2018 21st International Conference on Intelligent Transportation Systems (ITSC)
Maui, Hawaii, USA, November 4-7, 2018

Predictability in Human Mobility based on Geographical-boundary-free and Long-time Social Media Data*

Yuan Liao[1], *IEEE Student Member* and Sonia Yeh[1]

Abstract— **Understanding of predictability in human mobility benefits a broad spectrum such as urban planning and traffic forecasting. In human mobility studies, geotagged social media data are being gradually accepted as a user-contributed data source. It remains unclear to what extent we can use geotagged social media data to predict individual mobility. In the present study, a dataset is collected and applied which includes 652,945 geotagged tweets generated by 2,933 Swedish users covering time spans of more than one year (3.6 years on average). Based on such a dataset, human mobility predictability has been explored from three aspects: 1) time history of mobility range indicating how people diffuse in space, 2) entropy and the corresponding predictability of mobility, and 3) the limits of predictability dependent on geographical boundaries and mobility range. This study reveals a dataset that captures Twitter users' mobility where they routinely visit a couple of regions at most of the time and occasionally explore new regions. A 70% potential predictability is obtained by measuring the entropy of each individual's geotagged activity trajectory using a half-day time interval. The predictability's dependence on mobility range is prolonged when the observation of mobility is geographical-boundary-free which also decreases predictability.**

Index Terms— **Human mobility, geotagged activity trajectory, information theory, entropy, predictability.**

I. INTRODUCTION

Understanding of human mobility patterns benefits a broad spectrum from urban planning [1] to human virus prediction [2] and traffic forecasting [3]. The rapid development of information and communication technology (ICT) has broadened the understanding of human mobility patterns [4] through emerging data sources, such as banknote data [5], mobile phone data [6], and geotagged social media data [7]. The degree to which individual human whereabouts are predictable has been revealed [8], [9], [7]. However, the data sources applied in most previous studies suffer from geographical boundaries, short time span, and limited population coverage. It remains unclear that to what extent we can use geotagged social media data to predict individual mobility.

A. Related work

There have been a series of models characterising human mobility that are fundamentally stochastic [8] such as Lévy-walk models and their applications in viral dynamics. A high degree of temporal and spatial regularity of human mobility has been revealed [6] which has also been confirmed by the studies using banknote data [5] and mobile phone data [6]. People present a significant probability to return to a few highly frequented locations. Song et al further applied the information theory to reveal the role of randomness in human mobility behaviour and the degree to which individual human actions are predictable [8]. A 93% potential predictability in user mobility across the user base is obtained by measuring the entropy of each individual's trajectory [8]. And a remarkable lack of variability in predictability exists despite the significant differences in the travel patterns [8]. Another study reveals a 88% potential predictability in mobile phone user's mobility and it turns out such fundamental theoretical limit for potential predictive power is an approachable target for actual prediction accuracy [9].

Among the emerging data sources from ICT, social media data are being gradually accepted as user-contributed data sources in travel behaviour studies such as large-scale urban activity [10] and mobility patterns [11]. Geotagged tweets have proved a useful proxy for tracking and predicting human movement [7]. Jurdak et al. have applied a similar analysis framework, as shown in previous studies where the authors use mobile phone data [8], [9], to quantify the predictability of individual Tweet locations [7]. The major limitations of their study include that: 1) the analysis is limited to a certain area (Australia), 2) the dataset covers a short time range (8 months), and 3) the limits of predictability have not been revealed, meaning the predictability of using Twitter data has not been clearly quantified [7]. Compared to mobile phone data in a small area with limited population size and time span [8], [9], the utilisation of Twitter data has its pros and cons. Twitter data have low and irregular sampling frequency which means their time resolution is worse than mobile phone data. Despite that sampling issue, geotagged tweets have potentials to cover a long time period without any geographical boundaries [12]. The spatial resolution of geotagged tweets is as high as around 10 m [13], [7].

