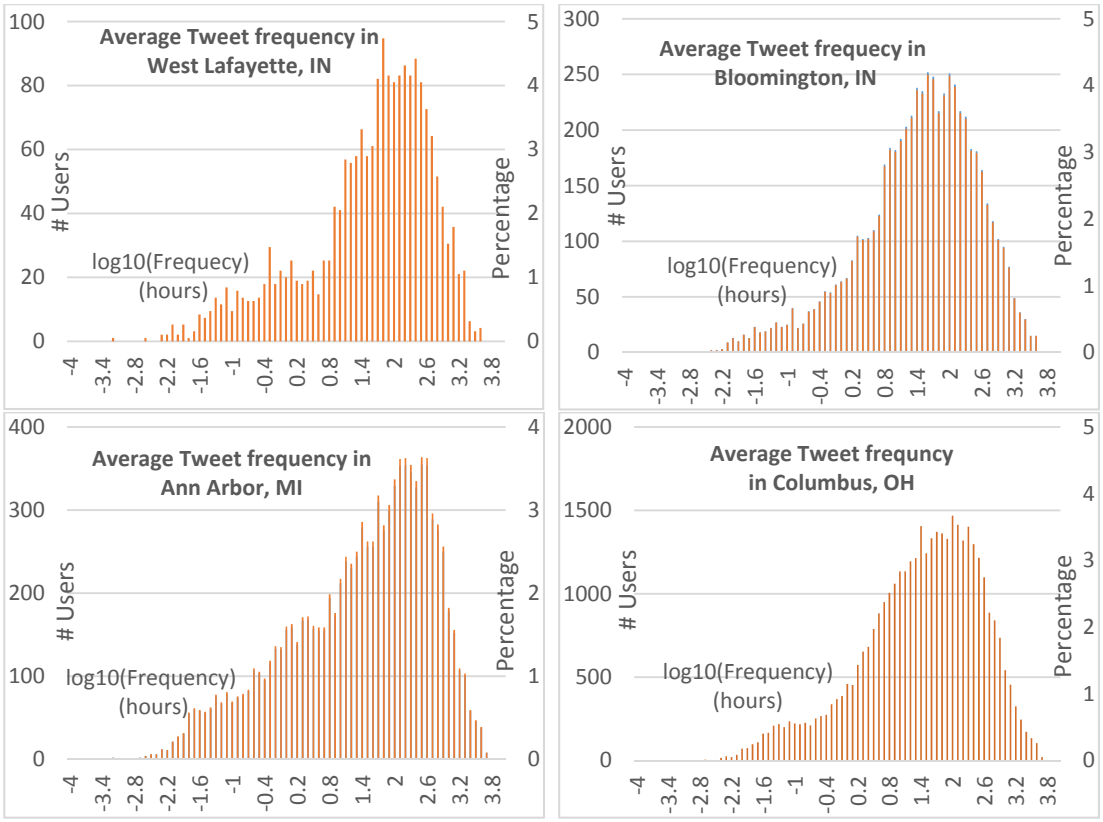

Figure 6. Number (in count and percentage) of users versus tweet frequency (in logarithm hours)

4.3. User Mobility

The Expectation-Maximization (EM) algorithm [58] was used for spatial clustering the tweets of individual users. For each individual Twitter user, the EM algorithm was applied to all locations of this user's tweets. The number of clusters for a user was defined as no more than 5 in this analysis. EM is a two-step iterative procedure to find maximum likelihood solutions [59]. For each iteration, the first step, i.e., the E-step (E-xpectation), assesses the probability of every point belonging to each cluster. Then, the second step, the M-step (M-aximiation), estimates the parameters for the probability distributions of the clusters. The algorithm is run until the distribution parameters converged or reached the maximum number of iterations [58]. EM algorithm is frequently used for data clustering

and is proved to be simple, stable and robust to noise [60]. Specifically, it has been used for spatial clustering [61]. In this subsection, only users with more than 100 tweets were used for mobility analysis, which counted to 4%–8% of the total users and made 40%–70% of all the tweets (Table 2).

