Frontiers of Information Technology & Electronic Engineering
www.zju.edu.cn/jzus; engineering.cae.cn; www.springerlink.com
ISSN 2095-9184 (print); ISSN 2095-9230 (online)
E-mail: jzus@zju.edu.cn

Home location inference from sparse and noisy data: models and applications*

Tian-ran HU, Jie-bo LUO‡, Henry KAUTZ, Adam SADILEK
(*Computer Science Department, University of Rochester, NY 14623, USA*)
E-mail: thu@cs.rochester.edu; jluo@cs.rochester.edu; kautz@cs.rochester.edu; sadilek@cs.rochester.edu

Received Nov. 7, 2015; Revision accepted Feb. 19, 2016; Crosschecked Apr. 11, 2016

Abstract: Accurate home location is increasingly important for urban computing. Existing methods either rely on continuous (and expensive) Global Positioning System (GPS) data or suffer from poor accuracy. In particular, the sparse and noisy nature of social media data poses serious challenges in pinpointing where people live at scale. We revisit this research topic and infer home location within 100 m×100 m squares at 70% accuracy for 76% and 71% of active users in New York City and the Bay Area, respectively. To the best of our knowledge, this is the first time home location has been detected at such a fine granularity using sparse and noisy data. Since people spend a large portion of their time at home, our model enables novel applications. As an example, we focus on modeling people's health at scale by linking their home locations with publicly available statistics, such as education disparity. Results in multiple geographic regions demonstrate both the effectiveness and added value of our home localization method and reveal insights that eluded earlier studies. In addition, we are able to discover the real buzz in the communities where people live.

Key words: Home location, Mobility patterns, Healthcare
http://dx.doi.org/10.1631/FITEE.1500385 **CLC number:** TP391

1 Introduction

Home, as one of the most important locations in people's mobility patterns, is the key to understanding many aspects of urban life and environments. With the knowledge of where people actually live, researchers are able to model the distribution of population, study human mobility patterns, infer life styles, and even discover the correlation between home location and other important aspects such as health conditions, disease diffusion, and environment changes.

Much of the research in the above mentioned areas is based on surveys and census, which are costly and often incur a delay that hampers real-time analysis and response. Fortunately, the wide adoption of geo-tagged social media provides us a new opportunity to feel the pulse of cities. In this paper, we present a machine learning based approach to detect home locations at the population level based only on geo-tagged tweets and use the estimated home locations to investigate these crucial aspects of urban life.

Indeed, any given dataset may carry certain biases and our Twitter dataset is no exception. In fact, the average sampling rate of the U.S. census in each state is about 3% (https://www.census.gov/acs/www/), which is similar to the percentage of users we covered out of all Twitter users. In addition, younger people and minorities are disproportionably present on Twitter as compared to the overall makeup of the

‡ Corresponding author
* Project supported by the Goergen Institute for Data Science, New York State and the Xerox Foundation
ORCID: Tian-ran HU, http://orcid.org/0000-0003-0086-2447

population (http://pewresearch.org/pubs/2007/twitter-users-cell-phone-2011-demographics). Nevertheless, prior work and our results demonstrate that tweets can provide powerful and fine-grained cues of what is going on in cities. There is a trend that more and more users are 'active' on Twitter and this would help alleviate the biases. To the best of our knowledge, this is the first time that urban life has been studied on such open source data at a fine granularity.

The practicability of a home detection method for urban studies depends on two metrics. The first is granularity, which indicates in what resolution a method can predict one's home; the second is applicable scope, which measures the ratio of population that a method is applicable to. In some prior work, granularity is also called 'resolution' (Lin *et al.*, 2012). In this paper, we use these two terms interchangeably. To look deep into city life, an acceptable method should not only precisely determine one's home, but also cover as many people as possible.

Significant work has been done to find where people live based on a wide variety of data sources, such as Global Positioning System (GPS) data (Hoh *et al.*, 2006; Krumm, 2007; Sadilek and Krumm, 2012), cellphone recording data (Cho *et al.*, 2011), and geo-tagged social network data (Scellato *et al.*, 2011a; Pontes *et al.*, 2012a; 2012b). It has been shown that, as an important stop in people's daily movements, home exhibits certain intrinsic characteristics. For example, it is often the place which corresponds to the most check-ins of a user's trace, and also is probably the daily final destination of a user. However, previous work suffers from either coarse granularity or low applicable scope. High quality data such as continuous GPS data and diary data are required to reduce the possible range of one's home location. Krumm (2007) reported that home can be located with a median error of about 60 m using GPS traces of vehicles. However, the difficulties in collecting GPS data lead to the low applicable scope of these types of methods. The wide adoption of social media can help us overcome the low applicable scope problem, but much of the existing work could locate home only at a low resolution (city level, state level, or even time zone level) based on social media data (Mahmud *et al.*, 2012).

In this paper, we investigate ways to balance granularity and applicable scope. In most previous work based on geo-tagged data, home was simply estimated as the place visited most frequently (most check-in place) (Cho *et al.*, 2011; Scellato *et al.*, 2011a; Pontes *et al.*, 2012a; 2012b). We will show that this method does not always work especially when a user visits several places with similar frequencies. In contrast, we extract the features of one's trajectory in terms of temporal, spatial, and other aspects from a Twitter user's sparse trace of locations based on the geo-tagged tweets. A machine learning method is employed to capture the inherent properties of home using these mobility features, and further detect one's precise home location. We evaluate our method on two large Twitter datasets from the Greater New York City (NYC) Area (Fig. 1) and the Bay Area, and the results show that our method is capable of locating homes within a 100 m×100 m square with an accuracy of 70% and applicable to 76% and 71% active Twitter users in NYC and the Bay Area, respectively. An active Twitter user is defined as one who sent at least five geo-tagged tweets using the same definition as in Song *et al.* (2010), Lin

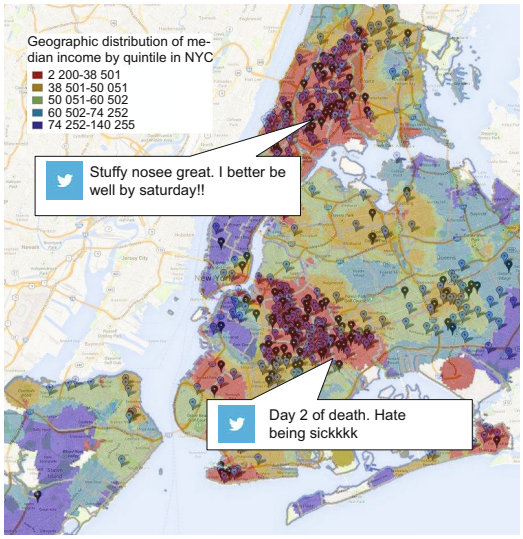

Fig. 1 Visualization of the health conditions and income levels of sample Twitter users in New York City. Two examples of 'sick tweets' and their origins are plotted in the figure. The income levels of zip code aligned areas are represented by a color scale. Blue colors indicate higher income levels and orange colors indicate lower income levels. This study explores to what extent online social media can be used to locate where people live, and predict the impact of a number of environmental and community demographic factors on health. References to color refer to the online version of this figure

et al. (2012), and Smith *et al.* (2014).

