Analyzing Human Mobility and Social Relationships from Cellular Network Data

Zheng Liu [1] , Yuanyuan Qiao [1]* , Siyan Tao [1] , Wenhui Lin [2] , Jie Yang [1]
[1] *Beijing Key Laboratory of Network System Architecture and Convergence, and Beijing Laboratory of Advanced Information Networks, BUPT, Beijing 100876, China*
[2] *Technology Research Institute, Aisino Corporation, Beijing 100876, China*

***Abstract*—Due to geographic and social constraints, human mobility shows a high degree of temporal and spatial regularity. More recently, emerging data tagged with geographical information can be used to study human mobility patterns. People usually spend most of their time at a few key important locations, such as home and work places. And for users with a certain social connection–co-workers or co-life, they often stay at the same locations. In this paper, firstly, we use an algorithm to identify important locations. After that, we find that the similarity between the two trajectories is closely related to their proximity in the location-based social network, where users having the same important locations are connected. In order to embody social contacts in mobility, for hourly variations in mobility similarity, we apply unsupervised clustering method to identify four categories of social ties. Finally, we further propose the unsupervised method and supervised method to predict which new links will develop in a social location-based network. We believe our finding can contribute to urban planning especially in areas of functional zone, transportation infrastructure deployment and mobile network facilities development.**

***Index Terms*—important places, location-based social network, user mobility, network proximity, link prediction**

I. Introduction

With the widespread use of mobile communication equipment and the rapid development of Internet technology, it becomes possible to acquire a variety of information at any time and places. According to the 39th "China Statistical Report on Internet Development" [1] reported in January 2017, the mobile Internet users maintained rapid growth in China. It is bound to produce a lot of information when people use mobile phones to exchange data. So it can be easy to capture the large volumes location data from mobile 2G/3G/4G data networks [2], Call Detail Records (CDRs) [3], [4], as well as Global Position System (GPS) tracks [5], [6]. And the research of user's network data makes price mobile and wireless services easily and more accurate.

Due to the identity and geography limit, people usually spend much of their time at some key locations [7]. For example, colleagues stay in the office on weekdays, the friends are tended to be together in the entertainment venues on the weekends, and family rest at home every night. Identifying these key places is thus central to understand human mobility and social pattern. To understand the interaction between social behavior and mobility, researchers began measuring the correlations between them. They found that social links are often driven by spatial proximity [8]. Understanding the social group's mobile pattern can offer solutions to plan urban functional area. And it also contributes to capture movements for those needing good estimates of travel demand in the city [9]. A deep understanding of the connections between us will help make the city we live more efficient and livable.

In this paper, firstly, in order to build the location-based social network, we identify individual important places from users' trajectories extracted from cellular network data. Secondly, we introduce the corresponding indicators to quantify social behavior and mobility. Then we find strong correlations between network proximity and mobility similarity and show that mobility similarity can be used to identify the social relationships. Finally, we also find that such correlation can effectively predict the new link in the location-based social network. Overall, the contributions of our work are as follows:

1) By analyzing the users' trajectory, we can build the offline social network - which represents the social networks formed by the users due to personal appearance in the same important locations. It reflects not only the user's mobile characteristics, but also the social behavior. Unlike online social networks, we identity social relationships by clustering hourly variations in mobility similarity. Especially, considering time and mobility variations, the clustering method can accurately identify the social groups to reconstruct users' mobility pattern.
2) In order to know how people are connected in the social network, we introduce several well-established measures of network proximity, based on the common neighbors in this network. And we adopt a series of co-location measures to quantify the mobility similarity. Then we find that the strong correlation exists between human mobility and network proximity, simultaneously, such correlation has considerable power to predict new links. Namely, these findings demonstrate that the social relationship that an individual maintains with other people in different groups can be predicted. It contributes to discover the interaction between social and mobility behavior.
3) Based on cellular network data, we use a clustering method to identify 793 active users' important loca-

*Corresponding author, email: yyqiao@bupt.edu.cn; First author, email: izheng@bupt.edu.cn

978-3-901882-98-2

tions in a region of the northern city in China. The data captured the more information with smaller time granularity comparing to the CDR. Since the cellular network exchanges data more frequently, it can be used to make a detailed description of people's daily mobile trajectory.

