Privacy-Preserving Social Tie Discovery Based on Cloaked Human Trajectories

Ye Tian, *Member, IEEE*, Wendong Wang, *Member, IEEE*, Jie Wu, *Fellow, IEEE*, Qinli Kou, Zheng Song, and Edith C.-H. Ngai, *Member, IEEE*

***Abstract*—The discovery of peoples' social connections is becoming a flourishing research topic, considering the rich social information inferable from human trajectories. Existing social tie detection methods often require mobile users to upload their accurate locations, causing serious privacy concerns. On the other hand, cloaking methods allow users to upload their obscured locations instead and can efficiently protect their location privacy. However, no existing social tie detection method can generate social relationships among users when only obscured trajectories are provided. To tackle the aforementioned problem, this paper proposes a novel semantic-tree-based algorithm. Specifically, we model the obscured regions from the cloaking algorithm as a semantic region tree and assign weight values for regions based on their popularity, further indicating the similarity between users based on their temporal and spatial relations. We evaluate our proposed approach using a real trajectory data set and show that our algorithm can identify social ties successfully with 20% higher accuracy than the existing approaches.**

***Index Terms*—Cloaked trajectory, privacy preserving, semantic similarity, social tie discovery.**

I. Introduction

Social link prediction has emerged as a hot topic in social network analysis because knowing the social ties among people would be beneficial to link prediction [1], rating prediction [2], product recommendation [3], and community discovery [4]. One commonly adopted method of inferring social ties is measuring the similarity of individuals' historical locations in both spatial and temporal dimensions because recent research studies have proven that people's social relationship may partly influence their mobility patterns. In particular, people's trajectory is an important clue for inferring social ties since we and our friends are likely to visit similar locations [5]. In recent years, location-based social network applications have become highly popular around the world. More people are using GPS-enabled devices to log their outdoor locations and activities [6], [7] and to share information about their current locations and activities with friends through remote servers. This kind of information sharing has a profound impact on social networks [8] and provides the basis of inferring social ties.

Location is sensitive information for individuals, and it is probable to be leaked out, by an untrusted server, to malicious third parties [9]. Extensive studies have been conducted to guard against individual's location privacy leaking in location-based services [9]–[13].

Most of the existing approaches are based on location perturbation and obfuscation, which employ well-known privacy metrics calculated at the server side to evaluate the level of personal data disclosure [14]. K-anonymity is a representative measure guaranteeing that a user is indistinguishable from at least $k - 1$ other users. Generally, to achieve k-anonymity, a centralized location anonymizer [15] is responsible for enlarging the queried location in a location-based service query to a larger region, which geographically covers at least $k - 1$ other users. This process is defined as *cloaking* [10], and the spatial cloaked area around user's actual location is defined as *cloaking region*, whereas a trajectory that is composed of a sequence of cloaking regions is then called a *cloaked trajectory*. As shown in Fig. 1, k is 4 and 7 for the two users, respectively. We notice that the cloaking regions have different sizes according to the setting of k. Sometimes, the cloaking region of a user with a large k might contain the cloaking region of another user with a smaller k (see the red rectangular region in Fig. 1).

Considering that cloaking-based approaches are widely adopted in location privacy preserving, identifying social ties from cloaked trajectories is pretty much a necessity. Based on k-anonymity cloaked locations, Tan *et al.* [16] proposed a social tie prediction algorithm, although it suffers from low accuracy. In our previous paper [17], we have introduced how to reveal the social connections via blurry trajectories processed by cloaking algorithms. Different from the work of Tan *et al.* [16], we take the different privacy protection levels into consideration and realize that the semantic meanings and hierarchical relations of cloaking regions would be very useful.

This paper tackles the problem of inferring social ties from obscured trajectories for privacy preservation purposes. Compared with traditional methods that ignore the location privacy of users, the distance between two cloaking regions can no

Manuscript received October 8, 2015; revised February 15, 2016; accepted April 9, 2016. Date of publication April 15, 2016; date of current version February 10, 2017. This work was supported in part by the National Natural Science Foundation of China under Grant 61370197 and Grant 61402045, by the Specialized Research Fund for the Doctoral Program of Higher Education under Grant 20130005110011, and by the National High Technology Research and Development Program of China under Grant 2013AA013301. The review of this paper was coordinated by Prof. J. Sun.

Y. Tian, W. Wang, Q. Kou, and Z. Song are with the State Key Laboratory of Networking and Switching Technology, Beijing University of Posts and Telecommunications, Beijing 100876, China (e-mail: yetian@bupt.edu.cn; wdwang@bupt.edu.cn; shuxi_kql@163.com; sonyyt@gmail.com).

J. Wu is with the Department of Computer and Information Sciences, Temple University, Philadelphia, PA 19122 USA (e-mail: jiewu@temple.edu).

E. C.-H. Ngai is with Uppsala University, 751 21 Uppsala, Sweden (e-mail: edith.ngai@it.uu.se).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TVT.2016.2554608

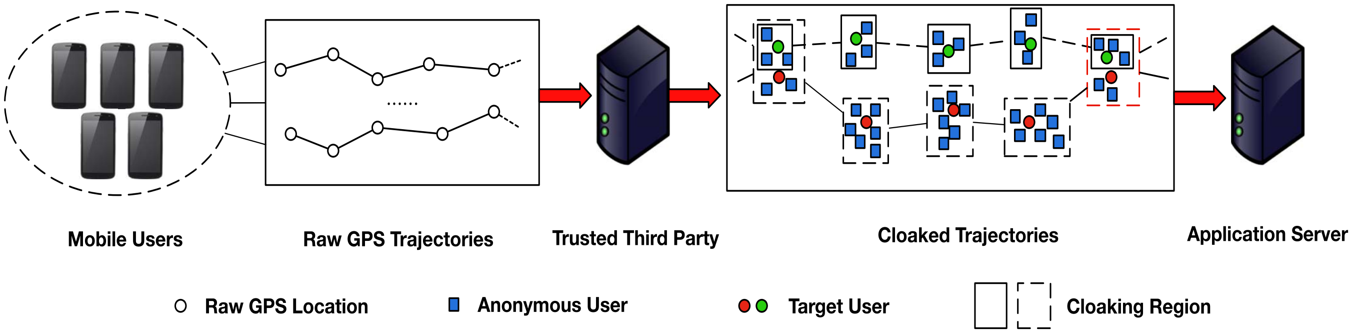


Fig. 1. Cloaked trajectories.

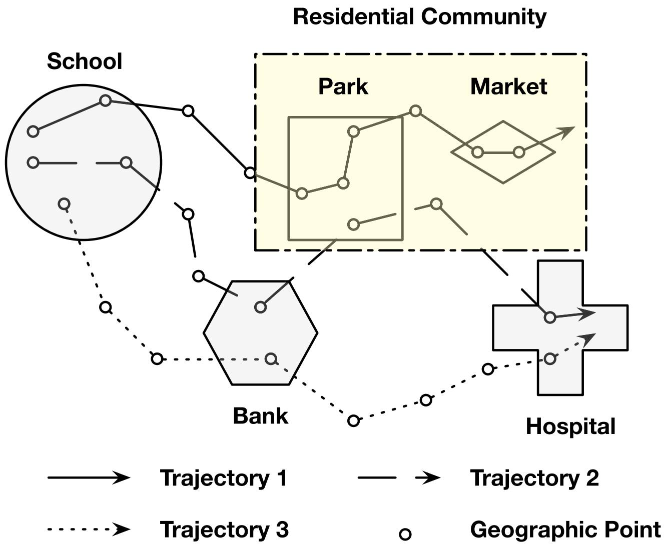


Fig. 2. Containment relationship of coverage areas.

longer reflect the actual distance between two users. To solve such problems, we proposed a *weighted hierarchical semantic tree model* in this paper. First, we transform the cloaking regions into semantic regions. Considering the different levels of privacy requirements of users, which can be reflected by their choices of k in k-anonymity cloaking, there exists a containment relationship among the semantic regions. The concept of containment relationship of semantic regions is illustrated in Fig. 2. In this figure, *residential community* contains the *park* and the *market*; for some users with high k values, the semantic region could be residential community, and for users with lower k values, the semantic region could be the park or the market. Furthermore, the difference of popularity of regions is taken into account in our paper. Since people appearing in some rarely visited regions are considered to have a larger chance of knowing each other, less popular regions are more successful in inferring social ties.