B. Objectives of this study

To date, there has been little work exploring the human mobility predictability in a larger spatial and temporal scale. This study attempts to quantify that to what extent we can predict human mobility using geotagged social media data. The dataset includes the 652,945 geotagged tweets generated by 2,933 Swedish users covering time spans of more than one year (3.6 years on average).

The remainder of this paper is organised as follows. Section 2 introduces the methods, including data collection

*This research is funded by the Swedish Research Council Formas (Project Number 2016-1326).

[1] Yuan Liao and Sonia Yeh are with Division of Physical Resource Theory, Department of Space, Earth and Environment, Chalmers University of Technology, Gothenburg 41296, Sweden yuan.liao@chalmers.se; sonia.yeh@chalmers.se

978-1-7281-0323-5/18/$31.00 ©2018 IEEE

and pre-processing, mobility range, identification of regions, and calculation of entropy and predictability. Section 3 presents the results and discussion where each subsection is organised as a description of results followed by a paragraph of discussion. Section 4 summarises the study.

II. METHODS

In this section, we first present the data collection and pre-processing methods. Based on geotagged locations, the radius of gyration (r_g), as an indicator of mobility range, is introduced together with the data processing method to produce a time history of r_g. Then we introduce the method that is used to identify regions of locations. Based on the regions rather than locations, the calculation of entropy and predictability are illustrated.

A. Data collection and pre-processing

In our previous study, we have identified 5000 non-commercial geo-users who geotagged their tweets most frequently during a six-month period (20 December 2015 - 20 June 2016) within the geographical bounding box of Sweden using Gnip database [14]. We extract these top geotagged users' historical tweets (without applying a spatial boundary limit) from user timeline by applying tweepy Python package [15]. The data is limited to 3200 tweets per user. This method produces varied time span and varied tweet number as not all users reached the 3200 tweets maximum. Besides time span and tweet number, the tweeting frequency also varies greatly among users.

We further apply the following rules to pre-process our data to ensure the studied individuals reside within Sweden and have a substantial number of geotagged tweets to reasonably capture their activity trajectories. The rules include: 1) the covered time span is above 1 year, 2) the geotweeting frequency (geotagged tweets /day) is above 0.1, or the total amount of collected geotagged tweets is above 50, and 3) the most frequently visited region is in Sweden. After screening, we identify 2933 users and 652,945 geotagged tweets.

The locations that a user visited is first captured using all geotagged tweets by that user with the time stamps: $A = (X1, X2, t)_k$, $k = 1, 2, ..., N$ where $X1$ is the decimal degree of Latitude, $X2$ is the decimal degree of Longitude, t the time stamp (UTC) of the $k-th$ location. N is the total number of locations visited by the user i through his/her geotagged tweets, and T is the total time span. The number of distinct locations of a user is usually smaller than the total number of locations being visited or in an extreme case is equal. Let n be the number of distinct locations, p_j be the visiting frequency rate of location j, and r_j be the rank of location j by the visiting frequency p_j. The vector of visited unique locations is therefore:

$$\mathbf{A} = (X1, X2, p, r)_j, j = 1, 2, , n \tag{1}$$

B. Mobility range indicated by the radius of gyration (r_g)

The radius of gyration, which combines the locations' geographical distribution and their visiting frequency, has been widely applied to characterising human mobility patterns [6], [8], [9], [7]. Someone who moves in a comparatively confined space will have a small radius of gyration even though he or she covers a large distance [9]. The total radius of gyration r_g is defined as:

$$r_g = \sqrt{\frac{1}{n}\sum_{q=1}^{n} p_q \cdot (\mathbf{r}_q - \mathbf{r}_{cm})^2} \tag{2}$$

where $\mathbf{r}_q = [X1, X2]_q$ and the mass centre of the visited locations $\mathbf{r}_{cm}$ is defined as

$$\mathbf{r}_{cm} = \left[\sum_{q=1}^{q=n}(X1 \cdot p_q)/\sum_{q=1}^{q=n} X1, \sum_{q=1}^{q=n}(X2 \cdot p_q)/\sum_{q=1}^{q=n} X2\right] \tag{3}$$