As another major contribution of this study, we demonstrate that highly precise and applicable scope home location estimation provides us the ability of looking deep into people's movements and habits. In particular, we study human mobility patterns in NYC. Furthermore, using the rich text content within tweets, we explore the health conditions of people in different zip code districts. Note that the zip code district in NYC has an average area of 3.6 km^2, and the radii of many of them are less than 1 km. Therefore, finer granularity methods are required. As we will show, our results correlate well ($r = 0.473$, p-value=0.006) with the data from the New York City Department of Health and Mental Hygiene (NYC DOHMH). We take a step further to study the correlation between the health conditions and various demographics of a zip code district. Consistent with previous social science studies, our results show that education level is the most important factor for health. This is the first time such analysis has been done in such a fine-grained fashion using social media data. Moreover, we discover the popular topics of the residents of different districts in the Bay Area. It is interesting to observe the distinctive lifestyles implied by the unique topics of certain districts. Not surprisingly, such distinctive signals of life style disappear when we extract topics based only on where the tweets come from.

2 Related work

2.1 Home location based user behavior understanding

Home location is crucial for modeling human mobility patterns. With the knowledge of home locations, we can gain a better insight into mobility patterns, as well as lifestyle in general. In Cho *et al.* (2011) and Scellato *et al.* (2011a; 2011b), home location is the key origin to calculate the distance that people travel and estimate the distance between social network users in a pairwise fashion. Researchers found that home location, as a crucial personal location, can be inferred from the information user posted online at a certain granularity (Krumm, 2007; Pontes *et al.*, 2012a; 2012b). Home location was also used to model individuals' living conditions and lifestyles by Sadilek and Kautz (2013).

2.2 Home location detection

2.2.1 Using social network data

Mahmud *et al.* (2012) used Twitter data attached with geo-information, especially tweet content, to infer the home locations at city, state, and time zone levels. In their study, the accuracies were 58%, 66%, 78% at city level, state level, and time zone level, respectively. Pontes *et al.* (2012a) developed 'single-attribute' models based on different social network features, for example, taking the value of users' 'employment' as their home city in Google+. They inferred the user home city using geo-tagged data from Foursquare, Google+, and Twitter with 67%, 72%, and 82% accuracy, respectively. Pontes *et al.* (2012b) also used geo-tagged social network data (Foursquare) to infer the home city within 50 km. A content-based method was used by Cheng *et al.* (2010) to detect Twitter users' home cities. They could place 51% of active Twitter users within 100 miles of their actual home locations. Cho *et al.* (2011) used a dataset containing the traces of two million mobile phone users from a European country to estimate the home locations according to the most visited places. They reported that by manual checking, the most check-in method can achieve 85% accuracy when they divided the area into 25 km×25 km cells. Scellato *et al.* (2011a) simply assigned the most check-in places as users' home locations, but they did not provide experiments to verify the accuracy of their method. The location of a person's home was estimated by fitting a two-dimensional Gaussian to all his/her locations between 1 A.M. and 6 A.M. (Sadilek and Kautz, 2013). The mean of this Gaussian was taken as the most likely home location. In summary, most of the home detection methods based on social media data require geo-tagged information. Although the accuracies reported in the above studies are reasonable, the granularity levels are so coarse that these methods are not applicable to district or finer level study.

2.2.2 Using GPS and diary data

GPS and diary data are much more dense and continuous than social network location data, which makes home detection more precise and easier.

However, such data are more difficult to obtain. Most of the work using GPS data suffered from the small number of users. In Krumm (2007), devices that recorded location every several seconds were installed on 172 subjects' vehicles. These devices collected the mobility data when the cars were moving. The ground truth of home location was filled by the drivers themselves. The author used four heuristic algorithms to compute the coordinates of each subject's home, and found that the best one is the 'last destination of a day'. We also include this feature in our extracted mobility features. The median distance error of the author's best algorithm was 60.7 m. Hoh *et al.* (2006) clustered the GPS traces of users agglomeratively until the clusters reached an average size of 100 m. Next, they eliminated clusters with no recorded points between 4 P.M. and midnight and clusters falling outside the residential areas by manual checking.

Semantically labeling places is another important topic related to home location detection. Sadilek and Krumm (2012) used GPS data from 307 people and 396 vehicles. They divided the world into 400 m×400 m squares, and assigned each GPS reading to the nearest cell. They found that the top 10 frequently visited locations usually can be semantically labeled as 'home', 'work', 'favorite restaurant', and so on. Krumm and Rouhana (2013) performed experiments using two diary datasets from the American Time Use Survey (ATUS) and the Puget Sound Regional Council (PSRC)'s Household Activity Survey, in which each location has a semantic label such as 'home' or 'school'. They extracted several features of a location and trained place classifiers using machine learning. They reported a classification accuracy above 90% on locations labeled as 'home'.

In summary, existing home detection methods suffer from either low applicable scope (GPS data and diary data) or coarse granularity problems (geotagged social media data), making them inadequate for urban computing problems that require both high applicable scope and fine granularity. Most check-in has been widely used to estimate user homes, and proved valid at a coarse granularity. However, its effectiveness is still doubtable when employed at finer granularity. We will show that most check-in does not always work for a more precise estimation. Our first contribution is in devising a method that can predict user home locations with a high accuracy using comprehensive features derived from sparse tweets. We also demonstrate the effectiveness of precise home location in representative applications in urban computing.