The rest of this paper is organized as follows. Section 2 presents some related works. And Section 3 describes the data we obtained as well as the measures we have taken to build the location-based social network. Section 4 calculates the correlation between network proximity and human mobility by introducing the related indexes. Section 5 embodies the social relationship in mobility. Section 6 gives a perspective for predicting the new link and Section 7 offers conclusions.

II. Related work

The recent availability of large-scale data sets, such as cell phone records, GPS data and cellular network data, have opened up new possibilities for studying pattern of human mobility. Gonzalez et al. [3] showed that human trajectories show a high degree of temporal and spatial regularity, and each individual is characterized by a time-independent travel distance and a significant probability to return to a few highly frequented locations, by analyzing cellphone records from an unnamed European country. The recent work [10] demonstrates different mobility patterns in two cities in the United States. Hightower et al. [11] and then Kim et al. [12] presented algorithms for discovering semantically meaningful places based on continuous GSM and WiFi data. Compared with the work presented in this paper, we focus on some particular locations when analyzing the human mobility to construct the offline location-based social network.

Krings, Lambiotte et al. [13], [14] show that two arbitrary people are likely connected by a short chain of intermediate friends and the probability that two customers are connected by a link follows a gravity model. One of the previous studies focused on geographical features and their impact on network topologies [15]. Consequently, we introduce some mobility similarity and network proximity measures to study the correlation exists between physical space and network structure. In this case, how human mobility and social behavior interact with each other can be addressed.

In the past years, previous studies contributed to finding various proximity metrics on network topologies, which can be used as factors to predict new link in the supervised [16], [17] or unsupervised frameworks [18]. However, these researches only consider the network proximity indicators, and Wang D et al. [19] began studying the interplay between mobility patterns and the structure of social ties to understand how individual mobility patterns shape and impact the social network. Based on previous researches, in order to focus on the impact of human mobility on link prediction, we design supervised and unsupervised prediction method combining with our mobility homogeneity and network proximity measures.

In this paper, compared with the previous researches, we first discover users' social relationships through the important positions, and next explore the correlation between the movement of these users and network characteristics to predict the new links. All of these can understand social relationships and discover social groups better based on user mobility.

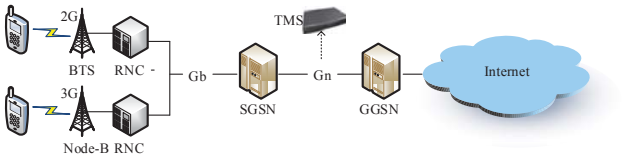

Fig. 1. The deployment of TMS.

III. Preliminaries

In this section, firstly we describe how the experimental data is collected, followed by a detailed description of the characteristics of the data. Secondly we extract the active users who appear more than 12 times in our mobile network data in hourly granularity every day, and finally we use these data to identify the important locations of active users.

A. Data Collection

The massive data we use is collected by the TMS (Traffic Monitoring System) deployed in the core node of a Chinese operator network. The deployment of TMS in the network is shown in Fig. 1.

The mobile terminals carried by mobile subscribers communicate with the base station, and the base station transmits the traffic data information to the RNC (Radio Network Controller). The RNC transmitter then passes the message to the SGSN (Service GPRS Support Node), and the SGSN forms a link channel Gn interface to finally transfer the data to the network when communicating with the GGSN (Gateway GPRS Support Node). To collect the data of the mobile terminal users, the traffic monitoring device TMS is deployed on the Gn interface. The data collected by TMS is resolved in the form of the flow records filled with abundant user access internet information through the DPI (Deep Packet Inspection) and DFI (Deep Flow Inspection) [20].

B. Data Description

As mentioned above, the data is the flow record which contains a variety of information, such as timestamp, user ID (in order to protect user's privacy, we cast the user phone number into user ID), LAC (Location Area Code), CI (Community Identity), etc. And the LAC and CI can identify a unique base station, then every user can be represented by 4-tuple $\{$time, ID, LAC, CI$\}$. In order to ensure the reliability of the analysis data, it is necessary to filter some noise data which is illegal or incomplete. In this paper, our data set obtains 794 active users and 770 base stations from one northern city in China. In our experiment, we only focus on active users due to following reasons: on the one hand, TMS runs all time without stopping to capture the data in the network, accompanied by