To get the time history of r_g, we first sort the distinct locations of each individual based on the visiting frequency. Then we anchor the time that the top-one location has been visited for the first time in one's trajectory of geotagged tweets. Assuming the top one location is part of the user's daily routine, the calculation of r_g time history starts from that time until 90 days later with a manner that updates r_g when a geotagged tweet appears. By doing so, we extract a 90-day episode from each user. To make the time sequence comparable, each episode is required to contain at least 10 instances of r_g. We regulate the individual time sequence into the same length (50 data points), applying either the nearest-neighbour interpolation to the sequence shorter than 50 or the randomly down-sampling the sequence longer than 50. Hence, we get a 90-day sequence of r_g for each user who satisfies the conditions above (2303 valid users in total).

C. Identification of regions

The resolution of geotagged locations is around 10 m. The distinct locations are too big to calculate the entropy with limited number of observations. The proximate regions covering the raw locations data can represent the area of interest with a lower spatial resolution. Therefore, DBSCAN clustering is applied to the overall geotagged locations merging which are geographically close into a cluster, i.e., region [16]. The advantage of DBSCAN is that it can identify clusters of arbitrary shape [16], [7]. The distance threshold (eps) for merging is set as 0.25 km, 1 km, and 10 km. The minimum number of location for a region is set as 1. The number of distinct regions is indicated by n'. In the calculation of entropy and the corresponding predictability, the identified regions are applied rather than the raw locations.

D. Calculation of entropy

The random entropy is $S^{rand} = \log n'$ capturing the degree of predictability of the user's whereabouts if each region is visited with equal probability [8]. The unconditional entropy can be expressed as $S^{unc} = -\sum_{j=1}^{n'} p(j) \log p(j)$ characterising the heterogeneity of visitation patterns.

To integrate the time dimension into a user's geotagged activity patterns, the actual entropy can be defined as $S^{real} = -\sum_{\mathbf{A}' \subset \mathbf{A}} P(A')\log[P(A')]$, where $P(\mathbf{A}')$ is the probability of finding a particular time-ordered sub-sequence $\mathbf{A}'$ in the trajectory of A. The mobility sub-sequence could be the 24-hour sequence representing the regions that the user is moved at each consecutive hourly interval, as defined in the previous study using mobile phone data [8]. However, such definition of sub-sequence requires a high frequency of dataset which is not satisfied in this study. Instead of focusing on the hour-interval, the sub-sequence of geotagged regions is defined as the regions that the user is observed at each half-day interval. For the case where there are more than 1 region reported during the half-day interval, the region is randomly select from the generated regions. For the case where there is no region reported during the half-day interval, the region is represented by the nearest reported region. This re-sampling process based on the individual geotagged activity trajectory makes the geotagged regions evenly spaced in time. However, it also introduces some uncertainties; the users who frequently geotweet get fewer instances than being observed, and those who less frequently do that get more interpolated instances. The block of regions (sub-sequence) is extracted daily with the length of sub-sequence (m) as 2. The block of regions is also extracted weekly with $m = 14$, and some length values in between.