Shown in Table 5 are the percentage users who had up to five spatial clusters. When a cluster had very few tweets, it was not considered as frequently visited places. Tweet clusters with less than 5% of the individual's total tweets were also excluded. Most of the users had two, three or four tweet clusters, while very few had one or five clusters. It is seen that most Tweet users had three clusters, while they visited and tweeted most often. These clusters were very likely users' homes, workplaces or places they have routinely visited.

Table 5. Percentage users of different spatial clusters and their distances.

City	# Centers					Avg. Distance between Centers	
	1	2	3	4	5	Mean (km)	Median (km)
West Lafayette	5.9	26.8	39.9	25.5	2.0	1.342	0.795
Bloomington	2.1	20.1	38.8	31.6	7.4	1.651	1.260
Ann Arbor	4.4	21.2	42.6	26.6	5.3	1.892	1.556
Columbus	12.2	31.9	33.4	18.1	4.5	5.763	4.879

The average distance between the cluster centers shown in Table 5 could approximate the user mobility range. For each Twitter user, the average of all the distances between any two centers was calculated as the user's average mobility range. The average mobility range for the users varied with the city in which they resided. This analysis determined that the larger the city is, the longer the mobile distance is. As shown in Table 5, users in West Lafayette had the smallest mean and median of the user's average mobility range, whereas users in Columbus had the largest. For the four cities, the mean values were larger than the median values (Table 5), indicating that half or more of the distances were smaller than the average distances. The city radius and median mobility range, and the city radius and mean mobility range are found linearly correlated in Figure 7. It should note that the city radius here is used as a measure of the city dimension or size. This is based on the consideration that most often people do not transport across the entire city and people's mobility occurs in all directions. The radius was calculated as the squared root of the city area divided by π. The coefficients of the two models indicate that the average mobility range is about 40% of the city radius. The R square values of these linear models are around 0.99, indicating that they are likely to be capable to predict the mobility range from the area of the city.

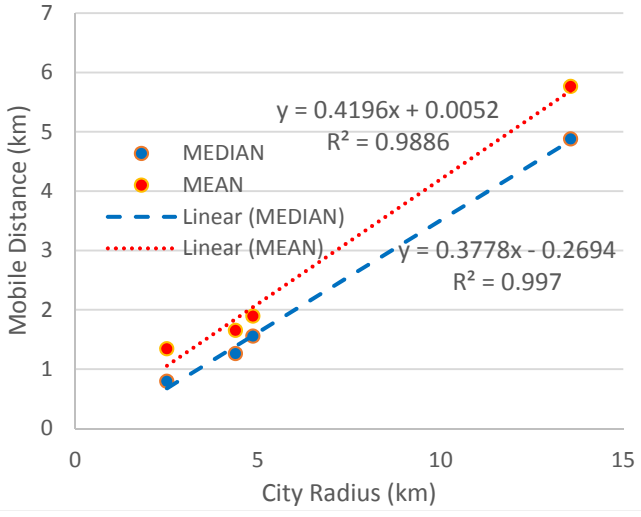

Figure 7. Relationship between city radius and users' average mobile distance.

Figure 8 plots the average mobility ranges (distances) for all frequent users. The distribution is skewed sharply to the larger distances, meaning the number of people living away from campus or workplace drops considerably. In West Lafayette, most users transported less than 1 km with a mean of 1.34 km and a median of 0.78 km (Table 5), inferring that the local residents took short transport to work or school. In Bloomington, the majority of the users' average mobility range was less than 3 km, with mean 1.65 km and median 1.26 km. The average mobility ranges of the users in Ann Arbor ranged from around 0.5 km to 4 km with mean 1.89 km and median 1.5 km. Although the mean and median mobility ranges for Bloomington and Ann Arbor were similar, the values aggregated around the median for Bloomington, while the values for Ann Arbor were more evenly distributed. The mobility ranges of users in Columbus were much longer than that in the other cities, with a mean of 5.76 km and a median of 4.87 km, likely due to the large size of this metropolitan city, and its zoning characteristics as well as the interstate and highway networks that connect the downtown district with neighborhood areas.