Next, we propose a novel algorithm to infer social ties among people using a weighted hierarchical semantic tree. The probability of the existence of social ties between two users is then measured by a *similarity score*.

On the basis of the previous work, we improve the accuracy in social tie discovery.

In particular, we improve the semantic tree model by further considering the impact of the popularity of semantic regions to the calculation of similarity among users. In this case, we divide regions into *popular regions* and *unpopular regions*.

For example, bars and stadiums could be regarded as the unpopular regions, whereas markets and restaurants could be regarded as popular regions. In our paper, the popularity of a region is mainly decided by its number of check-ins, and a region is more representative of people's hobbies and characteristics if it is an unpopular region. Thus, unpopular regions should be assigned a higher weight than popular regions when measuring similarity.

The contributions of this paper can be summarized as follows.

- We introduce a *weighted hierarchical semantic tree* model. In this model, different levels of user privacy preservation are taken into account. We also consider some important features of semantic regions to help construct our model, such as the containment relationship of regions and the popularity differences among regions.
- We propose a trajectory similarity measuring algorithm to discover social ties of users from the cloaked trajectories based on the weighted hierarchical semantic tree model.
- We use the Gowalla data set [5], [18] that records 196 591 users to evaluate the performance of the proposed scheme. The results show that our proposed method can improve the social tie inferring accuracy by almost 20% when compared with existing algorithms.

The rest of this paper is organized as follows. Section II reviews the related literature. Section III introduces the weighted hierarchical semantic tree model. Section IV gives full details of our algorithm for social ties detection. Section V evaluates our approach with a real-world data set and reports the evaluation results. Section VI concludes our work.

II. Related Work

The relation between social ties and human mobility has been widely explored in recent years [5], [6], [8], [19]. It is revealed that human trajectories and social ties are closely correlated [8]. Cho *et al.* [5] further studied the relation between social ties and human geographic and temporal dynamics and identified the strong indication between trajectory similarity and social tie.

Semantic trajectory data mining has emerged as an important tendency in recent studies [20], [21]. For example, Baratchi *et al.* constructed semantic location histories, with which they proposed a novel mechanism to estimate the probability if social tie exists between users by measuring their trajectory similarity [6], [22]. Liu *et al.* [23] captured landmarks on trajectory, which was composed of a sequence of locations labeled with semantic tags (called semantic locations). These semantic locations contain a wealth of information about individual's daily activities. For example, Alvares *et al.* [20] have found semantic trajectory patterns from users' mobility histories. They first mapped each stop in the trajectory to semantic landmark and then applied sequential pattern mining to find user's frequent behaviors. In [21], the user's next location on their trajectory is predicted by analyzing the geographic and semantic features.

Location information is important but sensitive; a mass of location privacy protection mechanisms have been proposed in social networking services. Zhang *et al.* proposed a suit of novel fine-grained private matching protocols to enable two users perform profile matching without disclosing privacy information for proximity-based mobile social networking [24], [25]. Sun *et al.* studied social tie discovery problem in mobile social network and adopted geographic cell index to record mobile user's location and further proposed a Private Set Intersection Cardinality (PSI-CA) protocol and a Bloom-filter-based protocol for privacy-preserving spatiotemporal matching [26]. Location perturbation and obfuscation are the most studied approaches [9], [27]. As a representative technology, k-anonymity, which was introduced by Gruteser and Grunwald [28], could reduce the probability of the target object being identified from a k-objects group to only $1/k$ [29]. Gruteser and Grunwald proposed an adaptive interval cloaking algorithm to construct spatial–temporal cloaking areas containing at least $k_{\min}$ users. After that, they sent only the cloaking areas to application servers for different kinds of services. In k-anonymity cloaking, the level of user anonymity, which is indicated by k, could be maintained by changing the size of the cloaking area [30]. A larger k corresponds to a higher privacy protection level, and *vice versa*.

A number of research studies also discussed the metrics of location privacy evaluation [31]–[34]. Intuitively, the degree of location privacy is defined as the accuracy with which an untrusted party can locate an individual. Since privacy is intrinsically related to uncertainty, entropy-based metrics are mostly adopted to evaluate the privacy protection level in anonymous communication [31], [32]. Typically, the privacy metric is defined as

$$H(k) = -\sum_{i=1}^{I} p_i \lg p_i \tag{1}$$

where p_i denotes the adversaries probabilities for different assignments of user identities to the observed position, and I indicates the total number of such assignment hypothesis. Hoh and Gruteser [31] also proposed an alternative metric with the expectation of distance error to capture how accurate an adversary can estimate individual's location. They gave the formulation as

$$E[d] = \frac{1}{NK}\sum_{k=1}^{K}\sum_{i=1}^{I} p_i(k) d_i(k) \tag{2}$$

where d_i describes the total distance error between the correct assignment hypothesis and the hypothesis i, whereas N is the number of users, and K denotes the total observation time.

In this paper, we try to identify social ties based on users' cloaked trajectory resulted from k-anonymity processing. Compared with previous studies, this approach extends the concept of semantic trajectory and explores hierarchical relationship of semantic regions for social tie identification; meanwhile, it preserves location privacy to users with different privacy protection requirement levels.

III. Proposed Model

Here, we would elaborate the detail of our proposed model. Intuitively, it is difficult to deduce social ties via obscured trajectories, since they are unable to reveal accurate location distance at the same time. In our approach, we first transform the cloaking regions to semantic regions in a preprocessing stage, to, e.g., make cloaking regions carry semantic information. Next, since users have different requirements in privacy protection, reflected in the different values of k, the granularity of cloaking regions varies. In this case, there exists hierarchical relationship among semantic regions. Users in two regions having a hierarchical relationship are likely to know each other. Third, the weight of semantic regions is allocated based on the popularity of regions. Two users have a higher probability of knowing each other if they appear in a less-visited region. This way, we can improve the accuracy of social tie discovery.

A. System Overview

A trusted anonymization server is deployed to transform raw trajectories to cloaked trajectories. Only the cloaking regions will be forwarded to the application server to support various kinds of services. Fig. 1 shows the procedure in detail. First, the trusted server collects raw GPS trajectories from the mobile devices. Then, the trusted server anonymizes the trajectories with k-anonymity to make users indistinguishable from $k-1$ other users. Different users may have different privacy protection requirements, which are determined by the value of k. A larger k brings a higher privacy protection level, and vice versa. In practice, a number of k options are presented to users. Users select suitable k empirically according to their privacy protection demand. When all trajectories are processed by the trusted server, cloaked trajectories could be shared with the application servers.