E. Definition of predictability

A measure of predictability could be defined as the probability Π that an appropriate predictive algorithm can predict correctly the user's future geotagged region [8]. Fano's inequality has been proposed to quantify such probability [17], [8], [7]. Let X be the set of messages containing K possible values, Y the corresponding estimation of X. Fano's inequality relates to the probability of incorrectly estimating X from Y. In the context of predicting the regions that a user may visit, X is equivalent to the set of regions where a user visits and Y is the set of regions where a user is observed through his/her geotagging activity. The maximum bound of the user's predictability $\Pi \leq \Pi^{\max}$ is determined by:

$$S = H(\Pi^{\max}) + (1 - \Pi^{\max})log(n' - 1) \tag{4}$$

where S refers to $S^{rand}/S^{unc}/S^{real}$ and the binary entropy function H can be expressed by the following equation.

$$H = -\Pi^{\max}log(\Pi^{\max}) - (1 - \Pi^{\max})log(1 - \Pi^{\max}) \tag{5}$$

$\Pi^{\max}$ represents the fundamental limit for each individual's predictability of their geotagged activity patterns [8]. For instance, a user has a $\Pi^{\max} = 0.3$ means that the user randomly choose his/her geotagged location at 70% of his/her time.

III. Results and Discussion

In this section, we first briefly summarise the characteristics of the applied geotagged activity dataset. Next, the diffusion process in space is illustrated in a 90-day period. To reveal the regularity of geotagged mobility, the distribution of different regions is displayed regarding their frequency of being visited. In Subsection C, we present the results of entropy predictability across three scales of regions with eps is set to 0.25 km, 1 km, and 10 km. At Subsection D, we focus on the regions that are produced using $eps = 0.25$ km. The maximum predictability of half-day time interval is presented. And such limits of predictability are explored depending on the mobility range and geographical bounding box.

A. Characteristics of geotagged activity dataset

Geotagged tweets of Swedish users are distributed globally without applying any geographical boundaries and a large proportion of geotagged locations are within Sweden (Figure 1A). The individual geotagged activity is unevenly distributed in time (Figure 1B).

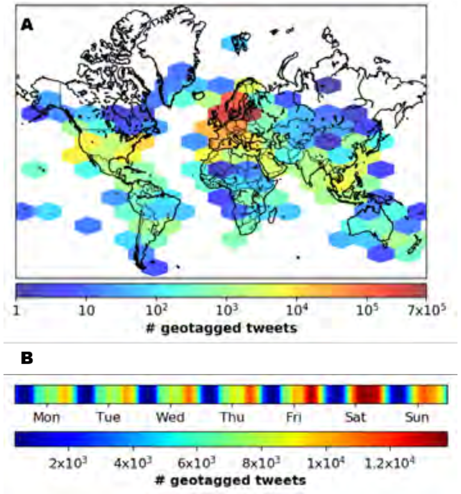

Fig. 1. Characteristics of geotagged activity trajectory of Swedish users. (A) Geotagged tweets on the map. (B) A week-long geotagging activity pattern that captures the time-dependent characteristic of geotagged locations. The frequency is calculated with a resolution of one-hour interval.

The 652,945 geotagged tweets contain 435,569 distinct locations (66.7%). Those distinct locations are further merged into regions on various scales. It turns out that the dataset contains 66,342 regions of 0.25 km ($eps = 0.25$, same following) with a compression rate $r_c = 84.8\%$, 32,773 regions of 1 km ($r_c = 92.5\%$), and 7,364 regions of 10 km ($r_c = 98.3\%$).

The percentage of distinct locations/regions quantifies the variance level of geotagged locations (Figure 2). The more geotagged locations that are outside the habitually visited locations, the larger variance level. The geotagged activity dataset captures both routine activities, e.g., return to home and work, and exploratory mobility. The proportion of distinct locations suggest that the geotagged tweets do not necessarily over-represent either one type of activity (routine vs. exploratory) but rather evenly spread out. When the locations are further merged into regions, the distribution shifts to smaller proportions of distinct regions among all visited regions. It suggests that this dataset captures both cases of Twitter users: routinely visit a couple of regions at most of time and occasionally explore new regions. It's worth noting that *eps* represents the maximum distance between locations within each identified region, not necessarily implying the diameter of such a region is equal to the value of *eps* setting.