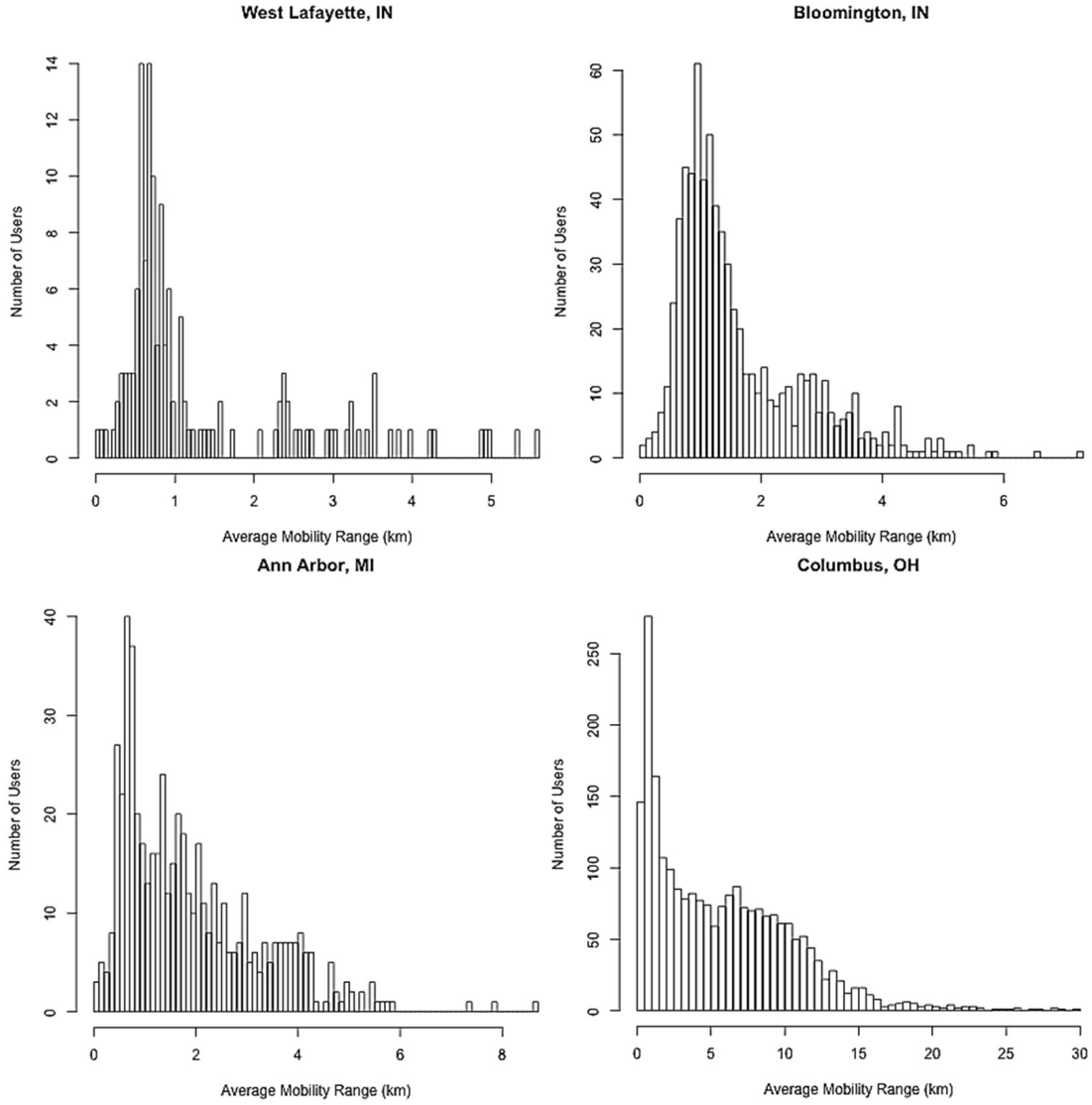

Figure 8. Distribution of average mobility ranges of frequent users.

5. Conclusions

The objective of this effort is to explore the spatial and temporal patterns of geo-tagged tweets from Midwestern college cities/towns and to infer the human mobility patterns of Twitter users. We expect to develop a framework for geospatial data mining for public social media data, e.g., Twitter data. We used various analytical and statistical methods to uncover the time, locations and tweeting behavior of Twitter users, and tested with four Midwest college cities. It was discovered that the majority of tweets are actually posted from a small portion of Twitter users. The four cities share a similar pattern in the time of the day the users' tweet, indicating tweeting time has little to do with the city size or characteristics. Twitter users are most active at 9:00 p.m. at night and 11:00 p.m. during the day. It is shown that there are more Twitter users during weekends than weekdays.