3 Methodology

3.1 Dataset and pre-processing

We collected all the geo-tagged tweets sent from the Greater New York City Area during July 2012 and also those sent from the Bay Area during the summer of 2013 through a vendor. A typical geotagged tweet contains the identity (ID) of the poster, the exact coordinates from where the tweet was sent, time stamp, and the text content. Due to the inherent noise in the geotags, we split the areas into 100 m×100 m squares and treat the center of each square as the target of home detection. We assign each tweet to its closest square, and each time a user tweet appeared in a square we say he/she had a check-in in this square. Therefore, the granularity of our square based home detection is around 70 m ($\sqrt{2} \times 100/2 \approx 70$). Similar to Song *et al.* (2010), Lin *et al.* (2012), and Smith *et al.* (2014), we focus only on those 'active users' who have sent at least five tweets. Also, following these studies, we use user's hourly traces (take only one location for each hour in our sampling duration) instead of taking account of every single check-in. If a user's location was not observed in an hour, the location for the corresponding hour is set to 'Null'; on the other hand, if a user appeared in several unique squares in an hour, we take the square with the highest number of check-ins as the location of this user in this hour. Typically, the hourly trace T_u of a user U looks like: $\boldsymbol{T} = [\text{Null}, L_i, \text{Null}, ..., L_j]$, where L_i does not have to be different from L_j. The lengths of the hourly traces of all users are the same, equaling the number of hours of our sampling period. We provide a snapshot of our dataset in Table 1.

3.2 Ground truth

The challenge is, without being told by the actual Twitter users, to know which is a person's home among all the places he/she visited. Almost all

Table 1 Statistics of our dataset

Item	NYC	Bay Area
Number of tweets	2 636 437	3 633 712
Total number of active users	55 237	53 314
Number of tweets labeled by AMT	5 000	5 000
Number of home locations (GT)	1 063	987

AMT: Amazon Mechanical Turk; GT: ground truth; NYC: New York City

previous work had the problem of obtaining a fine granularity ground truth. To obtain the ground truth of where people actually live, researchers relied on the information from user profiles (Mahmud *et al.*, 2012; Pontes *et al.*, 2012a; 2012b) or manually inspected the detection results (Cho *et al.*, 2011). Apparently, the location information in user profiles is coarse (at city level), while manual inspection is not scalable.

In this study, we rely on tweet content and human intelligence. For some tweets, a human can easily tell from where it was sent. For example, if a tweet said "The view from my office is awesome!" and included a picture of the view from a window, we can tell it was sent from a user's office. Some tweets are obviously sent from home. For example, "finally home!" or "home sweet home". This is the basis for us to design a mechanism to build the ground truth for home location. We polled some faithful Twitter users what they would like to post when at home. Based on their answers, we selected a set of keywords, each of which is likely to be mentioned in tweets sent from home. This set contains words like 'home', 'bath', 'sofa', 'TV', 'sleep', and so on. We ended up with a set of 50 unique words and their variants. Next, we used a simple keyword filter to obtain all the tweets that contain at least one of these keywords. From here we relied on human intelligence through crowdsourcing on Amazon Mechanical Turk (AMT) to find the 'home tweets', which were sent from home. We gathered these tweets into questionnaires. Each questionnaire contains five tweets, where we simply ask "Do you think these tweets are sent from home?" and the options include 'Yes', 'No', and 'Not sure'. We then posted these questionnaires to AMT. Each questionnaire was answered by three unique workers. To ensure the quality of the answers, we inserted several testing questions for which we knew the obvious answers. We found that workers have quite different thresholds on telling whether a tweet is sent from home. For example, a tweet said "The type of day where all I want is my bed and to not be at work!" and it received three different answers from three unique workers. Therefore, we retained only the tweets strictly for which all the three workers thought were sent from home. We checked these final tweets manually and found that virtually all of these tweets clearly indicated that the users were at home.

3.3 Models based on human mobility features

To study the inherent property of home, we extract several features of every unique location of one's hourly trace. In this section, we discuss these features in detail. Some of these temporal and spatial characteristics can be used as baseline methods to detect home location (e.g., check-in rate, PageRank score). We will show that although a single feature can be used to detect home location with a reasonable accuracy, it usually covers a limited amount of people. However, combining them appropriately using a machine learning method brings us significant gain in applicable scope.

3.3.1 Check-in rate

As we mentioned in Section 2, taking the place of most check-ins as home is a popular method. We call this method 'most check-in'. Due to the different tweet volumes of users, we do not use the absolute check-in amount. Although check-in based methods work well on GPS data (Krumm, 2007), it is not the case when it is employed on Twitter data. Unlike vehicle GPS devices which keep recording the location every several seconds, people tweet only when they feel like to do it. The place with most check-ins definitely is important to a user, but 'important' does not necessarily mean it is the home. We found that the effectiveness of most check-in is closely related to how much higher is the rate of check-ins of the most check-in place than the second most check-in place. Fig. 2 shows that the accuracy of most check-in increases linearly with the check-in rate margin. The accuracy is 70% only when the margin is significantly higher (50% or higher). Therefore, in addition to the check-in rate of a place, we include the margins of check-in rate between a place and its next higher and lower check-in places. Also, this is the

reason for the poor applicable scope of most check-in when high accuracy is required under our granularity setting (100 m×100 m squares). Fig. 3 is the cumulative distribution. It shows that only about 40% users have margins which are 50% or larger. The inset of Fig. 3 reports the distribution of check-in rate margin between the most check-in place and the second most check-in place.

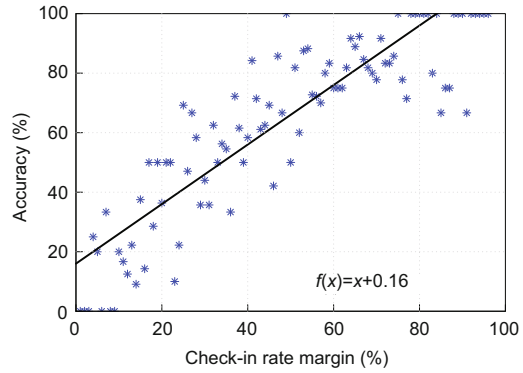

Fig. 2 Accuracy vs. the check-in rate margin between the most check-in place and the second most check-in place

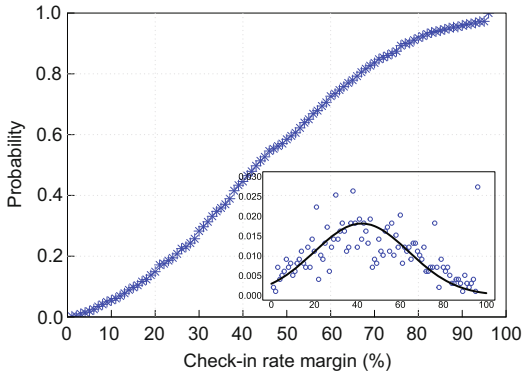

Fig. 3 Accumulated distribution of check-in rate margin. The inserted figure reports the distribution of check-in rate margin between the most check-in place and the second most check-in place. It is observed that the majority of margins are below 50%

3.3.2 Check-in rate during midnight

Intuitively, the places where people appear at midnight are probably their homes. Sadilek and Kautz (2013) took the places with the most check-ins during midnight (00:00-07:00) as people's homes. This method potentially alleviates the biases caused by other frequently visited places during daytime. Therefore, we take the check-in rate during midnight of a place as another feature, separated from the total check-in rate discussed above. However, as we will show later, it is not the case among active Twitter users. When an active user checks in at some place during midnight, this place is most probably not the home. This reflects the difference between Twitter data and GPS data. In Twitter data, a user has to be awake and active to report the current location at that moment, while in GPS data, the location recording is automatic even when the users are inactive. We will discuss this further later.