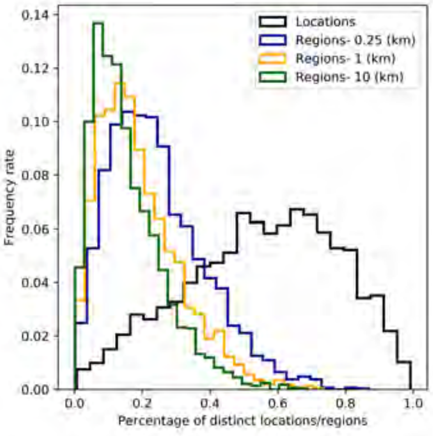

Fig. 2. The distribution of the proportion of distinct geotagged locations/regions across the users.

B. Diffusion process

Figure 3A shows the time history of an average diffusion process during 90 characterised by the radius of gyration. If individual trajectories follow the random walk [18], then the radius of gyration should follow the solid grey line $r_g(t) \sim t^{1/2}$. It suggests that, even after 90 days, the mobility range still remains increasing. Figure 3B shows the cumulative distribution function of the visiting frequency rate vs the ranking order of the most visited location/region. The cumulative frequency rate reflects the regularity of users' visiting behaviour. Compared to the curve of locations, the regions quickly saturate highlighting the returning effect of human mobility.

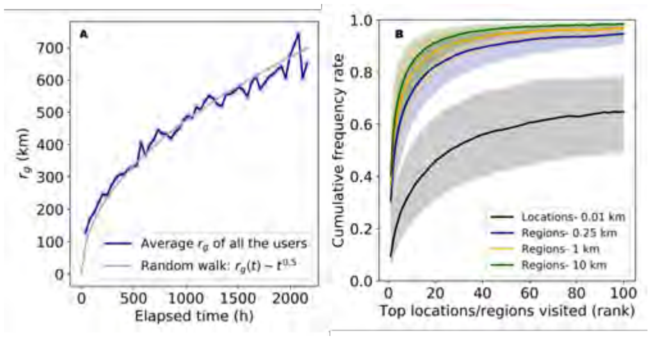

Fig. 3. Diffusion process. (A) Time history of radius of gyration within 90 days. The time history starts from the first time observing the most visited location; each data point indicates the mean value of r_g across all the users. (B) Visiting frequency by the ranking order of the most visited locations/regions. The shaded range indicates the upper bound (75%) and lower bound (25%) of the cumulative frequency rate of visits.

The time history of r_g reveals the stabilising effect of r_g. The mechanism stabilising r_g has been revealed as the strong tendency of humans to return to locations they visited before [6], which can be confirmed by Figure 3B.

C. Entropy and predictability

The distributions of users' random entropy (S^{rand}) and temporal-uncorrelated entropy (S^{unc}) are presented in Figure 4. The median value of S^{rand} for $eps = 0.25$ is around 5.7, indicating that if we assume that individuals randomly choose a region to geotag the next time, a typical explorer could be found in any of $2^{5.7} = 52$ regions. Not surprisingly, the probability calculated from S^{rand} is close to zero (Figure 4C) given the fact that people do not move around randomly. The temporal-uncorrelated entropy (S^{unc},), as implied by its definition, integrates the frequency and sequence order of the geotagged activity trajectory (Figure 4B). The median value of S^{unc} is around 4.1 decreasing the uncertainty of their whereabouts to $2^{4.1} \approx 17$ regions ($eps = 0.25$). Correspondingly, the predictability using S^{unc} increases (Figure 4D) where we can get around 45%.