For convenience, the cloaking region is represented with a rectangle indicated by the x, y coordinates of its top left and bottom right corners. Considering that users may have different privacy protection levels, the size of cloaking regions varies accordingly. Let $\mathcal{U} = \{U_i : i = 1, 2, \ldots, U\}$ denote the set of users and $\mathcal{T}_u$ denote the corresponding trajectories. Each

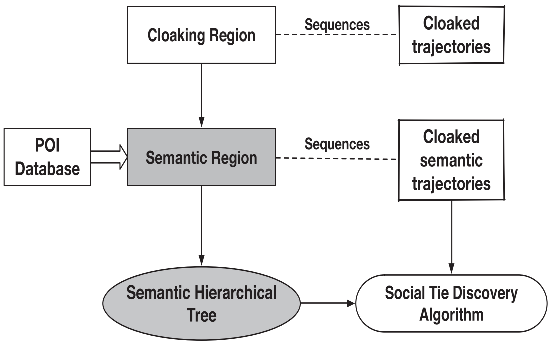

Cloaking Region
Sequences
Cloaked trajectories
POI Database
Semantic Region
Sequences
Cloaked semantic trajectories
Semantic Hierarchical Tree
Social Tie Discovery Algorithm

Fig. 3. Procedure of our model.

trajectory is composed of a set of triples $\{R_p, E_p, L_p\}$, where E_p and L_p denote the time stamp when user u arrived and left region R_p, respectively, $\mathcal{R} = \{R_i : i = 1, 2, \ldots, R\}$ is the set of cloaking regions.

B. Semantic Regions

Here, we will elaborate the procedure of labeling cloaking regions with semantic tags. The semantic meaning is represented by its detailed address rather than its GPS location. The procedure of transforming cloaking region to semantic cloaking region is illustrated in the first gray box in Fig. 3. Services such as Google Maps API provide basic functions to geographic locations to their corresponding toponym with accurate GPS information. However, a cloaked region cannot be processed with this service since it lacks exact latitude and longitude information.

It is worth noting that a semantic region transformed from a cloaking region should carry one and only one semantic meaning. If a semantic region could not meet this requirement, it needs to be expanded to carry only one semantic meaning. The operation is to include more regions.

To implement the process of transforming a cloaking region to a semantic region, we select several locations in the cloaking region as a sample set. This is conducted by dividing the cloaking region into disjoint rectangles with equal sizes and selecting the center point of each rectangle as a sample location. We denote all the sample locations in a cloaking region with $\mathcal{L} = \{L_i : i = 1, 2, \ldots, L\}$. For each location $L_i \in \mathcal{L}$, the semantic meaning associated to it could be derived by reverse geocoding using Google Maps API.[1] Let $\mathcal{S} = \{S_i : i = 1, 2, \ldots, S\}$ be the set of semantic meanings associated to the locations in $\mathcal{L}$, such that S_i represents the semantic information of location L_i.

If all the subregions in $\mathcal{L}$ carry a same semantic meaning, which is denoted by $\mathbf{A}$, it could be deduced that the semantic meaning of the cloaking region is $\mathbf{A}$. Otherwise, none of the semantic meanings in $\mathcal{S}$ could satisfy the user's anonymity level. Thus, a region with a bigger size, covering all of the semantic meaning as the semantic region of this cloaking region, should be selected as the representative region. As the structure of each semantic meaning obtained by reverse geocoding using Google Maps API is hierarchical, we search them along with the hierarchical structure to the upper levels until the semantic meaning is the same. For example, the semantic meanings in a cloaking region are different, including *Library of BUPT*, *Canteen of BUPT*, and *Basketball Court of BUPT*; thus, we should select a larger region covering all of these semantic meanings as the semantic region. In this example, the semantic region is *BUPT campus*, which is a common upper level semantic meaning. This way, we could transform all the cloaking regions into semantic regions. For each cloaking region, A_p represents the semantic meaning of the spatial–temporal portion $\{R_p, E_p, L_p\}$.

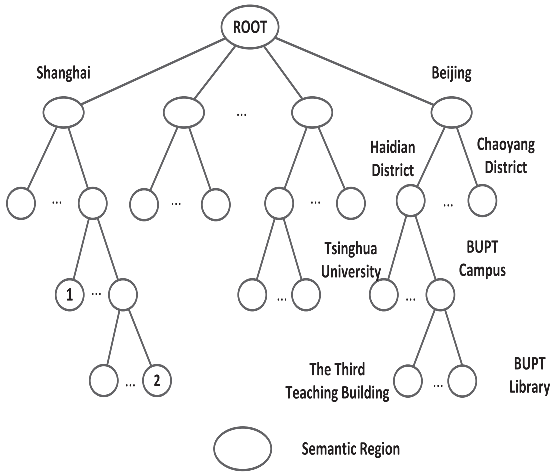


Fig. 4. Hierarchical semantic tree.

C. Hierarchical Semantic Tree

In the previous discussion, we have transformed cloaking regions into semantic regions. As the anonymity levels of individuals are different, the varying sizes of semantic regions may result in internal containment relationship. Taking Fig. 2 as an example, we assume that *Jack* and *John* appear at a park and that the anonymity level of John is higher than that of Jack. Moreover, the number of users in the park within a given time slot is too few to satisfy John's needs in privacy preservation. Thus, the semantic region should be expanded to a *residential community*, as in Fig. 2. Intuitively, a person with larger k is more likely to be allocated a semantic region with a larger size, and this region would probably contain the region that is allocated to a person holding smaller k.

Based on the containment relationship, a hierarchical semantic tree is constructed. Each node in the tree is associated with a semantic region. As illustrated in Fig. 4, nodes located in higher layer geographically contain those in lower layers. For example, *BUPT Campus* geographically covers the *BUPT library* and *the third teaching building*; thus, the node representing *BUPT library* is the parent node of the latter two. Similarly, since *BUPT Campus* and *Tsinghua University* are both located in *Haidian District*, they are both child nodes of *Haidian District*. Population density is unevenly distributed in different regions; thus, persons with constant privacy level k are not associated to a certain node fixedly. For example, a user in region 1 may transfer to region 2 in the next timestamp (see Fig. 4).

[1] https://developers.google.com/maps/documentation/geocoding

D. Weights of Semantic Regions

In our previous work, we consider that all the regions expose the same influence on similarity calculation. Different regions should contribute different influences when measuring the similarity of users' trajectories. For example, people checking in a bar may share more similar characteristics. Since a bar is a less popular region, people here are more likely to have social ties. In our model, we consider that the popularity of a region decides its weight in similarity calculations, and less popular regions should be assigned a higher weight.

To determine the weight of each region, we propose an approach to set regions' weights based on mathematical expectation and standard deviation. Mathematical expectation could reflect the probability-weighted average of all possible values of a discrete random variable. Standard deviation is used to quantify the amount of variation or dispersion of a set of data values.

First of all, we need to calculate the probability of every region being checked in. We count the number of check-ins of each region and then calculate the percentage of the check-ins of each region on each level by

$$P_{ji} = \frac{C_{ji}}{\sum_{i=1}^{|\mathcal{C}_j|} C_{ji}} \tag{3}$$

where the set of the check-ins in the jth level is denoted by $\mathcal{C}_j = \{C_{ji} : i = 1, 2, \ldots, |\mathcal{C}_j|\}$, and C_{ji} represents the number of check-ins of the ith node in the jth level. The set of percentages is denoted by $\mathcal{P}_j = \{P_{ji} : i = 1, 2, \ldots, |\mathcal{P}_j|\}$, where $|\mathcal{P}_j|$ denotes the number of nodes in the jth level. This percentage could represent the probability that a region has been checked in.

Next, we calculate the expectation and the standard deviation, which are denoted by $\mathbf{EV}$ and $\mathbf{SD}$, respectively, in each level of the tree. For the jth level, we could obtain the expectation value via

$$\mathbf{EV} = \sum_{i=1}^{|\mathcal{C}_j|} P_{ji} \times C_{ji} \tag{4}$$

where $|\mathcal{C}_j|$ is the number of nodes in the jth level of the tree, and obtain the standard deviation by

$$\mathbf{SD} = \sqrt{\frac{1}{|\mathcal{C}_j|} \sum_{i=1}^{|\mathcal{C}_j|} (C_{ji} - \mathbf{EV})^2} \tag{5}$$

where $\mathbf{EV}$ could be obtained from (4). The mathematical expectation is the probability-weighted average of different amounts of check-ins of all regions in this level.