3.3.3 Last destination of a day

According to the research by Krumm (2007) on GPS data, the last destination of a user in a day (no later than 03:00 in the morning) is probably the home. It reveals that people's daily movements end at their homes. We include this as another mobility feature after minor modification. First, we extract all the final destinations of a user over the entire sampling period. We then sum up the number of days when a place had been the final destination. This value, the times of a place being the last stop of a day, is taken as one of the mobility features of this place.

3.3.4 Last destination with inactive midnight

Since 'last destination' might suffer from the check-ins sent from non-home places especially when the night was spent outside, we also consider a variant feature of last destination. We consider only tweets sent on the days when people are inactive during midnight (00:00-07:00). We exclude the days with active midnight and find the last destination among the remaining days, and then count the times of a place being the last destination.

The three features above introduce extra human behavior information into the original check-in feature. This helps reduce the applicable scope limitation of the simple check-in rate feature.

3.3.5 Spatial features

As the center of everyday activities of most users, home is one of the most important starting points and destinations of their movements. We use weighted PageRank (Xing and Ghorbani, 2004) and reversed PageRank scores to model the importance of a place being an origin point and a destination. To use PageRank related algorithms, we transfer one's trace into a directed graph. Vertices of the graph are

the locations visited. A directed edge from location L_i to L_j represents that L_j is visited directly after L_i. To quantify the certainty and importance of transitions between locations, we assign weight to each edge. The weight should be inversely proportional to the length of blank idle between two locations, and also proportional to the times of a transition appearing in one's trace. Formally, let $t(L_i, L_j)$ represent the transition between L_i and L_j and $w_{t(L_i,L_j)}$ represent the weight. The definition of the weight is as follows:

$$w_{t(L_i,L_j)} = \sum_{k\text{th } t(L_i,L_j) \text{ in } \boldsymbol{T}} \frac{1}{\text{Number of idle hours in the } k\text{th } t(L_i,L_j)}. \quad (1)$$

After constructing a user's movement graph, we apply the PageRank algorithm to calculate the importance of being a destination for each location in one's trace. Meanwhile, to study the importance of being an origin, we propose a reversed PageRank score. We reverse each edge's direction in the movement graph, with the weights of edges unchanged. The same reversed calculation is performed with the weighted PageRank algorithm. Compared with the earlier features, the PageRank score and reversed PageRank score describe the spatial characteristics of movements.

3.3.6 Temporal features

According to Krumm (2007), the probability of being at home varies over time. We extract the check-in rates of a place in different hours. These time-related features help us capture the property of home in terms of temporal patterns.

3.4 Multi-feature prediction

We believe that, as a distinctive place of one's trace, home permeates its influence into all the mobility features discussed above. Indeed, one can use a single feature to detect home location. However, a single feature captures only one type of characteristic, and thus will lead to low applicable scope of these methods. We apply a machine learning method to combine all the features. Because of the complementary effect between features, an appropriate combination will significantly increase the method's applicable scope without loss of accuracy.

Our goal is to distinguish home from other locations that one has visited. Since we obtain various features for every place of one's trace, the original problem can be transferred to an equivalent classification problem: given locations and corresponding feature values, we want to train a model to predict home among them. The inputs of the model are transactions identified uniquely by user ID and location ID, followed by features calculated from this user's hourly trace and a label as 'home' or 'non-home'. We use a linear support vector machine (SVM) model to exploit how these features are combined. Given the places and their features for a given user, the model outputs a score for each place. If the highest score exceeds a threshold, we take the corresponding place as the user's home. Otherwise, this user cannot be covered by our model. In Table 2, we present significantly positive and negative features and their weights. Not surprisingly, check-in rate, PageRank score, and reversed PageRank score related features are more significant than others. Note that all features contribute to the better overall applicable scope.

4 Home location evaluation

4.1 Applicable scope vs. accuracy

To guarantee the practicability of our home detection method, we need to balance granularity and applicable scope. Because of the natural trade-off between granularity and detection accuracy, we fix the

Table 2 Significantly positive and negative features and their weights

Positive feature	Weight	Negative feature	Weight
Check-in rate	2.03	Margin below next higher check-in	−0.30
Margin over the second highest check-in	0.19	Margin under next higher PageRank	−0.28
PageRank score	0.19	Margin under next higher reversed PageRank	−0.21
Last destination on inactive midnight	0.12	Rank of reversed PageRank	−0.07
Reversed PageRank score	0.09	Rank of PageRank	−0.07

granularity as 100 m×100 m squares and explore the relationship between accuracy and applicable scope. The accuracy of each single feature can be adjusted through the threshold, which also affects applicable scope. In this section, on both NYC and Bay Area data we compare our method with three other intuitive single-feature based methods: (1) most check-in (Due to the statistical insignificance of too few check-ins, we also set a constraint on this method: the absolute check-in number of the most check-in place is at least 3); (2) highest PageRank score (Similarly, the threshold is how much higher the highest PageRank score compared with the second highest one); (3) highest reversed PageRank score. Figs. 4 and 5 indicate the trend of applicable scope along with accuracy. Applicable scope decreases rapidly as the accuracy rises. It shows that, at every accuracy level, our method covers more users. Specifically, when we set the accuracy of each method at 70% (which, we think, is acceptable for urban computing), our method obtains 76% and 71% applicable scope in NYC and Bay Area, respectively. As to other methods, none is able to detect home for more than 50% users when the accuracy is 70% or higher. It proves that an intelligent combination model leads to a significant increase in applicable scope. Most check-in works better than PageRank score and reversed PageRank score, but still suffers from low applicable scope. The higher applicable scope of our method implies the complementary effect between these mobility features. This method dose not depend on any one type of information but combines the information of several aspects. For example, when a user's check-in features do not provide enough cues to predict the home, other types of features may pick up the slack and naturally lead to a higher applicable scope. The balance of applicable scope and accuracy facilitates more extensive and deep urban life studies, which we will describe in the following subsection.