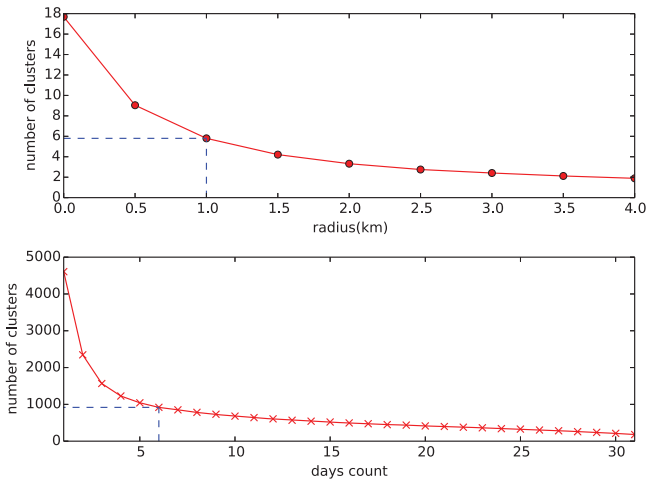

Fig. 2. The number of clusters changing with different parameter values.

zombies traffic in the background. So analysis of these traffic makes no sense. On the other hand, only considering active user of a region can effectively discover the social relationships between them. And our data sets can be divided into two parts, one is the sample data set for constructing the location-based social network (from August 1st to September 9th in 2015), the other part is the testing data set for the new network (from September 10th to 20th).

C. Identify Important Locations

To construct the location-based social network, the primary work is to identify users' important locations. So we apply a systemic clustering method called Hartigan [7]. The method consists of the following two steps: in the first step, for each user, we count the number of days they have access to each base station. And these base station are arranged in descending order based on this number. The second step is that we consider the first base station in the sorted list as the cluster center, for each subsequent base station, it is added to the existing cluster if the distance between them is less than the set threshold. Otherwise, the base station becomes a new cluster center. Therefore, we can obtain the base station category by combining the time and distance factors. In order to determine the threshold of clustering radius and ensure which base station is the important location of users, we observe the change of the clusters' number with different parameter values. And the appropriate value is the "elbow value" of the curve.

As can be seen from Fig. 2, the optimal clustering radius threshold is 1km, and we regard the locations where users appear for more than 6 days as important places in our clustering result. Averagely, users spend nearly 17 hours at above important places cluster in a day, and only stay at other places clusters for 7 hours. So it can be inferred that these positions of the important places clusters are home or work place for users.

IV. Network proximity and mobility homogeneity

We have obtained the users' important locations in the section 3, so we can build the location-based social network based on these locations, and the link is formed when two users visited the same important places. Namely, The $G = \{V, E\}$ represents the established network. V denotes the set of users, and E means the set of edges. For each user pair x and y ($x \in V$, $y \in V$), there is an edge $(x, y) \in E$ when they have the same important positions.

In order to analyze the mobility characteristics of users in location-based social networks and predict links, we introduce a series of indicators that describe network proximity and mobility homogeneity. By calculating these indicators, we can find the correlation between them and understand how they impact on each other.

A. Network proximity

Most of the link prediction studies focus on the proximity of the network, because the two nodes are more likely to have a connection in the future when they are close enough but not yet connected, so we have selected three common indexes to represent the network proximity.

- Common Neighbor. The number of neighbors shared by two users.

$$CN \equiv |\Gamma(x) \cap \Gamma(y)| \tag{1}$$

  $\Gamma(x)$ represents the neighbor set of user x.
- Adamic-Adar. It sums the inverse logarithm of common neighbors' degree.

$$AA(x, y) \equiv \sum_{z \in \Gamma(x) \cap \Gamma(y)} \frac{1}{log\Gamma(z)} \tag{2}$$

- Jaccard's Coefficient. The number of neighbors shared by two users divided by the size of their neighbors' union, characterizing the similarity between their sets of neighbors.

$$J(x, y) \equiv |\Gamma(x) \cap \Gamma(y)| / |\Gamma(x) \cup \Gamma(y)| \tag{3}$$

B. Mobility homogeneity

Similar to the above network proximity, the link prediction can be carried out by the user's mobility homogeneity. The more similar the user's movement trajectories are, the higher the likelihood of their contact will be in the future.