Observing the distribution of the check-in numbers could help classify the regions into different grades. The distribution curve of the number of check-ins in certain level of the tree is demonstrated in Fig. 5. We found that the distribution curve is similar to the normal distribution, and the peak of the curve corresponds to the expectation value. In Fig. 5, the horizontal axis is the number of check-ins, and the vertical axis is the percentage. The area encompassed by the curve and two lines perpendicular to the horizontal axis represents the percentage of regions whose check-ins numbers fall into the range of the two lines. For example, the area of the shaded part could represent the percentage of regions whose number of check-ins falls into between $x = 0$ and $x = \mathrm{EV} - 2\mathrm{SD}$, which is $p\%$ in this case.

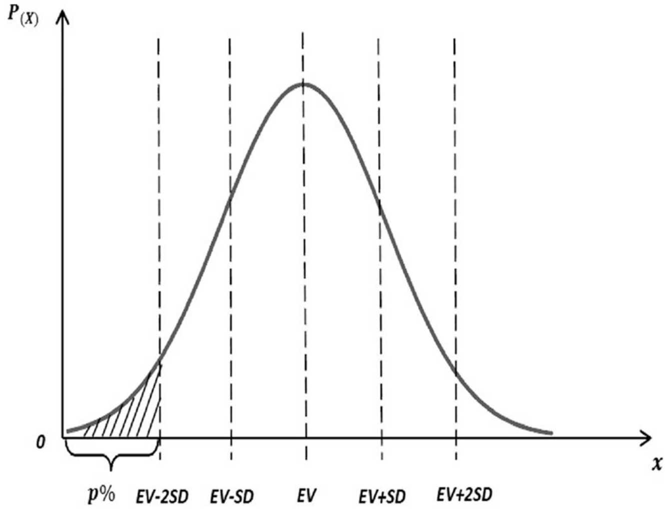


Fig. 5. Distribution curve of check-in numbers.

According to the mathematical expectation value, the standard deviation, and the distribution analysis, all of the regions at the same level could be clustered into multiple grades, and different weight values should be assigned to different grades accordingly. This way, regions could be differentiated based on popularity.

Next, we would illustrate the details of classifying regions under the premise of not changing the structure of the hierarchical semantic tree. When clustering regions of a level, the k-means method is adopted. In further details, we set the clusters and weights as follows.

- *Determine typical values as the center points of k-means.* Instead of selecting initial centering points randomly in a traditional k-means method, we select them according to $\mathbf{EV}$ and $\mathbf{SD}$. It is known that the standard deviation does measure how far typical values tend to be from the expectation value, and thus, the initial center points could be set based on the standard deviations of the expectation value (mathematically, $\mathbf{EV} \pm \mathbf{SD}$ or $\mathbf{EV} \pm 2\mathbf{SD}$). This way, some errors caused by selecting center points randomly could be reduced.
- *Cluster regions for each level.* In each level of the tree, we cluster regions by a k-means method. In the k-means method, items are clustered mainly by the distance between the item and the center points. In our work, the distance is the difference between the number of check-ins in every region and that of center points.
- *Assign different weight values to different clusters.* We rank the clusters according to the percentage of their corresponding center point, and this rank could reflect their popularity. A lower value of a center point indicates that the regions in this cluster are relatively less popular;

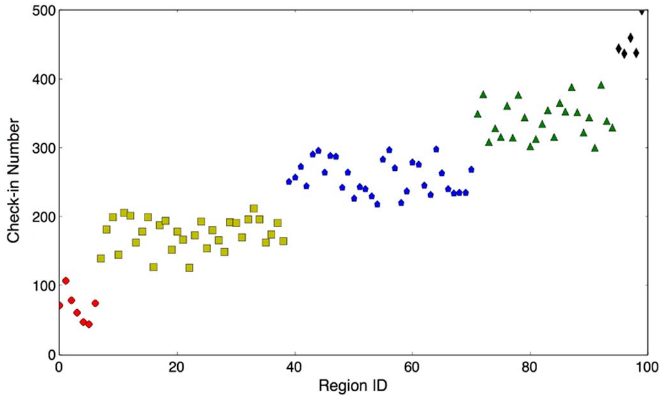


Fig. 6. Example of clustering regions.

thus, we should assign a higher weight value to this class, and *vice versa*.

As shown in Fig. 6, regions of certain level are clustered into five clusters according to the check-in numbers. In this example, $\mathbf{EV}$ is equal to 260, and $\mathbf{SD}$ is equal to 150. The points with the same color and shape belong to a cluster. In Fig. 6, we could observe that the numbers of regions in different clusters are uneven. Furthermore, a cluster contains more regions if the value of center point is closer to the expectation value, and *vice versa*. In this case, we should assign different weight values to five clusters. Regions in the cluster within which the check-in numbers of the center point are the lowest are assigned the biggest weight, and vice versa. For example, if the center point values of five clusters are v_1, v_2, v_3, v_4, and v_5, respectively, where $v_1 < v_2 < v_3 < v_4 < v_5$, their weights should be w_1, w_2, w_3, w_4, and w_5, respectively, where $w_1 > w_2 > w_3 > w_4 > w_5$.

E. Temporal Similarity

When calculating similarity of two trajectories, it is necessary to ensure that they are aligned in temporal dimension. It makes sense only if the two regions to be compared fall in a same time slot.

Definition 3.1–(Pair Regions): Suppose p_i and p_j are two spatial–temporal semantic portions from two trajectories t_1 and t_2, respectively; they are defined as *pair regions* if and only if the temporal constraint TimeDiff is satisfied. More specifically, $\text{TimeDiff}(E_{p_i}, E_{p_j}) \leqslant \delta_t$ and $\text{TimeDiff}(L_{p_i}, L_{p_j}) \leqslant \delta_t$ should be satisfied simultaneously for p_i and p_j, where δ_t is a given temporal implying the optimal time span.

To guarantee the mobility of individuals and the amount of pair regions on two trajectories, δ_t is set as 1 h. Two considerations should be fine-tuned to better determine the parameter δ_t. For one thing, to ensure that the semantic regions in a trajectory are meaningful, the time span should not be too long. For example, making a comparison between trajectories in different years makes no sense. For another, the alignment of different trajectories on the same day is not mandatory. This ensures that regions with similar patterns in different days should still be considered as pair regions. For example, Jack and John are colleagues and know each other. However, Jack goes to the company every workday, whereas John does not come to the company every day. Although Jack and John may not meet every day, we still observe similar mobility patterns between them.

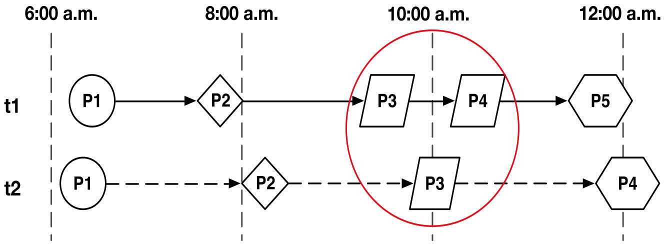


Fig. 7. Pair regions in two trajectories.

Fig. 7 illustrates the concept of pair regions, where pair regions are indicated with same shape. It is possible that a region on a trajectory may be paired with more than one region on another trajectory, i.e., p_4, p_5, and p_6 circled in Fig. 7. This is because multiple regions may satisfy the same temporal constraint δ_t. For this case, similarity between all pair regions is measured.