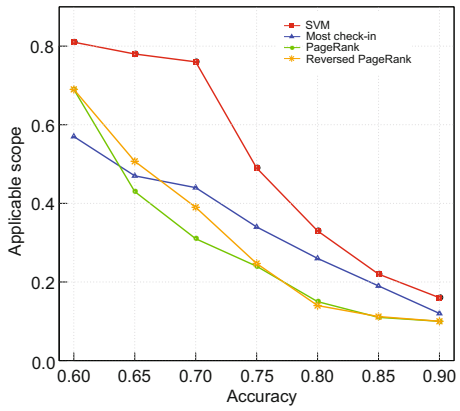

Fig. 4 The applicable scope and accuracies of different methods on the NYC dataset

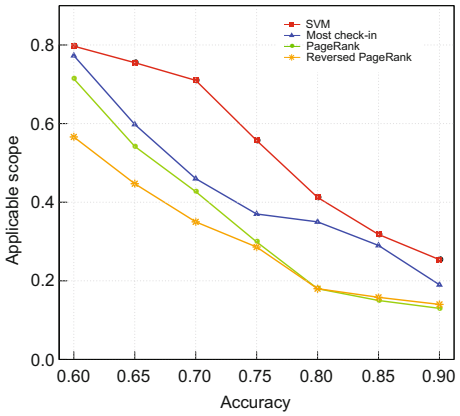

Fig. 5 The applicable scope and accuracies of different methods on the Bay Area dataset

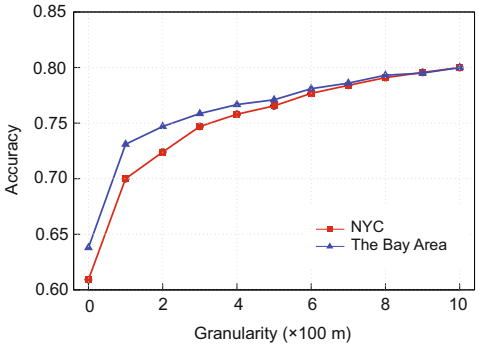

Fig. 6 Granularity vs. accuracy on NYC and Bay Area datasets

4.2 Resolution of home detection

The detection of home is performed in 100 m×100 m squares, and thus the granularity of the method is around 70 m. Different urban studies require different resolutions. To explore this, we fix the applicable scope of our method at 80%, and vary the granularity from 100 m to 1000 m to evaluate the change in accuracy of our method. Fig. 6 indicates that the accuracy of our method increases when the granularity increases. As the granularity becomes coarser, the increase in the accuracy slows down,

and peaks at around 80% when the granularity is about 1000 m. Compared with Pontes *et al.* (2012a), our method provides higher granularity with similar accuracy (around 80%).

In summary, our method achieves a desirable balance between resolution and applicable scope, and for geographically and demographically diverse regions in U.S. Consequently, it will prove instrumental for higher-level urban informatics. Next, we will show that such precise home location information enables us to study mobility patterns and health conditions and reveals significant insights that eluded earlier studies on similar subjects.

5 Mobility patterns in New York City

Human mobility pattern is one of the most studied subjects in urban computing research. However, without knowing precise locations of people's homes, most work is limited to a coarse granularity (Cheng *et al.*, 2010; Mahmud *et al.*, 2012; Pontes *et al.*, 2012a; 2012b). We overcome this limitation and dig deeper on people's mobility patterns, daily habits, and even their check-in patterns at home. Due to the space limit, we report only the results for NYC.

5.1 Active user mobility

Knowing the precise home location of a user, we can calculate the distance from each check-in to the home. Note that when a user checks in somewhere this user must be active at that time. Self-reported location is one of the most important features of geo-tagged Twitter data. When the user is not posting locations, though one may infer locations given user's historic data (Ashbrook and Starner, 2003; Backstrom *et al.*, 2010; Cranshaw *et al.*, 2010), such inference cannot obtain precise knowledge of their locations. We report our findings in Fig. 7. There are roughly two peaks in Fig. 7, one at midnight and the other at mid-day. As we discussed above, though surprising at first glance, the check-ins at midnight are most likely not at home. In contrast, active Twitter users are actually quite far from their homes late at night. At 01:00 A.M., the average distance from home is at the minimum value. We can also see that, on Saturday people move further away from home than on any other day of the week. This tendency continues to early Sunday morning. Also, on late Friday (after 04:00 P.M.) people tend to move further than on other workdays.

5.2 Likelihood of returning home

We also model the probability of returning home. We find that people return home in a daily pattern. In other words, the probability of returning home increases significantly every 24 h (Fig. 8). Cheng *et al.* (2011) and Pontes *et al.* (2012b) did similar work. They modeled the probability of returning to any place (not limited to home), and found that in addition to a daily pattern, revisiting also follows a weekly pattern. In other words, the probability of revisit also increases every 168 h (one week). However, in this study we show that when the place is limited to 'home', the weekly pattern disappears. This makes obvious sense as people normally return to their homes every day but not many return to their homes every week. The probability of returning home decreases over longer periods (e.g., 48 h) because this scenario is naturally less frequent.

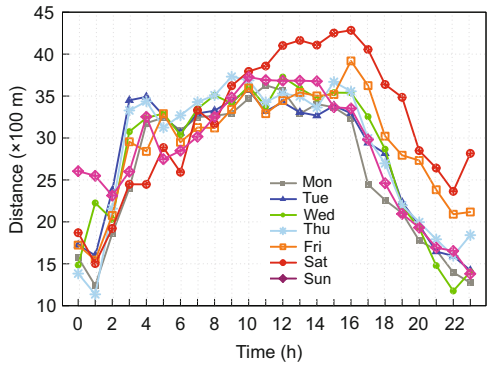

Fig. 7 The average distance of active users from home in every hour of a week

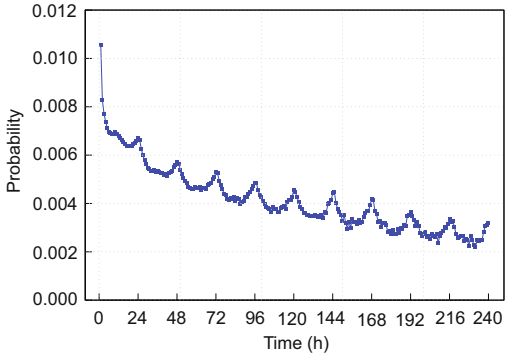

Fig. 8 The probability of returning home over time

5.3 Likelihood of being at home

We also estimate the probability of a user being at home (Fig. 9). The daily pattern from our result is similar to that of Krumm (2007), which was obtained by recording the continuous GPS trace of people and calculation based on self-reported home locations. This proves the effectiveness of our machine learning based home location method on sparse and noisy data. To take a deeper look into people's mobility patterns, we also model the probability of being at home over a week. The differences between our results and those in Krumm (2007) are: (1) The probabilities are much lower after 03:00 A.M. in our results, as we discussed above, showing that 'being away from home at midnight' is an interesting characteristic of active Twitter users; (2) The probabilities of being at home during the daytime are slightly higher in our study. This is probably because some rarely tweeting users cannot be covered in our method compared with studies using continuous GPS data. We also observe the difference between weekdays and weekends: the probability of being at home starts decreasing from Friday nights and keeps decreasing on Saturday nights, suggesting that people are more likely at home during the night on other days in the week.