The spatial pattern of tweets distribution varies with the size of the city. In smaller cities, tweets in institutional areas make up the majority of tweets. On the other hand, tweets in residential areas of bigger cities account for the most. Only in big cities tweet clusters are found at shopping malls, while tweets usually aggregate on campuses and apartment complexes in small cities.

Adopting a user-centered view, we also developed a methodology to find the places that people frequently visit and calculated their mobile distances. We found that majority of Twitter users have two to four places of frequent visits. Moreover, the median or mean mobile distance of all users in a city is positively correlated to the size of the city. The average mobile distance is about 40% of the city radius.

However, there are limitations in using tweets in social science studies since the data may be biased for various reasons. There is no current quantitative information available on the socioeconomic structure of Twitter users due to privacy restrictions, though we reasonably believe most Twitter users in West Lafayette, Bloomington and Ann Arbor are related to the local colleges. In addition, since Twitter requires users to opt-in to enable the geo-tag function, the motivation to do this varies with their social behaviors and personalities, or even the rewards of doing so. Thus, Twitter users may not be a complete representative of the public. Furthermore, the tweet data used in this study are only the geo-tagged ones, which are a small part of all tweets that the Streaming API can collect and even a tiny part of the entire Twitter dataset. As such, the discovered knowledge may only reflect the human activity and mobility patterns for a portion of the total population, e.g., college students. The findings found here in college cities in the Midwest of the U.S. may not be fully observed for other places with different urban and demographic structures.

We foresee that the findings of this research can be used to train advanced machine learning techniques to infer the patterns and activities of a larger population of Twitter users, including those who opt-out geo-tagging. As a result, the established method can then be expected to study human dynamics of general public. In addition, inspired from this research using Twitter data, when more social media, cellphone, card transaction, and other data from a broader population are available, we could build deeper and more complete insights on mobility patterns of general public. Another possible and necessary future direction of this type of research is to utilize the content of tweets for text mining, topic modeling, and natural language processing so as to discover deeper knowledge and patterns about human and human behavior. This can facilitate the understanding on the nature of user's activities and the functions of places where frequent tweeting occurs. It can also help to detect space-time tweet clusters and infer the types of gatherings or events. Finally, we may investigate the possibilities of applying the found spatial and temporal patterns into broader fields such as traffic planning, market analysis, urban development, politics, and social science studies.

Acknowledgments: The authors would like to acknowledge the anonymous reviewers for their helpful, detailed and constructive comments, which considerably improved the presentation of the work.