F. Relationship Between Trajectories' Similarity and Social Ties

It is supposed that the similarity of individuals' trajectories can imply, to a large extent, their common preference of behaviors and mobility patterns. Therefore, a high similarity score of trajectories may indicate that social tie may exist between users.

Hierarchical semantic tree is adopted as a feasible model for inferring social ties in this paper. The underlying concept is twofold. First, if users' trajectories match well, they are likely to share same pattern in the hierarchical semantic tree. Second, to say the least, although their trajectories are not exactly the same, social tie may still exist if their trajectories are close in the hierarchical semantic tree or if there exists containment relationship between their semantic regions.

IV. Proposed Social Tie Discovery Algorithm

The existence of social ties mainly depends on the similarity of trajectories between users. We can calculate the similarity of two trajectories based on pair regions. Several relevant metrics, including the lowest common ancestor node and the length of the shortest path, can be obtained from the hierarchical semantic tree, as follows.

- The level of the lowest common ancestor node. The level of the lowest common ancestor node of pair regions determines how relevant the pair regions are in the hierarchical semantic tree. Lower level may lead to greater impact on similarity measurement since a lower level node represents a more specific semantic region. If the pair regions are located in the same node in the tree, it, itself, is the lowest common ancestor node.
- The shortest length between two semantic regions. The shortest length between two regions in the hierarchical semantic tree reflects how close the two regions are.

Intuitively, they are closer if they are more geographically adjacent.
- The level of the semantic regions themselves. If the semantic regions are located in a lower level in the hierarchical semantic tree, their locations are more accurately indicated.
- The popularity of each region. We consider that persons checking in the less popular region are more likely to have social ties.

Based on the aforementioned considerations, the metrics are defined accordingly as follows.

1) $\mathrm{len}(R_i, R_j)$: the length of the shortest path between region R_i and region R_j;
2) $\mathrm{lca}(R_i, R_j)$: the lowest common ancestor node of R_i and R_j;
3) $\mathrm{depth}(R_i)$: the level of region R_i in the hierarchical semantic tree;
4) $\mathrm{deep\_max}$: the maximum level of the tree;
5) $\mathrm{sim}(R_i, R_j)$: the similarity between cloaking regions R_i and R_j in the tree;
6) $\inf(R_i)$: the influence of R_i in social ties prediction, which is decided by the level of region R_i. Intuitively, $\inf(R_i)$ increases with $\mathrm{depth}(R_i)$ monotonically;
7) $W(R_i)$: the weight of R_i, which reflects the popularity of this region.

We measure the similarity of trajectories based on these three metrics: the length of the shortest path, the level of the lowest common ancestor, and the influence of the semantic region. It is illustrated in the following equation:

$$\begin{aligned}\mathrm{simPair}(R_i, R_j) &\\ = e^{-\alpha\times \mathrm{len}(R_i,R_j)} &\\ \times \Bigg\{ \inf(R_i) \times \inf(R_j) \times e^{-\gamma(W(R_i)\times W(R_j))} &\\ \times \frac{e^{\beta\times \mathrm{depth}(\mathrm{lca}(R_i,R_j))} - e^{-\beta\times \mathrm{depth}(\mathrm{lca}(R_i,R_j))}}{e^{\beta\times \mathrm{depth}(\mathrm{lca}(R_i,Rj))} + e^{-\beta\times \mathrm{depth}(\mathrm{lca}(R_i,Rj))}} \Bigg\}. &\end{aligned} \tag{6}$$

Apparently, the value of (6) increases monotonically with respect to $\mathrm{depth}(\mathrm{lca}(R_i, R_j))$, but decreases with $\mathrm{len}(R_i, R_j)$. The smaller $\mathrm{len}(R_i, R_j)$ is or the greater $\mathrm{depth}(\mathrm{lca}(R_i, R_j))$ is, user *A* and user *B* are more likely to locate in regions of a lower level in the hierarchical semantic tree, such that they will have greater influence and achieve a maximum $\mathrm{sim}(R_i, R_j)$. On the contrary, if $\mathrm{len}(R_i, R_j)$ is closer to $2 \times \mathrm{deep\_max}$ and $\mathrm{depth}(\mathrm{lca}(R_i, R_j))$ is equal to 1 (i.e., root node), then $\mathrm{sim}(R_i, R_j)$ will be close to 0. In addition, α and β are parameters scaling the contribution of the length of the shortest path and the level of the lowest common ancestor, respectively; and γ is scaling the contribution of the weight of a region. The optimal setting of α, β, and γ should be decided by the experimental results. In addition, $\inf(R_i) \times \inf(R_j)$ represents the mutual influence contributed to the similarity. To derive the similarity of two cloaked trajectories, the hierarchical semantic tree and the set of pair regions in two trajectories (t_1 and t_2), which is denoted by $\mathcal{PR}$, are taken as input. Then, the metrics defined earlier are derived for each pair of regions. According to (6), the similarity of each pair region is then obtained. Finally, by taking the weighted average of the similarities of all pair regions, the similarity of two cloaked trajectories is achieved. The detail description is illustrated in Algorithm 1.

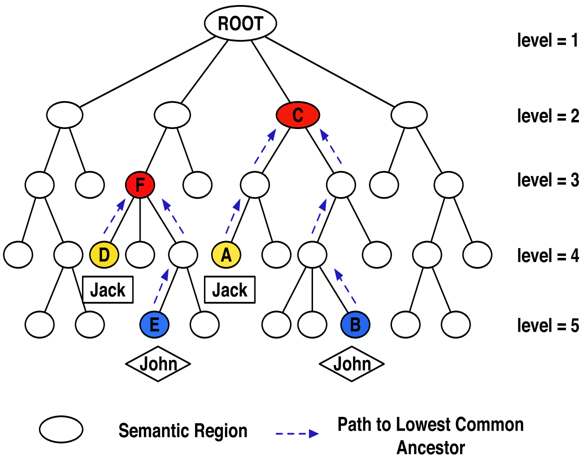


Fig. 8. Calculating the similarity of regions.

Algorithm 1 The algorithm of calculating two trajectories' similarity

```
Require:
    The hierarchical semantic tree, HST;
    The set of pair regions of trajectories $t_1$ and $t_2$, $\mathcal{PR}$;
Ensure:
    The similarity of trajectories $t_1$ and $t_2$, Sim;
Sim = 0;
Num = $|\mathcal{PR}|$;
for all $(R_i, R_j) \in \mathcal{PR}$ do
    LCANode = searchLCA($R_i$, $R_j$, HST);
    SP = ShortestPathLength($R_i$, $R_j$, HST);
    LCALevel = LocatedLevel(LCANode, HST);
    Level1 = LocatedLevel($R_i$, HST);
    Level2 = LocatedLevel($R_j$, HST);
    $W_1$ = Weight($R_i$)
    $W_2$ = Weight($R_j$)
    Inf1 = Influence(Level1);
    Inf2 = Influence(Level2);
    Update simPair according to (6);
    Sim = Sim + simPair;
end for
Sim = Sim/Num;
return Sim;
```

We take Fig. 8 as an example. Two users, namely, *Jack* and *John*, are involved in this scenario. At time m_1, the semantic region of Jack is covered by region *A*, and the semantic region of John is in region *B*. Region *C* is their lowest common ancestor, which is in the second level of the semantic tree. According to this tree, the length of the shortest path is equal to 5. Then, we can calculate the similarity of the first pair regions of these two trajectories. In the next timestamp m_2, Jack and John arrive at regions *D* and *E*, respectively. The second similarity score can also be calculated based on (6) earlier. After calculating the similarity scores of all pair regions, we set $\overrightarrow{\mathrm{s}}$ as the similarity vector to record the similarity score of each pair of regions from the two trajectories, which is denoted by $\overrightarrow{\mathrm{s}} = (\mathrm{simPair}_1, \mathrm{simPair}_2, \ldots, \mathrm{simPair}_n)$.