- Distance. The minimum of the shortest distance sets between the users' most frequent visited locations.

$$d(x, y) \equiv min\{d | d = dist(ML(x), ML(y))\} \tag{4}$$

$$ML(x) \equiv argmax_{l \in Loc} PV(x, l) \tag{5}$$

$$PV(x, l) = \sum_{i=1}^{n(x)} \frac{\delta(l_i, L_i(x))}{n(x)} \tag{6}$$

  $\delta(x, y) = 1$, if $x = y$, 0 otherwise.
  Loc denotes the set of locations someone visited. $n(x)$ is the size of this set. So $L_i(x)$ is x's i-th visited location.

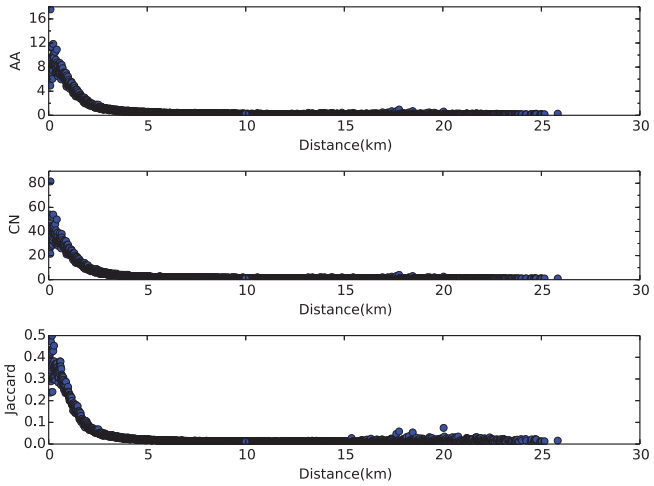

Fig. 3. Correlation between network proximity measures (AA, CN and Jaccard) and Distance.

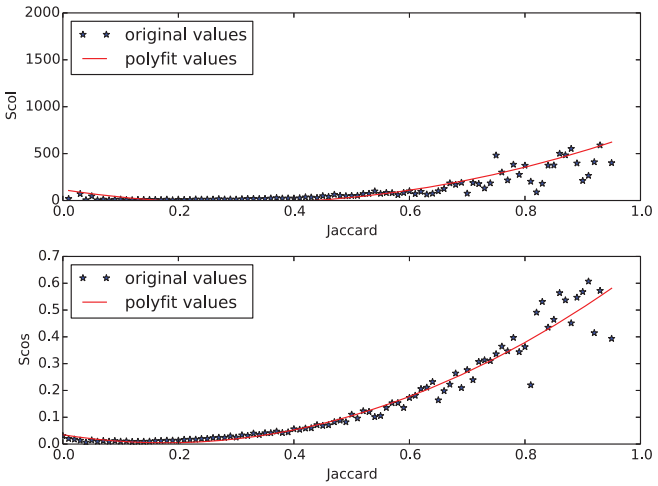

Fig. 4. Correlation between mobility measures (Scol, Scos) and Jaccard's coefficient.

- Spatial Co-Location Rate. The probability that two users appear at the same location is not limit to the same time.

$$SCol\,(x,y) \equiv \sum_{l \in Loc} PV\,(x,l) \times PV\,(y,l) \qquad (7)$$

- Spatial Cosine Similarity. The cosine similarity of users' trajectories, representing the normalization of spatial co-location rate based on mold.

$$SCos\,(x,y) \equiv \sum_{l \in Loc} \frac{PV\,(x,l) \times PV\,(y,l)}{\|PV\,(x,l)\| \times \|PV\,(y,l)\|} \qquad (8)$$

C. Correlation between network proximity and mobility homogeneity

The network proximity indexes characterize the users' degree of their tightness and possibility of connected in the network, meanwhile the mobility homogeneity indicators indicate the similarity of their trajectories and activity range. In Fig. 3 we plot the median values of Common neighbors, Adamic-Adar, Jaccard's coefficient for different values of Distance to observe the correlation between each other. And it shows that the network proximity between two individuals decays with geographical distance.

Fig. 4 demonstrates that users who are not connected with each other but have a larger proportion of common neighbors in their neighbors' set, will have the higher degree of trajectory overlap, namely their similar mobility patterns is more similar.

For the space limitation, the figure can't fully show such correlation in all cases, and we calculate the Pearson coefficients of each pairs of variables, the results are listed in Table I.