Author Contributions: Y.L. and J.S. conceived and designed the experiments. Y.L. performed the experiments; Q.L. implemented the EM algorithm. Y.L. and J.S. analyzed the data and wrote the paper.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Milstein, S.; Lorica, B.; Magoulas, R.; Hochmuth, G.; Chowdhury, A.; O'Reilly, T. *Twitter and the Micro-Messaging Revolution: Communication, Connections, and Immediacy–140 Characters at a Time*; O'Reilly Media, Inc.: Sebastopol, CA, USA, 2008.
2. Twitter. Available online: http://en.wikipedia.org/wiki/Twitter (accessed on 13 November 2016).
3. Stone, B. Twitter, the startup that wouldn't die. *Bloom. Bus. Week* **2012**, *1*, 62–67.
4. Lunden, I. Analyst: Twitter passed 500m users in June 2012, 140m of them in US; Jakarta 'Biggest Tweeting' city. Available online: https://techcrunch.com/2012/07/30/analyst-twitter-passed-500m-users-in-june-2012-140m-of-them-in-us-jakarta-biggest-tweeting-city/ (accessed on 9 February 2017).
5. Leetaru, K.; Wang, S.; Cao, G.; Padmanabhan, A.; Shook, E. Mapping the global Twitter heartbeat: The geography of Twitter. *First Monday* **2013**, *18*. [CrossRef]
6. Miller, G. Social scientists wade into the tweet stream. *Science* **2011**, *333*, 1814–1815. [CrossRef] [PubMed]
7. Steiger, E.; Albuquerque, J.P.; Zipf, A. An Advanced Systematic Literature Review on Spatiotemporal Analyses of Twitter Data. *Trans. GIS* **2015**, *19*, 809–834. [CrossRef]
8. Goodchild, M.F.; Glennon, J.A. Crowdsourcing geographic information for disaster response: A research frontier. *Int. J. Digit. Earth* **2010**, *3*, 231–241. [CrossRef]
9. Jurgens, D.; Finethy, T.; McCorriston, J.; Xu, Y.T.; Ruths, D. Geolocation Prediction in Twitter Using Social Networks: A Critical Analysis and Review of Current Practice. In Proceedings of the 9th International AAAI Conference on Web and Social Media (ICWSM), Oxford, UK, 26–29 May 2015.
10. Preoţiuc-Pietro, D.; Cohn, T. Mining user behaviours: A study of check-in patterns in location based social networks. In Proceedings of the 5th Annual ACM Web Science Conference, Paris, France, 2–4 May 2013; pp. 306–315.
11. Fujisaka, T.; Lee, R.; Sumiya, K. Discovery of user behavior patterns from geo-tagged micro-blogs. In Proceedings of the 4th International Conference on Ubiquitous Information Management and Communication, Suwon, Korea, 14–15 January 2010.
12. Goodchild, M.F. Citizens as sensors: The world of volunteered geography. *GeoJournal* **2007**, *69*, 211–221. [CrossRef]
13. Carley, K.M.; Malik, M.; Landwehr, P.M.; Pfeffer, J.; Kowalchuck, M. Crowd sourcing disaster management: The complex nature of Twitter usage in Padang Indonesia. *Saf. Sci.* **2016**, *90*, 48–61. [CrossRef]
14. Stevenson, J.R.; Emrich, C.T.; Mitchell, J.T.; Cutter, S.L. Using building permits to monitor disaster recovery: A spatio-temporal case study of coastal Mississippi following Hurricane Katrina. *Cartogr. Geogr. Inf. Sci.* **2010**, *37*, 57–68. [CrossRef]
15. Granell, C.; Ostermann, F.O. Beyond data collection: Objectives and methods of research using VGI and geo-social media for disaster management. *Comput. Environ. Urban Syst.* **2016**, *59*, 231–243. [CrossRef]
16. Zook, M.; Graham, M.; Shelton, T.; Gorman, S. Volunteered geographic information and crowdsourcing disaster relief: A case study of the Haitian earthquake. *World Med. Health Policy* **2010**, *2*, 7–33. [CrossRef]
17. Arcaini, P.; Bordogna, G.; Ienco, D.; Sterlacchini, S. User-driven geo-temporal density-based exploration of periodic and not periodic events reported in social networks. *Inf. Sci.* **2016**, *340*, 122–143. [CrossRef]
18. Nakaji, Y.; Yanai, K. Visualization of real-world events with geotagged tweet photos. In Proceedings of the IEEE International Conference on Multimedia and Expo Workshops (ICMEW), Melbourne, Australia, 9–13 July 2012; pp. 272–277.
19. Crampton, J.W.; Graham, M.; Poorthuis, A.; Shelton, T.; Stephens, M.; Wilson, M.W.; Zook, M. Beyond the geotag: Situating 'big data' and leveraging the potential of the geoweb. *Cartogr. Geogr. Inf. Sci.* **2013**, *40*, 130–139. [CrossRef]
20. Tsou, M.H.; Leitner, M. Visualization of social media: Seeing a mirage or a message? *Cartogr. Geogr. Inf. Sci.* **2013**, *40*, 55–60. [CrossRef]
21. Tsou, M.H.; Yang, J.A.; Lusher, D.; Han, S.; Spitzberg, B.; Gawron, J.M.; An, L. Mapping social activities and concepts with social media (Twitter) and web search engines (Yahoo and Bing): A case study in 2012 US Presidential Election. *Cartogr. Geogr. Inf. Sci.* **2013**, *40*, 337–348. [CrossRef]
22. Ghosh, D.; Guha, R. What are we 'tweeting' about obesity? Mapping tweets with topic modeling and Geographic Information System. *Cartogr. Geogr. Inf. Sci.* **2013**, *40*, 90–102. [CrossRef] [PubMed]