Finally, the similarity score of the two trajectories (t_1, t_2) can be calculated by

$$\mathrm{Sim}(t_1, t_2) = \frac{1}{|\overrightarrow{\mathrm{s}}|} \times \sum_{i=1}^{n} \mathrm{simPair}_i. \tag{7}$$

If more than one trajectory exists for an individual, each pair of trajectories between users should be compared. We assume that the set of trajectories of Jack is $\mathcal{T}_1$ and the set of trajectories of John is $\mathcal{T}_2$. For each trajectory $t_i \in \mathcal{T}_1$, we calculate a trajectory similarity score of t_i and every trajectory $t_j \in \mathcal{T}_2$. Then, an average value of all the known trajectory similarity scores can be combined to obtain an overall similarity score of two individuals. This score can be used to measure their degree of closeness. To decide whether there exist social ties between two individuals, a threshold δ_s is set. If the score is greater than threshold δ_s, we consider that it is more likely that they have social ties between them. δ_s is determined by estimating the F-measure when it achieves optimal value.

V. Performance Evaluation

A. Setup

In this paper, the proposed algorithm is verified with a real-world data set collected by Gowalla [5], [18]. It is a location-based social networking website where users share their locations by check-in. The friendship network is undirected and was collected using their public API. This data set consists of 196 591 nodes and 950 327 edges. A total of 6 442 890 check-ins from these users has been collected from February 2009 to October 2010.

Wang *et al.* have found that the similarity between two individuals' movements strongly correlates with their relationship in the social network, and the probability could reach almost 80% [8]. Thus, we select some users with high movement similarity from the original data set, and it is considered that these users really have social ties. Thus, we use the data set of these users as our data set. We adopt the following procedures to set up our simulation.

- We consider that users' trajectories with too few records could not reflect their mobility routine. It will be difficult to discover social ties among the users with little information. Thus, we ignore users with less than 20 check-ins.
- For all of the retained users, we randomly set k for each of them to represent individual privacy protection level $(3 \leqslant k \leqslant 10)$. We select consecutive check-ins within a defined time interval as a trajectory and only choose trajectories having at least 20 locations. Then, we store all retained trajectory information, friendship information, and the correspondences between users and k in a local MySQL database.
- For a given trajectory composed of GPS check-in locations, we transform it to a cloaked trajectory. Next, we transform all of the cloaking regions to semantic regions.
- We divide the data set into two partitions: a training set containing the previous 75% of consecutive records in each trajectory and a testing set containing the remaining 25% of records in each trajectory. All of the trajectories are processed by the k-anonymity cloaking method.

Precision, recall, and F-measure are the main measurements for the experimental evaluation. They are usually used in information retrieval tasks. In information retrieval contexts, precision and recall are defined in terms of a set of retrieved documents and a set of relevant documents. Precision is the fraction of retrieved instances that are relevant, whereas recall is the fraction of relevant instances that are retrieved. F-measure is the harmonic mean of precision and recall. In this paper, the precision rate, the recall rate, and the F-measure are redefined by

$$\mathrm{Precision} = \frac{p^+}{p^+ + p^-} \tag{8}$$

$$\mathrm{Recall} = \frac{p^+}{|R|} \tag{9}$$

$$\text{F-measure} = \frac{2 \times \mathrm{Precision} \times \mathrm{Recall}}{\mathrm{Precision} + \mathrm{Recall}} \tag{10}$$

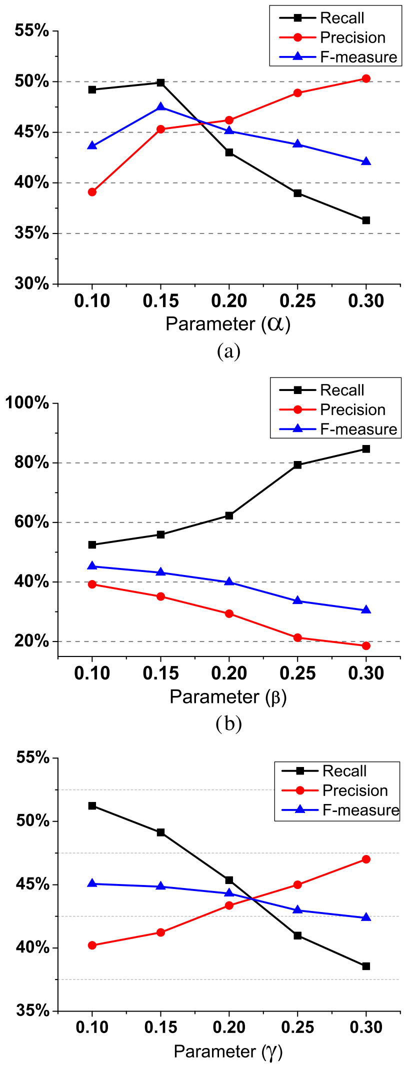
where p^+ and p^- indicate the numbers of correct and incorrect predictions of the existence of social ties, respectively. $|R|$ indicates the total number of social tie records in the social network.

Experiments are divided into two parts: sensitivity tests and performance comparison. The sensitivity tests evaluate the proposed algorithm under various parameter settings (i.e., α, β, and γ) in (6). In performance comparison, we mainly evaluate the performance from three aspects. First, we analyze the performance of our approach with different groups of people holding different values of k. Second, we contrast our model with the k-anonymous spatial–temporal cloaking model (KSTCM) [16]. The anonymity levels for different users are varied in our paper. Finally, we evaluate the performance by using semantic trajectories without cloaking and make a comparison with the performance using cloaked trajectories.

B. Sensitivity Tests

This test aims to evaluate the performance of the proposed algorithm under different parameter settings. To investigate the impact of different factors, in each step, only one parameter would be fine-tuned, whereas the others remain unchanged.

As can be observed from Fig. 9, when α and γ grow, precision improves, but recall decreases. Conversely, precision decreases, but recall improves when β grows. Meanwhile, it is interesting to find β gains better impact to influence both precision and recall, which suggests that the level of the lowest common ancestor node plays an important role in determining the performance of our proposed mechanism. In addition, it could be concluded that the containment relationship among



(a)

(b)

(c)

Fig. 9. Performance in various parameter settings. (a) Performance changing with α. (b) Performance changing with β. (c) Performance changing with γ.

semantic regions is influenced in similarity calculation. In our previous work [17], we consider that all regions have the same weight values. In this paper, we could observe that the performance improves greatly when the region popularity is taken into consideration.

To explore the best combination of α, β, and γ, more experimental results with various parameter settings are reported.

Although there is no causal relation between precision and recall, they are regarded as two interconstraint measures. In this situation, the best performance would be obtained when F-measure achieves the highest value. This is because F-measure takes both precision and recall into consideration. Precision and recall get closer when F-measure increases. From the parameter setting experiment, we set $\alpha = 0.15$, $\beta = 0.1$, and $\gamma = 0.15$ after this experiment, since it achieves the best performance of F-measure and precision and recall are most harmonious. In this situation, our approach could achieve 47.22% in terms of precision and 49.05% in terms of recall, which reflects the performance when considering all anonymity levels of privacy protection.

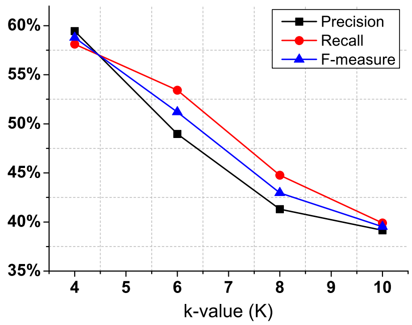


Fig. 10. Influence of k on performance.