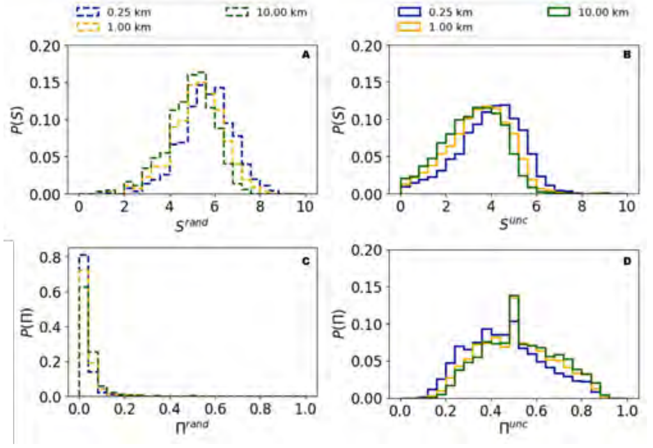

Fig. 4. Random entropy (S^{rand}) & temporal-uncorrelated entropy (S^{unc}) and their corresponding predictability. (A) The distribution of random entropy (S^{rand}). (B) The distribution of temporal-uncorrelated entropy (S^{unc}). (C) The distribution of random predictability (Π^{rand}). (D) The distribution of temporal-uncorrelated predictability (Π^{unc}).

The real entropy (S^{real}_m) is presented in Figure 5 in a comparison with random entropy (S^{rand}) and temporal-uncorrelated entropy (S^{unc}), and the corresponding predictability (Figure 5). For most users, the predictability peaks around the block length of 2, i.e., the two locations daily generated. We get finite length of location sequence from each user where the realisation number of a sub-sequence decreases with increasing length of sub-sequence (m). The larger the m is, the more difficult it is to find the patterns. This explains why the increasing rate of entropy decreases when m increases, however, the decreasing rate of predictability does not follow the same tendency.

The random entropy only depends on the number of distinct regions. The mobility is naturally bounded by the returning mechanism [6], suggesting the limited number of regions that are visited more frequently. The temporal-uncorrelated entropy adds the visiting frequency rate to

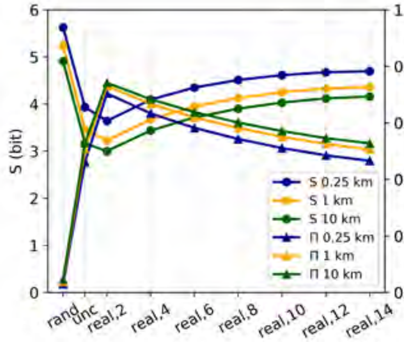

Fig. 5. Real entropy (S_m^{real}), the corresponding predictability, and a comparison with S^{rand} and S^{unc}.

describe the stationary probabilistic distribution of visited regions. Adding time dimension, the real entropy describes the block entropy of daily visits. Humans are habitual animals; the real entropy further increases the knowledge on their daily mobility. Not surprisingly, the predictability elicited from the real entropy is higher than the random entropy and the stationary entropy. When the block length increases, the real entropy increases and slowly saturates and stabilises. However, with finite length of data points, increased length means decreased reliability of the probabilistic description brought by the decreased sample size.

D. Limits of predictability with half-day time interval

Taking $eps = 0.25$ km for instance, the maximum predictability of half-day mobility ($m = 2$) is individually defined ($\Pi_{\max}$), which does not follow the fat-tailed distribution (Figure 6). $\Pi_{\max}$ of all geotagged regions is around 70%, which is significantly lower than the $\Pi_{\max}$ where the geographical bounding box exists (73%) indicated by Mann-Whitney U test ($p < 0.001$).

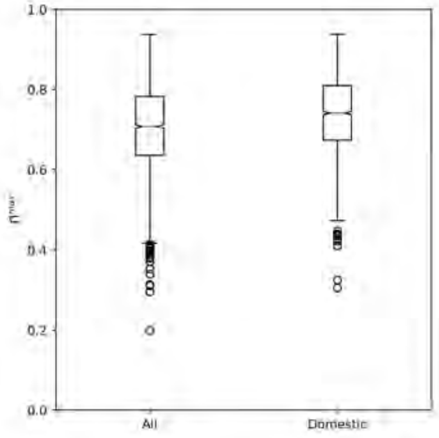

Fig. 6. Distribution of maximum predictability ($\Pi_{\max}$) with $eps = 0.25$ km: all geotagged tweets included vs domestic geotagged tweets filtered by the geographical bounding box of Sweden.