As mentioned earlier, these mobility measures (Distance, Spatial Co-Location Rate, Spatial Cosine Similarity) aim at describing the degree of trajectories overlap or geographic closeness of two users, and the network proximity measures (Common Neighbor, Adamic-Adar, Jaccard's Coefficient) show the possibility that two not yet connected users will be contacted in the future. As can be seen from Table I, the correlation between the same kind of indicators is higher than that of other categories. However, mobility homogeneity also correlates with network proximity, especially such correlation strongly exists between spatial cosine similarity and jaccard coefficient, which is 0.81. It means that even if two individuals are not connected be colleagues and friends, namely, they do not work or live at the same places in reality, they will have high possibility to be connected in the future as long as their mobility pattern are similar enough. Simultaneously, the connection strength not only correlates with others network proximity measures, but also has a certain correlation with mobility homogeneity. So we can predict the new links in the network by the above two categories of indicators.

V. Embodying Social Relationship in Mobility

From section 4 we have concluded that the two individuals' social connection related with their mobility. And this section focuses on how to further embody the social relationship in users' mobility. For hourly variations in mobility similarity, we next use unsupervised clustering method to identify four categories of social ties. And we divide the time into weekdays (1∼24) and weekends (25∼48) based on hour granularity, defining the $scos(t)$ for the cosine similarity of trajectories

TABLE I
Pearson Coefficients

	D	Scol	Scos	CN	AA	J	W
D	1.00	-0.10	-0.13	-0.58	-0.57	-0.79	-0.14
Scol	-0.10	1.00	0.81	0.33	0.02	0.57	0.02
Scos	-0.13	0.81	1.00	0.74	0.44	0.81	0.41
CN	-0.58	0.33	0.74	1.00	0.99	0.93	0.93
AA	-0.57	0.02	0.44	0.99	1.00	0.94	0.83
J	-0.79	0.57	0.81	0.93	0.94	1.00	0.82
W	-0.14	0.02	0.41	0.93	0.83	0.82	1.00

[a]D : Distance, J : Jaccard Coefficient, W : Connection Strength.

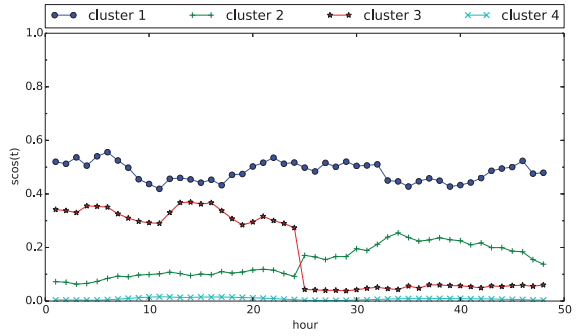

Fig. 5. Characterizing social relationship based on similarity of trajectories over time.

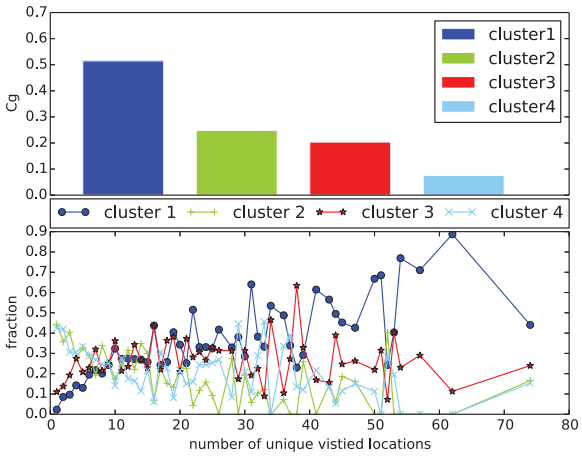

Fig. 6. Clustering coefficient and unique visited location of each cluster.

between the user pairs in different time periods. To identify the groups, we use the k-means clustering algorithm on these $scos(t)$ similarity time series. The optimal clustering number is 4 according to calculating the WSS (sum of squared error) value and result are as follows.

- Cluster1: family/friends with high mobility similarity on nights and weekends. And their mobility similarity has been reduced at daytime, but also kept at a higher value, indicating that their workplace is also in the same area.
- Cluster2: co-workers with high mobility similarity on workdays and low mobility similarity on weekends.
- Cluster3: friends with low mobility similarity on weekdays, and high mobility similarity on weekends.
- Cluster4: unfamiliar users with low level mobility similarity.