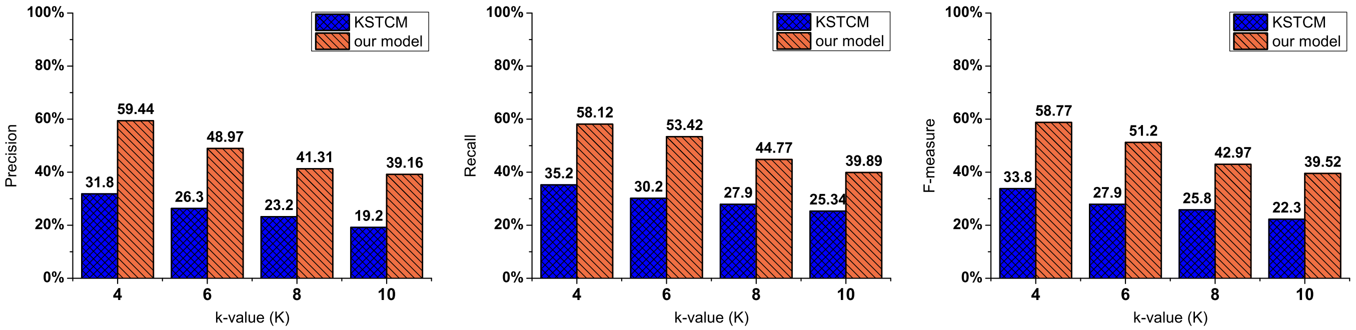
C. Performance Comparisons

To validate the performance of our proposed model, comparisons are conducted with previous studies. More specifically, the experiments are divided into the following steps.

1) We first validate the performance in terms of anonymity levels by grouping all users according to their anonymity level k.
2) The performance is compared between our approach and KSTCM [16]. We calculate precision, recall, and F-measure when k is set to 4, 6, 8, and 10.
3) To validate the influence of the weights of semantic regions, we compare the performance with our previous work [17], which does not consider the impact of semantic regions' popularity.
4) To evaluate the proposed algorithm under the condition of ignoring the privacy protection, we evaluate the performance using purely semantic trajectories without k-anonymity cloaking. In this comparison, we transform the raw GPS trajectories to semantic trajectories directly without cloaking.

Apparently, the size of the cloaking regions grows when k increases. The growing size of the cloaking regions will inevitably lead to an indefinite representation of the region and finally cause a degradation of performance. As illustrated in Fig. 10, precision and recall gradually deteriorate as k increases, which implies that a better performance is available for users with low privacy requirements.

In KSTCM [16], the privacy protection levels of all the users are the same by default. A comparison of two models is given in Fig. 11. As can be seen, our approach outperforms KSTCM in terms of precision, recall, and F-measure. It demonstrates that our approach using semantic regions is more capable than that using raw cloaking regions. The result implies that semantic regions could reveal more individuals' interests and preferences, and individuals usually have closer social ties when sharing similar semantic regions or locating in the less-popular regions. A comparison between this model and our previous work [17], which does not take the difference of popularity of regions into account, is shown in Fig. 12. As shown in Fig. 12, the performance improves by almost 5% when considering the weight of semantic regions based on their popularity. It suggests that the popularity of regions is helpful to discover social ties,



(a) (b) (c)

Fig. 11. Comparison of precision, recall, and F-measure under different values of k. (a) Precision rate comparison. (b) Recall rate comparison. (c) F-measure comparison.

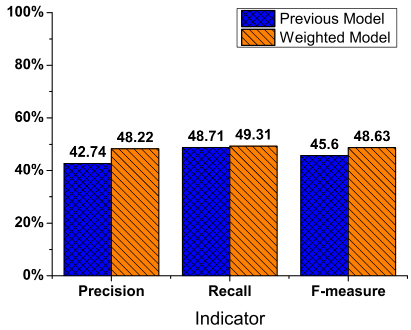


Fig. 12. Performances of our previous model and of the weighted hierarchical semantic tree model.

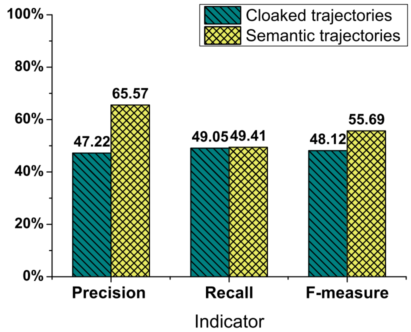


Fig. 13. Performances using cloaked trajectories and semantic trajectories.

and less-popular regions could better represent users' interests and characteristics.

Finally, we evaluate our proposed algorithm with semantic trajectories without cloaking by transforming raw GPS trajectories to semantic regions directly. This way, we compare the performance of our approach under the situations with and without providing privacy preservation. From Fig. 13, we observe that there is a significant improvement in precision rate using accurate semantic trajectories rather than cloaked trajectories. It suggests that accurate locations benefit in improving the accuracy of social ties detection. Intuitively, raw GPS locations carry more accurate semantic meanings, which would probably make the semantic regions locate in lower levels of the hierarchical semantic tree. From the results, we see that the three metrics do not degrade significantly, considering the challenge of using cloaked trajectories. We validate that our algorithm is capable of discovering social ties with a reasonable performance while preserving location privacy effectively.

VI. Conclusions and Future Work

In this paper, we have proposed a novel approach to infer social ties with cloaked trajectories instead of accurate GPS trajectories for privacy preserving. Instead of matching historical locations directly in geographic space, we presented a novel model to transform cloaking regions to semantic regions and further proposed a weighted hierarchical semantic tree model to make the containment relationship visible. The weighted hierarchical semantic tree is then used to calculate the similarity of the trajectories of individuals and predict the existence of social ties. We have conducted extensive experiments to evaluate the performance of our approach with a real data set. The evaluation results demonstrated that our approach could infer social ties and effectively preserve privacy of users. We have compared our proposed model with existing work and demonstrated that our approach could achieve much higher performance in social tie detection.

In the future, we would like to further improve the accuracy of discovering social ties by clustering similar users based on their semantic patterns. In addition, we plan to apply privacy preservation techniques to new applications that require location privacy protection, such as in participation sensing.

References

[1] J. Zhang, C. Wang, and J. Wang, "Who proposed the relationship?: Recovering the hidden directions of undirected social networks," in *Proc. WWW*, 2014, pp. 807–818.

[2] P. Symeonidis, E. Tiakas, and Y. Manolopoulos, "Product recommendation and rating prediction based on multi-modal social networks," in *Proc. 4th ACM Conf. RecSys*, 2011, pp. 61–68.

[3] H. Ma, T. C. Zhou, M. R. Lyu, and I. King, "Improving recommender systems by incorporating social contextual information," *Trans. Inf. Syst.*, vol. 29, no. 2, p. 9, 2011.

[4] S. Parthasarathy, Y. Ruan, and V. Satuluri, "Community discovery in social networks: Applications, methods and emerging trends," in *Social Network Data Analytics*. Berlin, Germany: Springer-Verlag, 2011, pp. 79–113.

[5] E. Cho, S. A. Myers, and J. Leskovec, "Friendship and mobility: User movement in location-based social networks," in *Proc. KDD*, 2011, pp. 1082–1090.

[6] X. Xiao, Y. Zheng, Q. Luo, and X. Xie, "Inferring social ties between users with human location history," *J. Ambient Intell. Human. Comput.*, vol. 5, no. 1, pp. 3–19, 2014.

[7] A. Y. Xue *et al.*, "Destination prediction by sub-trajectory synthesis and privacy protection against such prediction," in *Proc. IEEE ICDE*, 2013, pp. 254–265.