As shown in Figure 7A, there is a steady decrease of $\Pi_{\max}$ when r_g increases. When the r_g keeps increasing, $\Pi_{\max}$ gradually stays around 0.66, indicating the independence of $\Pi_{\max}$ on the large mobility range and users randomly choose their geotagged location in 34% of their time on average. Figure 7B shows that $\Pi_{\max}$ increases when the ratio of domestic geotagged tweets increases, which explains why the geographical bounding box leads to a higher predictability shown in Figure 6 (Domestic vs All).

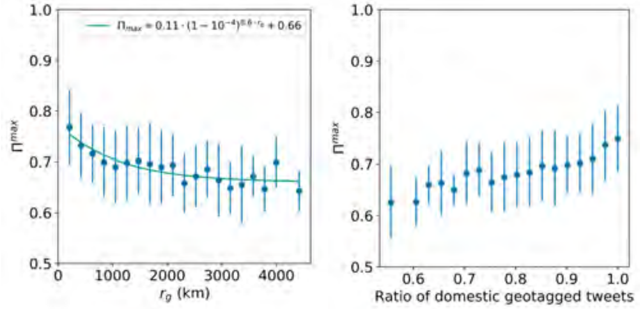

Fig. 7. Dependence of predictability ($eps = 0.25$). (A) The dependence on the user's radius of gyration r_g, capturing the distances regularly covered by each user. (B) The dependence on the user's ratio of domestic geotagged tweets. Each data point represents the median value of more than 25 users.

An overall 70% potential predictability in user mobility is obtained. It's worth noting that compared to the previous studies [8], [9], this result is obtained without any geographical boundaries covering the time period of 3.6 years on average. The range of r_g is only covered with $[0, 1000]$ km by previous studies while in the present study, the application of geotagged social media data extends that range to 4000 km. Although the previous studies suggest an independence of high predictability on the radius of gyration [8], the current study presents the predictability's dependence on r_g is prolonged when the observation of human mobility is geographical-boundary-free. The positive correlation between $\Pi_{\max}$ and ratio of domestic geotagged tweets further displays the effect of geographical bounding box. When most regions are observed in Sweden, the dataset of such users captures more regular mobility therefore the corresponding $\Pi_{\max}$.

IV. CONCLUSIONS

Focusing on geotagged social media data, this study has quantified that to what extent we can predict human mobility as geotagged by the Twitter user. The main contributions include:

1) The applied dataset spans a long time period (3.6 years on average), and without any geographical boundaries. The dataset captures Twitter users' mobility where they routinely visit a couple of regions at most of the time and occasionally explore new regions.

2) A 70% potential predictability in user mobility is obtained by measuring the entropy of each individual's geotagged activity trajectory using a half-day time interval.

3) Geographical boundary affects the results of mobility predictability. The predictability's dependence on r_g is prolonged when the observation of human mobility is

geographical-boundary-free. The existence of geographical bounding box increases predictability.

One limitation of the present study stems from the low and irregular sampling frequency. Although the users that we analyse present highly active geotagging behaviour, the number of geotagged tweet per day is 0.32 on average. Compared to the 5 to 6 locations daily visited identified by the mobile phone data [8], individually predicting human mobility with geotagged tweets is constrained by the poor time resolution. There is a possibility that long sampling period can compensate for this shortcoming. The other limitation is the lack of actual human mobility trajectories serving as the ground truth. Hence it is hard to generalise the findings from the geotagged human mobility to general human mobility.