We extract subgraphs containing only edges in a single cluster, finding that these subgraphs retain high clustering coefficient (C_g) with family, friends and co-workers groups.

For those who have a certain social relationship between the users, they are more concentrated in the network, that is, their important positions are more overlapping. Each user's activity range is different, the fraction of edges that belong to each identified groups based on their unique visited locations are shown in Fig. 6. Friends with the common interest have relatively large range of activities compared to the colleagues with some fixed locations. Hence, we explore the categories of locations the users tend to visit in each clusters by using Baidu map application programming interface. The results are shown in Table II.

The table shows the first six categories, we can conclude that for family or colleagues their common important positions are generally real estate (i.e. residential or office area), in contrast, friends' common important place are generally tourist attractions or leisure places, and unfamiliar individuals do not have a specific type of common locations because of their different interest, so they may encounter in a traffic hub coincidentally.

VI. Link Prediction

Section 4 and 5 indicate that the two individuals' daily movement can determine their social relationship, and the mobility and network proximity measures correlate connection strength. We next focus on the link prediction by considering both unsupervised and supervised methods with these factors. In this section, link prediction is formalized as a binary classification problem in the collection of all the potential links. We consider the network proximity and mobility homogeneity measures introduced in section 4 to specify the classification category.

In the sample data set, it consists of $n = 793$ users and $m = 11,043$ old links, producing $(n(n-1)/2)m = 314,821$ potential links. And the actual new links are 2,338 in the testing data set. According to above formulation, our goal is to predict whether a potential link becomes a new link in the testing data set based on the sample data set.

A. Unsupervised method

For unsupervised method, we first rank the set of potential links by using one of the network proximity and mobility homogeneity indexes, then select the top k ranked potential links as new links, where k is 2,388 in our experiments. The rest are classified as missing links.

We study the precision in this case, and the precision for different quantities is computed by considering the fraction of actual new links in the set of top-ranked 2,388 potential links.

The result of table III shows that the Distance measure has a high precision. Because the network link is built when the users have the same important places, if users are closer,

TABLE II
Location Categories

Cluster1	Cluster2	Cluster3	Cluster4
Real estate	Real estate	Tourist	Transportation
Education	Education	Leisure	Car service
Business	Medical	Shopping	Hotel
Life service	Sports	Food	Medical
Tourist	Car service	Financial	Shopping
Leisure	Financial	Hotel	Finance

they are more likely to have a new link. And other mobility and network measures also have considerable predictive power compared to the random prediction.

B. Supervised method

For supervised method, we construct a learning classifier named decision tree, which can classify the potential links as new links or missing links according to network proximity and mobility homogeneity measures.

In order to improve the precision of prediction, we apply the subset of potential links under the constraint distance < 1 and common neighbors > 20 (or Scos > 0.4). In this case, our tree's precision is 59.3%, but the recall is 21.3%. Traditionally, the precision and recall can draw the system PR (Precision-Recall) curve, and the curve determines the merits of the system. However, we are more concerned on precision, as the most challenging task is to classify some potential links as new links at a high probability, even with the price of a nonnegligible number of false negatives. So from our defined classifier, considering both network proximity and mobility homogeneity measures appropriately can effectively improve the prediction precision.

VII. Conclusion

In this paper, by analyzing the collected cellular network data, we show that the locations of the individuals indicate their social relationships. For instance, family stay at home together at night, colleagues work together in the same work area during the day, and friends meet in a leisure place on weekends. So it's feasible to build the location-based social network by identifying these users' important locations. The rest finding consists of two aspects. Firstly, the network proximity, mobility homogeneity and connection strength strongly correlate with each other. Secondly, mobility measures have considerable prediction power with traditional network measure, and combining the two measures can significantly improve the precision of the prediction. In the future, we can take the time factors into account to describe people's social relationships more accurately or we can predict their locations by considering the users' social relationships and activity models.

So based on these, we can determine the relationship between users by their trajectories, and obtain locations' categories which they usually go. And it is possible to speculate whether unfamiliar people will meet in these places in the future. Then, according to users' groups and their mobility, we can plan the function areas in the city rationally. All of above findings have important implications for the exploration of social groups, traffic infrastructure and urban planning.