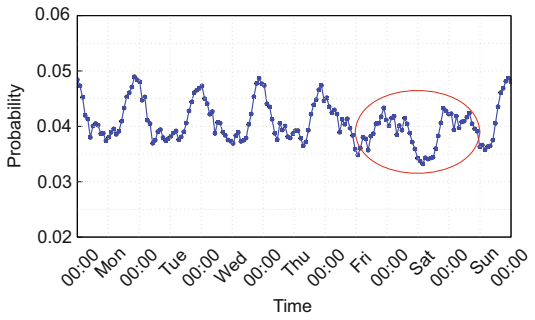

Fig. 9 The probabilities of being at home. The red circle emphasizes the decrease of the probability of being at home on Friday and Saturday nights. References to color refer to the online version of this figure

6 Buzz by home location

Previous studies on topics of conversation on Twitter are mostly based on the geolocation from where the tweets are posted (http://projects.knightlab.com/projects/neighborhood-buzz). In this paper, we call the topics of conversation based on geolocation 'buzz by location'. Clearly, tweets sent from a region do not necessarily have to be sent by people who live there. We call the topics of conversation of local residents 'buzz by home'. What intrigues us is, are these two kinds of buzz the same? According to our experiments, the answer is negative. To find out the difference, we group tweets from the Bay Area by the cities from which they were tweeted and also by the estimated home city of their posters, respectively. Note that 'cities' in the Bay Area are some small areas, and the radii of some of them are only around 1 km (e.g., Emeryville). Therefore, to find buzz by home in such areas, fine-grained home detection methods are required. By extracting the topics following hashtags, we gather topics from an area. We are interested only in the topics that are distinctive to a certain city, and thus we filter out topics appearing in more than three cities. We find notable differences between buzz by location and buzz by home. Due to the limit of space, we report only the most interesting ones in Table 3. Buzz by home is more life-like than buzz by location. In several cities, '#healthcare' and '#hospitality' are widely discussed by people living there, and some topics are closely related to local population, such as '#qa' (referring to quality assurance) in Mountain View and '#webdesign' in San Mateo. Buzz by home is also related to local events; e.g., '#backbenoit' in Emeryville refers to a famous baseball game in that summer held near Emeryville. On the other hand, buzz by location is more monotonic. Most topics are just the names of the cities where tweets are sent.

7 Community health condition estimation

With the precise knowledge of where people live, we are further interested in the relationship between people's health conditions and their home locations. Let us start with how we model the health conditions of Twitter users.

7.1 Inferring health state

We select prediction features upon previous work on classification of short text messages (Culotta, 2010; Paul and Dredze, 2011; Sadilek *et al.*, 2012) and learn an SVM classifier C_s (the subscript

Table 3 Buzz by home and buzz by location in different cities of the Bay Area

City name	Home buzz	Location buzz
Mountain View	#qa	#mountainview
		#followmeskip
		#skipfollowme
Fremont	#hospitality	#fremont
San Mateo	#webdesign	#sanmateo
Emeryville	#backbenoit	#emeryville
	#emeryville	
Sunnyvale	#engineering	#sunnyvale
	#healthcare	#apocono
	#apocono	
Santa Clara	#marketing	#santaclara
San Leandro	#love	#sanleandro
	#fashion	
Redwood	#manufacturing	#redwoodcity
	#geekcamp	#tryhard
	#tryhard	#scifi
Hayward	#healthcare	#hayward
	#accounting	#green
	#letsgooakland	

's' indicates sickness to differentiate it from the earlier SVM used for home location) which identifies tweets that indicate their author is ill. C_{s} is trained by directly optimizing the area under the receiver operating characteristic (ROC) curve. It is robust even in the presence of strong class imbalance, where for every health-related message there are more than 1000 irrelevant ones. We use C_{s} to distinguish between tweets indicating the author is affected by an ailment (we call such tweets 'sick' tweets) and all other tweets (called 'other' or 'normal' tweets). For SVM features, we use all unigram, bigram, and trigram word tokens that appear in the training data. For example, 'so sick of' is represented by the feature vector $(\text{so, sick, of, so sick, sick of, so sick of})^{\mathrm{T}}$. As a result, our SVM operates in more than 1.7 million dimensions, where each dimension represents a word or a phrase extracted from the training data. Before tokenization, we convert all text to lower case, strip punctuation, and special characters, and remove mentions of user names (the '@' tag) and retweets (analogous to email forwarding). However, we do keep hashtags (such as '#sick'), as those are often relevant to the author's health state, and are particularly useful for disambiguation of short or ill-formed messages. Table 4 lists examples of significant features found in the process of learning C_{s}. We use the SVM cascade learning procedure described in Sadilek *et al.* (2012). The evaluation of C_{s} on a held-out set shows 0.98 precision and 0.97 recall with respect to labels agreed upon by human annotators. The ground truth for each tweet is obtained by asking AMT workers to label the tweet as either 'sick' or 'other' and subsequently extracting the majority vote.

Table 4 Examples of positively and negatively weighted significant features of our SVM model C_{s}

Positive	Weight	Negative	Weight
Sick	0.9579	Sick of	−0.4005
Headache	0.5249	You	−0.3662
Flu	0.5051	Lol	−0.3017
Fever	0.3870	Love	−0.1750
Coughing	0.2910	So sick of	−0.0800
Being sick	0.1910	Bieber fever	−0.1000
Better	0.1980	Smoking	−0.0980
Being	0.1940	I'm sick of	−0.0890
Stomach	0.1700	Pressure	−0.0830
Infection	0.1680	I love	−0.0710