To address the above limitations, the integration of multiple data sources remains to be done, such as survey, twitter, mobile phone, and GPS data to reach a deeper understanding of the relationship between geotagged locations and actual travel trajectories. With that understanding and the awareness of the limits of predictability in human mobility, as revealed in this study, we can further develop mobility models leveraging geotagged social media data for real-world applications.

References

[1] A. Noulas, S. Scellato, R. Lambiotte, M. Pontil, and C. Mascolo, "A tale of many cities: universal patterns in human urban mobility," *PloS one*, vol. 7, no. 5, p. e37027, 2012.

[2] D. Balcan, V. Colizza, B. Gonçalves, H. Hu, J. J. Ramasco, and A. Vespignani, "Multiscale mobility networks and the spatial spreading of infectious diseases," *Proceedings of the National Academy of Sciences*, vol. 106, no. 51, pp. 21 484–21 489, 2009.

[3] A. Kesting and M. Treiber, "Traffic flow dynamics: data, models and simulation," *no. Book, Whole)(Springer Berlin Heidelberg, Berlin, Heidelberg, 2013)*, 2013.

[4] Z. Zhang, Q. He, and S. Zhu, "Potentials of using social media to infer the longitudinal travel behavior: A sequential model-based clustering method," *Transportation Research Part C: Emerging Technologies*, vol. 85, pp. 396–414, 2017.

[5] D. Brockmann, L. Hufnagel, and T. Geisel, "The scaling laws of human travel," *Nature*, vol. 439, no. 7075, pp. 462–465, Jan. 2006. [Online]. Available: https://www.nature.com/nature/journal/v439/n7075/full/nature04292.html

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

[7] R. Jurdak, K. Zhao, J. Liu, M. AbouJaoude, M. Cameron, and D. Newth, "Understanding human mobility from twitter," *PloS one*, vol. 10, no. 7, p. e0131469, 2015.

[8] C. Song, Z. Qu, N. Blumm, and A.-L. Barabási, "Limits of predictability in human mobility," *Science*, vol. 327, no. 5968, pp. 1018–1021, 2010.

[9] X. Lu, E. Wetter, N. Bharti, A. J. Tatem, and L. Bengtsson, "Approaching the limit of predictability in human mobility," *Scientific reports*, vol. 3, p. srep02923, 2013.

[10] S. Gao, J. A. Yang, B. Yan, Y. Hu, K. Janowicz, and G. McKenzie, "Detecting origin-destination mobility flows from geotagged tweets in greater los angeles area," 2014.

[11] S. Hasan, C. Schneider, S. Ukkusuri, and M. Gonzlez, "Spatiotemporal patterns of urban human mobility," *Journal of Statistical Physics*, vol. 151, no. 1-2, pp. 304–318, 2013. [Online]. Available: http://dx.doi.org/10.1007/s10955-012-0645-0

[12] Y. Liao, S. Yeh, and G. Jeuken, "From individual to collective behaviours: Exploring features of human mobility in space and time based on social media data," *Transportation Research Part C: Emerging Technologies*, Submitted.

[13] H. Bar-Gera, "Evaluation of a cellular phone-based system for measurements of traffic speeds and travel times: A case study from Israel," *Transportation Research Part C: Emerging Technologies*, vol. 15, no. 6, pp. 380–391, Dec. 2007. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0968090X07000393

[14] G. Jeuken, S. Yeh, and Y. Liao, "Revealing patterns of mobility and travel distance from social media with twitter," *Working Paper*.

[15] The Tweepy project developers , "Tweepy: v3.5.0," Apr. 2017. [Online]. Available: http://tweepy.readthedocs.io/en/v3.5.0/

[16] M. Ester, H.-P. Kriegel, J. Sander, X. Xu, *et al.*, "A density-based algorithm for discovering clusters in large spatial databases with noise." in *Kdd*, vol. 96, no. 34, 1996, pp. 226–231.

[17] R. M. Fano and W. Wintringham, "Transmission of information," *Physics Today*, vol. 14, p. 56, 1961.

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