TABLE III
Precision

Index	Distance	Scol	Scos	CN	AA	Jaccard	Random
Precision	29.65%	2.26%	3.22%	3.12%	3.48%	4.61%	0.74%

Acknowledgment

This work is supported in part by the National Natural Science Foundation of China (61671078, 61701031), Director Funds of Beijing Key Laboratory of Network System Architecture and Convergence (2017BKL-NSAC-ZJ-06), and 111 Project of China (B08004, B17007). This work is conducted on the platform of Center for Data Science of Beijing University of Posts and Telecommunications.

References

[1] China Internet Network Information Center, China Statistical Report on Internet Development, January 2017.

[2] Zhang Y. User mobility from the view of cellular data networks[C]//INFOCOM, 2014 Proceedings IEEE. IEEE, 2014: 1348-1356.

[3] Gonzalez M C, Hidalgo C A, Barabasi A L. Understanding individual human mobility patterns[J]. Nature, 2008, 453(7196): 779-782.

[4] Jiang S, Ferreira J, Gonzlez M C. Activity-based human mobility patterns inferred from mobile phone data: A case study of Singapore[J]. IEEE Transactions on Big Data, 2017, 3(2): 208-219.

[5] Giannotti F, Nanni M, Pinelli F, et al. Trajectory pattern mining[C]//Proceedings of the 13th ACM SIGKDD international conference on Knowledge discovery and data mining. ACM, 2007: 330-339.

[6] Sia-Nowicka K, Vandrol J, Oshan T, et al. Analysis of human mobility patterns from GPS trajectories and contextual information[J]. International Journal of Geographical Information Science, 2016, 30(5): 881-906.

[7] Isaacman S, Becker R, Cceres R, et al. Identifying important places in people's lives from cellular network data[C]//International Conference on Pervasive Computing. Springer Berlin Heidelberg, 2011: 133-151.

[8] Rivera M T, Soderstrom S B, Uzzi B. Dynamics of dyads in social networks: Assortative, relational, and proximity mechanisms[J]. annual Review of Sociology, 2010, 36: 91-115.

[9] Toole J L, Herrera-Yaqe C, Schneider C M, et al. Coupling human mobility and social ties[J]. Journal of The Royal Society Interface, 2015, 12(105): 20141128.

[10] Isaacman S, Becker R, Cceres R, et al. A tale of two cities[C]//Proceedings of the Eleventh Workshop on Mobile Computing Systems & Applications. ACM, 2010: 19-24.

[11] Hightower J, Consolvo S, LaMarca A, et al. Learning and recognizing the places we go[C]//International Conference on Ubiquitous Computing. Springer Berlin Heidelberg, 2005: 159-176.

[12] Kim D H, Hightower J, Govindan R, et al. Discovering semantically meaningful places from pervasive RF-beacons[C]//Proceedings of the 11th international conference on Ubiquitous computing. ACM, 2009: 21-30.

[13] G Krings G, Calabrese F, Ratti C, et al. Urban gravity: a model for inter-city telecommunication flows[J]. Journal of Statistical Mechanics: Theory and Experiment, 2009, 2009(07): L07003.

[14] Lambiotte R, Blondel V D, De Kerchove C, et al. Geographical dispersal of mobile communication networks[J]. Physica A: Statistical Mechanics and its Applications, 2008, 387(21): 5317-5325.

[15] Grabowicz P A, Ramasco J J, Gonalves B, et al. Entangling mobility and interactions in social media[J]. PloS one, 2014, 9(3): e92196.

[16] Al Hasan M, Chaoji V, Salem S, et al. Link prediction using supervised learning[C]//SDM06: workshop on link analysis, counter-terrorism and security. 2006.

[17] Chen H, Li X, Huang Z. Link prediction approach to collaborative filtering[C]//Digital Libraries, 2005. JCDL'05. Proceedings of the 5th ACM/IEEE-CS Joint Conference on. IEEE, 2005: 141-142.

[18] Wang C, Satuluri V, Parthasarathy S. Local probabilistic models for link prediction[C]//Data Mining, 2007. ICDM 2007. Seventh IEEE International Conference on. IEEE, 2007: 322-331.

[19] Wang D, Pedreschi D, Song C, et al. Human mobility, social ties, and link prediction[C]//Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining. ACM, 2011: 1100-1108.

[20] Wang C, Zhou X, You F, et al. Design of P2P traffic identification based on DPI and DFI[C]//Computer Network and Multimedia Technology, 2009. CNMT 2009. International Symposium on. IEEE, 2009: 1-4.