23. Holderness, T.; Kennedy-Walker, R.; Alderson, D.; Evans, B. An evaluation of spatial network modelling to aid sanitation planning in informal settlements using crowd-sourced data. In Proceedings of the International Symposium for Next Generation Infrastructure, Wollongong, Australia, 1–4 October 2013; pp. 185–192.
24. Roick, O.; Heuser, S. Location based social networks–definition, current state of the art and research Agenda. *Trans. GIS* **2013**, *17*, 763–784. [CrossRef]
25. Gonzalez, M.C.; Hidalgo, C.A.; Barabasi, A.L. Understanding individual human mobility patterns. *Nature* **2008**, *453*, 779–782. [CrossRef] [PubMed]
26. Hawelka, B.; Sitko, I.; Beinat, E.; Sobolevsky, S.; Kazakopoulos, P.; Ratti, C. Geo-located Twitter as proxy for global mobility patterns. *Cartogr. Geogr. Inf. Sci.* **2014**, *41*, 260–271. [CrossRef] [PubMed]
27. Caverlee, J.; Cheng, Z.; Sui, D.Z.; Kamath, K.Y. Towards Geo-Social Intelligence: Mining, Analyzing, and Leveraging Geospatial Footprints in Social Media. *IEEE Data Eng. Bull.* **2013**, *36*, 33–41.
28. Peng, C.; Jin, X.; Wong, K.; Shi, M.; Lio, P. Collective human mobility pattern from taxi trips in urban area. *PLoS ONE* **2012**, *7*, e34487.
29. Liu, Y.; Wang, F.; Xiao, Y.; Gao, S. Urban land uses and traffic 'source-sink areas': Evidence from GPS-enabled taxi data in Shanghai. *Landsc. Urban Plan.* **2012**, *106*, 73–87. [CrossRef]
30. Zheng, V.; Zheng, Y.; Xie, X.; Yang, Q. Towards mobile intelligence: Learning from GPS history data for collaborative recommendation. *Artif. Intell.* **2012**, *184*, 17–37. [CrossRef]
31. Cheng, Z.; Caverlee, J.; Lee, K.; Sui, D. Exploring millions of footprints in location sharing services. In Proceedings of the Fifth International AAAI Conference on Weblogs and Social Media, Barcelona, Spain, 17–21 July 2011 ; pp. 81–88.
32. Liang, X.; Zhao, J.; Dong, L.; Xu, K. Unraveling the origin of exponential law in intra-urban human mobility. *Sci. Rep.* **2013**, *3*, 2983. [CrossRef] [PubMed]
33. Bagrow, J.P.; Lin, Y. Mesoscopic structure and social aspects of human mobility. *PLoS ONE* **2012**, *7*, e37676. [CrossRef] [PubMed]
34. Bayir, M.A.; Demirbas, M.; Eagle, N. Discovering spatiotemporal mobility profiles of cellphone users. In Proceedings of the 2009 IEEE International Symposium on a "World of Wireless, Mobile and Multimedia Networks & Workshops" (WoWMoM 2009), Kos, Greece, 15–19 June 2009; pp. 1–9.
35. Yang, X.; Fang, Z.; Xu, Y.; Shaw, S.L.; Zhao, Z.; Yin, L.; Zhang, T.; Lin, Y. Understanding Spatiotemporal Patterns of Human Convergence and Divergence Using Mobile Phone Location Data. *ISPRS Int. J. Geo-Inf.* **2016**, *5*, 177. [CrossRef]
36. Hasan, S.; Schneider, C.M.; Ukkusuri, S.V.; González, M.C. Spatiotemporal patterns of urban human mobility. *J. Stat. Phys.* **2013**, *151*, 304–318. [CrossRef]
37. Gong, L.; Liu, X.; Wu, L.; Liu, Y. Inferring trip purposes and uncovering travel patterns from taxi trajectory data. *Cartogr. Geogr. Inf. Sci.* **2016**, *43*, 103–114. [CrossRef]
38. Huang, Q.; Wong, D.W. Activity patterns, socioeconomic status and urban spatial structure: What can social media data tell us? *Int. J. Geogr. Inf. Sci.* **2016**, *30*, 1873–1898. [CrossRef]
39. Gao, S.; Yang, J.A.; Yan, B.; Hu, Y.; Janowicz, K.; McKenzie, G. Detecting Origin-Destination Mobility Flows From Geotagged Tweets in Greater Los Angeles Area. In Proceedings of the Eighth International Conference on Geographic Information Science (GIScience'14), Vienna, Austria, 24–26 September 2014.
40. Girardin, F.; Calabrese, F.; Fiore, F.D.; Ratti, C.; Blat, J. Digital footprinting: Uncovering tourists with user-generated content. *IEEE Pervasive Comput.* **2008**, *7*, 36–43. [CrossRef]
41. Leung, R.; Vu, H.Q.; Rong, J.; Miao, Y. Tourists Visit and Photo Sharing Behavior Analysis: A Case Study of Hong Kong Temples. In *Information and Communication Technologies in Tourism 2016*; Springer International Publishing: Cham, Switzerland, 2016; pp. 197–209.
42. Popescu, A.; Grefenstette, G.; Moëllic, P.A. Mining tourist information from user-supplied collections. In Proceedings of the 18th ACM conference on Information and knowledge management, Hong Kong, China, 2–6 November 2009; pp. 1713–1716.
43. Li, L.; Goodchild, M.F.; Xu, B. Spatial, temporal, and socioeconomic patterns in the use of Twitter and Flickr. *Cartogr. Geogr. Inf. Sci.* **2013**, *40*, 61–77. [CrossRef]
44. Li, Y.; Shan, J. Understanding the Spatio-temporal Pattern of Tweets. *Photogramm. Eng. Remote Sens.* **2013**, *79*, 769–773.
45. West Lafayette, Indiana. Available online: http://en.wikipedia.org/wiki/West_Lafayette,_Indiana (accessed on 5 November 2014).