7.2 Zip code district health condition

7.2.1 Benchmarking with DOHMH data

We define 'sickness score' of a district as the percentage of 'sick' people who live in it. Therefore, the higher the sickness score is, the worse the health condition of this district is. To evaluate our home location method and our health inference model, we compare our sickness score with the data from NYC DOHMH. In the dataset provided by DOHMH, NYC is divided into 34 areas. The health condition of individuals has four levels, 'excellent', 'very good', 'good', and 'fairly good'. DOHMH provides the percentage of each level of every area. We calculate the correlation between our sickness score and the percentages of 'excellent', 'very good', 'good' levels (since the sum of four percentages is one, there is no need to calculate the correlation for all), respectively. Our sickness score is highly negatively correlated with the 'excellent' percentage ($r = -0.383$, p-value=0.030), and positively correlated with the 'good' percentage of each area (r=0.473, p-value=0.006). This makes sense because our health state inference method is based on the percentage of sick Twitter users, and thus it indicates a rough degree of relatively unhealthy people in an area. Intuitively, this degree should be negatively correlated with the percentage of people whose health is in 'excellent' condition. Because there are

only four levels in the health survey and the 'good' level is the second worst level among the four levels, we consider this level as a 'relatively unhealthy' metric. Therefore, the positive correlation between our sickness score and this level makes sense. Although our method can provide the highest applicable scope among all the methods, it still suffers from the problem of small numbers of samples in some districts. For better data reliability, we exclude all the districts in which there are fewer than 200 residents detected by our method. Table 5 shows the correlations with 'excellent' ($r = -0.569$, p-value=0.017) and 'good' (r=0.601, p-value=0.010).

7.2.2 Factors that affect health

The United States has one of the world's largest health inequalities across its society, where the gap in life expectancy of the most and the least advantaged segments of the population is over 20 years. It has been reported that this difference is partly due to a difference in social status, but many aspects of the phenomenon remain unexplained (Sapolsky, 2004). What we are interested in here is what factors (such as the poverty level, education level, and race percentage) of a community may affect residents' health condition most? The data provided by DOHMH help us little on this as the whole city is divided into 34 areas spatially, each containing several zip code districts. However, there are about 170 zip code districts in NYC and the conditions of the zip code districts within an area may vary dramatically. The granularity of such division is too coarse for accurate analysis about the relationship between health and factors we are interested in. This is why we again need precise home locations. The average NYC zip code zone has an area of 3.6 km^2 and can be walked across in under 20 min. The zip code areas are as shown in Fig. 1. We can now associate each person with the context derived from the 2010 census, the most recent census available. We focus on three broad characteristics of a person's neighborhood: poverty, education, and race. Poverty is measured in terms of the fraction of families and individuals below the poverty line, the number of abandoned housing units, and the prevalence of social security dependence. Education captures the proportion of people over 25 years old with various levels of education (from elementary school to a doctorate). The race factor includes the proportion of different races and ethnic groups. We first use our method to accurately put each resident into the right zip code district. With the census data telling us the income and poverty level of each district, we plot Fig. 10, which lists the correlations between the factors of race, education, poverty, income, and the sickness score. Consistent with Winkleby *et al.* (1992), education is the best predictor for good health. Among our results, high-level education percentage of a district strongly negatively correlates with the sickness score of that district. On the other hand, the low-level education percentage strongly positively correlates with the sickness score. We also model the correlation between heath and factors including poverty, race, and income. It

Table 5 The sickness score and percentages of people whose health is in 'excellent' or 'good' condition

Area name	Sickness score	Percentage (%)	
		Excellent	Good
Upper West	0.046	30.7	26.3
Chelsea	0.018	29.3	20.0
Gramercy	0.029	26.4	19.8
Flatbush	0.100	23.6	30.5
Central Harlem	0.062	23.2	23.9
Lower Man.	0.019	23.2	22.0
Southeast Q.	0.043	22.1	27.6
Astoria	0.042	21.0	35.1
Crown Heights	0.066	20.8	34.7
Heights/Slope	0.061	20.5	27.1
Inwood	0.049	20.2	38.8
Bushwick	0.070	20.0	30.5
Southwest Q.	0.084	16.9	33.5
South Bronx	0.050	13.2	38.6
Fordham	0.083	13.0	46.1
Pelham	0.085	10.5	38.7
Correlation	NA	−0.569	**0.601**

The bottom row shows the correlation between our score and the ratio. The table is sorted by the percentage of people whose health is in 'excellent' condition in descending order

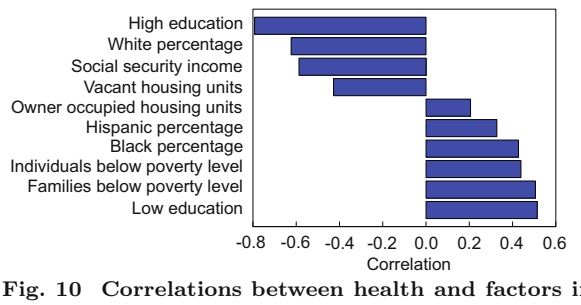

Fig. 10 Correlations between health and factors including poverty level, education, race, and income

is shown that the poverty level is negatively related to health condition, which is supported by Winkleby *et al.* (1992).

It is worth noting that in the earlier study in Sadilek *et al.* (2012), little correlation was found between the health conditions and factors such as income and education levels. We believe that the lack of signals in that study was primarily due to the lack of precision on home location estimation for Twitter users as other aspects of the analyses are equivalent between the two studies. Therefore, precise home location is instrumental for such studies in urban computing.

8 Conclusions and future work

In this paper, we propose a machine learning based multi-feature method which can precisely locate people's home locations. Compared with previous work, we do not require continuous GPS trace data but instead use noisy and sparse Twitter data. By evaluation on the ground truth obtained by human annotation, our method has achieved 76% and 71% applicable scope on the Twitter data that we have collected from New York City and the Bay Area, respectively. To the best of our knowledge, this is the first time that urban life has been studied on such open source data at a fine granularity.

Our method achieves a desirable balance between resolution and applicable scope, and for geographically and demographically diverse regions in U.S. With such a balance, we are able to study human mobility patterns to an extent that was not feasible before. We have shown interesting mobility patterns extracted from Twitter data. Furthermore, we use a health state model to estimate Twitter users' health condition. We relate people's health to their home locations and compare our estimated sickness scores with data from NYC DOHMH. Highly correlated results have validated the effectiveness of both our home location estimation method and our health state inference model. More importantly, with the precise location information and accurate estimation of people's health conditions, we have found correlation between the average health state and community demographics (such as income, poverty level, education level, and race distribution) of different districts in the city. Our findings match well with those of independent studies based on census data. We also discover the true buzz in communities where people live. Overall, we believe our precise home location detection method provides a crucial piece of context that enables helpful and pervasive applications.

In the future, we will investigate other domains of interest in urban computing given the knowledge of home locations, such as economic activities, resource consumption, urban planing, and emergency management. It will also be interesting to extend our method for home location to general place recognition based on user movement behaviors.