46. Purdue University. Available online: http://en.wikipedia.org/wiki/Purdue_University (accessed on 1 August 2013).
47. Bloomington, Indiana. Available online: http://en.wikipedia.org/wiki/Bloomington,_Indiana (accessed on 5 November 2014).
48. Housing. Available online: http://www.iub.edu/student/housing.shtml (accessed on 5 November 2014).
49. Ann Arbor, Michigan. Available online: http://en.wikipedia.org/wiki/Ann_Arbor,_Michigan (accessed on 6 November 2014).
50. Housing Options. Available online: http://housing.umich.edu/options (accessed on 6 November 2014).
51. Columbus, Ohio. Available online: http://en.wikipedia.org/wiki/Columbus,_Ohio (accessed on 6 November 2014).
52. The Ohio State University. Available online: http://en.wikipedia.org/wiki/Ohio_State_University (accessed on 6 November 2014).
53. The Streaming APIs Overview. Available online: https://dev.twitter.com/streaming/overview (accessed on 6 November 2014).
54. Tweepy. Available online: http://www.tweepy.org/ (accessed on 4 February 2017).
55. Twitter-Streamer. Available online: https://github.com/inactivist/twitter-streamer (accessed on 4 February 2017).
56. API Rate Limit. Available online: https://dev.twitter.com/rest/public/rate-limiting (accessed on 4 February 2017).
57. Pesyna, K.M., Jr.; Heath, R.W., Jr.; Humphreys, T.E. Centimeter Positioning with a Smartphone-Quality GNSS Antenna. GPS World, 2 February 2015. Available online: http://gpsworld.com/accuracy-in-the-palm-of-your-hand/ (accessed on 10 January 2017).
58. Dempster, A.P.; Laird, N.M.; Rubin, D.B. Maximum likelihood from incomplete data via the EM algorithm. *J. R. Stat. Soci.* **1977**, *39*, 1–38.
59. Aggarwal, C.C.; Reddy, C.K. *Data Clustering: Algorithms and Applications*; CRC Press: Boca Raton, FL, USA, 2013.
60. Gan, G.; Ma, C.; Wu, J. *Data Clustering: Theory, Algorithms, and Applications*; SIAM: Philadelphia, PA, USA, 2007; Volume 20.
61. Ambroise, C.; Dang, M.; Govaert, G. Clustering of spatial data by the EM algorithm. In *geoENV I—Geostatistics for Environmental Applications*; Springer Science+Business Media B.V.: Dordrecht, The Netherlands, 1997; pp. 493–504.