References

Ashbrook, D., Starner, T., 2003. Using GPS to learn significant locations and predict movement across multiple users. *Pers. Ubiq. Comput.*, **7**(5):275-286. http://dx.doi.org/10.1007/s00779-003-0240-0

Backstrom, L., Sun, E., Marlow, C., 2010. Find me if you can: improving geographical prediction with social and spatial proximity. Proc. 19th Int. Conf. on World Wide Web, p.61-70. http://dx.doi.org/10.1145/1772690.1772698

Cheng, Z., Caverlee, J., Lee, K., 2010. You are where you tweet: a content-based approach to geo-locating twitter users. Proc. 19th ACM Int. Conf. on Information and Knowledge Management, p.759-768. http://dx.doi.org/10.1145/1871437.1871535

Cheng, Z., Caverlee, J., Lee, K., *et al.*, 2011. Exploring millions of footprints in location sharing services. Proc. 5th Int. AAAI Conf. on Weblogs and Social Media, p.81-88.

Cho, E., Myers, S.A., Leskovec, J., 2011. Friendship and mobility: user movement in location-based social networks. Proc. 17th ACM SIGKDD Int. Conf. on Knowledge Discovery and Data Mining, p.1082-1090. http://dx.doi.org/10.1145/2020408.2020579

Cranshaw, J., Toch, E., Hong, J., *et al.*, 2010. Bridging the gap between physical location and online social networks. Proc. 12th ACM Int. Conf. on Ubiquitous Computing, p.119-128. http://dx.doi.org/10.1145/1864349.1864380

Culotta, A., 2010. Towards detecting influenza epidemics by analyzing Twitter messages. Proc. 1st Workshop on Social Media Analytics, p.115-122. http://dx.doi.org/10.1145/1964858.1964874

Hoh, B., Gruteser, M., Xiong, H., *et al.*, 2006. Enhancing security and privacy in traffic-monitoring systems. *IEEE Perv. Comput.*, **5**(4):38-46. http://dx.doi.org/10.1109/MPRV.2006.69

Krumm, J., 2007. Inference attacks on location tracks. Proc. 5th Int. Conf. on Pervasive Computing, p.127-143. http://dx.doi.org/10.1007/978-3-540-72037-9_8

Krumm, J., Rouhana, D., 2013. Placer: semantic place labels from diary data. Proc. ACM Int. Joint Conf. on Pervasive and Ubiquitous Computing, p.163-172. http://dx.doi.org/10.1145/2493432.2493504

Lin, M., Hsu, W., Lee, Z., 2012. Predictability of individuals' mobility with high-resolution positioning data. Proc. ACM Conf. on Ubiquitous Computing, p.381-390. http://dx.doi.org/10.1145/2370216.2370274

Mahmud, J., Nichols, J., Drews, C., 2012. Where is this tweet from? Inferring home locations of Twitter users. Proc. 6th Int. AAAI Conf. on Weblogs and Social Media, p.511-514.

Paul, M.J., Dredze, M., 2011. A Model for Mining Public Health Topics from Twitter. Technical Report, Johns Hopkins University, USA.

Pontes, T., Magno, G., Vasconcelos, M., *et al.*, 2012a. Beware of what you share: inferring home location in social networks. Proc. IEEE 12th Int. Conf. on Data Mining Workshops, p.571-578. http://dx.doi.org/10.1109/ICDMW.2012.106

Pontes, T., Vasconcelos, M., Almeida, J., *et al.*, 2012b. We know where you live: privacy characterization of Foursquare behavior. Proc. ACM Conf. on Ubiquitous Computing, p.898-905. http://dx.doi.org/10.1145/2370216.2370419

Sadilek, A., Krumm, J., 2012. Far out: predicting long-term human mobility. Proc. 26th AAAI Conf. on Artificial Intelligence, p.814-820.

Sadilek, A., Kautz, H., 2013. Modeling the impact of lifestyle on health at scale. Proc. 6th ACM Int. Conf. on Web Search and Data Mining, p.637-646. http://dx.doi.org/10.1145/2433396.2433476

Sadilek, A., Kautz, H., Silenzio, V., 2012. Modeling spread of disease from social interactions. Proc. 6th Int. AAAI Conf. on Weblogs and Social Media.

Sapolsky, R.M., 2004. Social status and health in humans and other animals. *Ann. Rev. Anthropol.*, **33**:393-418.

Scellato, S., Noulas, A., Lambiotte, R., *et al.*, 2011a. Socio-spatial properties of online location-based social networks. Proc. 5th Int. AAAI Conf. on Weblogs and Social Media, p.329-336.

Scellato, S., Noulas, A., Mascolo, C., 2011b. Exploiting place features in link prediction on location-based social networks. Proc. 17th ACM SIGKDD Int. Conf. on Knowledge Discovery and Data Mining, p.1046-1054. http://dx.doi.org/10.1145/2020408.2020575

Smith, G., Wieser, R., Goulding, J., *et al.*, 2014. A refined limit on the predictability of human mobility. Proc. IEEE Int. Conf. on Pervasive Computing and Communications, p.88-94. http://dx.doi.org/10.1109/PerCom.2014.6813948

Song, C., Qu, Z., Blumm, N., *et al.*, 2010. Limits of predictability in human mobility. *Science*, **327**(5968):1018-1021. http://dx.doi.org/10.1126/science.1177170

Winkleby, M.A., Jatulis, D.E., Frank, E., *et al.*, 1992. Socioeconomic status and health: how education, income, and occupation contribute to risk factors for cardiovascular disease. *Am. J. Public Health*, **82**(6):816-820. http://dx.doi.org/10.2105/AJPH.82.6.816

Xing, W., Ghorbani, A., 2004. Weighted pagerank algorithm. Proc. 2nd Annual Conf. on Communication Networks and Services Research, p.305-314. http://dx.doi.org/10.1109/DNSR.2004.1344743

Dr. Jie-bo LUO, corresponding author of this invited research article, joined the University of Rochester in Fall 2011 after over 15 years at Kodak Research Laboratories, where he was a senior principal scientist leading research and advanced development. He has been involved in numerous technical conferences, including serving as the program co-chair of ACM Multimedia 2010 and IEEE CVPR 2012. He is the Editor-in-Chief of *Journal of Multimedia*, and has served on the editorial boards of *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *IEEE Transactions on Multimedia*, *IEEE Transactions on Circuits and Systems for Video Technology*, *Pattern Recognition*, *Machine Vision and Applications*, and *Journal of Electronic Imaging*. He is a Fellow of the SPIE, IEEE, and IAPR.