[8] D. Wang, D. Pedreschi, C. Song, F. Giannotti, and A.-L. Barabasi, "Human mobility, social ties, and link prediction," in *Proc. 17th SIGKDD Int. Conf. Knowl. Discovery Data Mining*, 2011, pp. 1100–1108.

[9] B. Niu, Q. Li, X. Zhu, G. Cao, and H. Li, "Achieving k-anonymity in privacy-aware location-based services," in *Proc. INFOCOM*, 2014, pp. 754–762.

[10] J. G. Khuong Vu and R. Zheng, "Efficient algorithms for k-anonymous location privacy in participatory sensing," in *Proc. IEEE INFOCOM*, 2012, pp. 2399–2407.

[11] G. Ghinita, P. Kalnis, A. Khoshgozaran, C. Shahabi, and K.-L. Tan, "Private queries in location based services: Anonymizers are not necessary," in *Proc. SIGMOD*, 2008, pp. 121–132.

[12] K. P. Puttaswamy *et al.*, "Preserving location privacy in geosocial applications," *IEEE Trans. Mobile Comput.*, vol. 13, no. 1, pp. 159–173, Jan. 2014.

[13] X. Pan, J. Xu, and X. Meng, "Protecting location privacy against location-dependent attacks in mobile services," *IEEE Trans. Knowl. Data Eng.*, vol. 24, no. 8, pp. 1506–1519, Aug. 2012.

[14] M. Li, S. Salinas, A. Thapa, and P. Li, "n-CD: A geometric approach to preserving location privacy in location-based services," in *Proc. IEEE INFOCOM*, 2013, pp. 3012–3020.

[15] M. F. Mokbel, C.-Y. Chow, and W. G. Aref, "The new Casper: Query processing for location services without compromising privacy," in *Proc. 32nd Int. Conf. Very Large Data Bases*, 2006, pp. 763–774.

[16] R. Tan, J. Gu, P. Chen, and Z. Zhong, "Link prediction using protected location history," in *Proc. IEEE 5th ICCIS*, 2013, pp. 795–798.

[17] Q. Kou, Y. Tian, Z. Song, E. Ngai, and W. Wang, "Privacy preserving social tie discovery based on cloaked human trajectories," in *Proc. HOTPOST*, 2015, pp. 13–18.

[18] J. Leskovec and A. Krevl, "SNAP Datasets: Stanford Large Network Dataset Collection," Jun. 2014. [Online]. Available: http://snap.stanford.edu/data

[19] S. Scellato, A. Noulas, and C. Mascolo, "Exploiting place features in link prediction on location-based social networks," in *Proc. 17th SIGKDD Int. Conf. Knowl. Discovery Data Mining*, 2011, pp. 1046–1054.

[20] L. O. Alvares *et al.*, "Towards semantic trajectory knowledge discovery," *Data Mining Knowl. Discovery*, 2007, to be published.

[21] J. J.-C. Ying, W.-C. Lee, T.-C. Weng, and V. S. Tseng, "Semantic trajectory mining for location prediction," in *Proc. 19th ACM SIGSPATIAL Int. Conf. Adv. Geograph. Inf. Syst.*, 2011, pp. 34–43.

[22] M. Baratchi, N. Meratnia, and P. J. Havinga, "On the use of mobility data for discovery and description of social ties," in *Proc. IEEE/ACM Int. Conf. Adv. Soc. Netw. Anal. Mining*, 2013, pp. 1229–1236.

[23] J. Liu, O. Wolfson, and H. Yin, "Extracting semantic location from outdoor positioning systems," in *Proc. IEEE 7th Int. Conf. MDM*, 2006, p. 73.

[24] R. Zhang, J. Zhang, Y. Zhang, J. Sun, and G. Yan, "Privacy-preserving profile matching for proximity-based mobile social networking," *IEEE J. Sel. Areas Commun.*, vol. 31, no. 9, pp. 656–668, May 2013.

[25] R. Zhang, R. Zhang, J. Sun, and U. Yan, "Fine-grained private matching for proximity-based mobile social networking," in *Proc. IEEE INFOCOM*, 2012, pp. 1969–1977.

[26] J. Sun, R. Zhang, and Y. Zhang, "Privacy-preserving spatiotemporal matching," in *Proc. IEEE INFOCOM*, 2013, pp. 800–808.

[27] B. Gedik and L. Liu, "Protecting location privacy with personalized k-anonymity: Architecture and algorithms," *IEEE Trans. Mobile Comput.*, vol. 7, no. 1, pp. 1–18, Jan. 2008.

[28] M. Gruteser and D. Grunwald, "Anonymous usage of location-based services through spatial and temporal cloaking," in *Proc. 1st Int. Conf. Mobile Syst., Appl. Services*, 2003, pp. 31–42.

[29] M. Wernke, P. Skvortsov, F. Dürr, and K. Rothermel, "A classification of location privacy attacks and approaches," *Personal Ubiquitous Comput.*, vol. 18, no. 1, pp. 163–175, 2014.

[30] B. Gedik and L. Liu, "Location privacy in mobile systems: A personalized anonymization model," in *Proc. ICDCS*, 2005, pp. 620–629.

[31] B. Hoh and M. Gruteser, "Protecting location privacy through path confusion," in *Proc. IEEE 1st Int. Conf. SecureComm*, 2005, pp. 194–205.

[32] R. Shokri, "Quantifying and protecting location privacy," Ph.D. dissertation, Comput. Commun. Sci., École Polytech. Fédérale de Lausanne, Lausanne, Switzerland, 2013.

[33] A. Serjantov and G. Danezis, "Towards an information theoretic metric for anonymity," in *Privacy Enhancing Technologies*. New York, NY, USA: Springer-Verlag, 2002, pp. 41–53.

[34] A. R. Beresford and F. Stajano, "Location privacy in pervasive computing," *IEEE Pervasive Comput.*, vol. 2, no. 1, pp. 46–55, Jan.–Mar. 2003.

Ye Tian (M'15) received the B.S. degree from Nanjing University of Posts and Telecommunications (NUPT), Nanjing, China, in 2003 and the M.S. and Ph.D. degrees in computer science and technology from Beijing University of Posts and Telecommunications (BUPT), Beijing, China, in 2007 and 2013, respectively.

He is currently a Postdoctoral Researcher with BUPT. He has authored or coauthored over 20 papers. He is the holder of 13 U.S./China patents. His research interests include social network analysis, text mining, and mobile computing.

Wendong Wang (M'02) received the B.S. and M.S. degrees from Beijing University of Posts and Telecommunications (BUPT), Beijing, China, in 1985 and 1991, respectively.

He is currently a Full Professor with BUPT. He is also currently on the Assessment Panel of the National Natural Science Foundation Program and the National High Technology Research and Development Program of China. He has authored or coauthored over 200 papers in various journals and conference proceedings. He is the holder of 14 U.S./China patents. His current research interests include next-generation network architecture; Internet of Things; participatory sensing; wireless ad hoc, sensor, and mesh networks; and mobile Internet.

Jie Wu (M'89–SM'94–F'09) received the B.S. and M.S. degrees from Shanghai University of Science and Technology (currently Shanghai University), Shanghai, China, in 1982 and 1985, respectively, and the Ph.D. degree in computer engineering from Florida Atlantic University, Boca Raton, FL, USA, in 1989.

He is currently the Chair of and a Laura H. Carnell Professor with the Department of Computer and Information Sciences, Temple University, Philadelphia, PA, USA. Prior to joining Temple University, he was a Program Director with the National Science Foundation and a Distinguished Professor with Florida Atlantic University. He has regularly published scholarly journals, conference proceedings, and books. His current research interests include mobile computing and wireless networks, routing protocols, cloud and green computing, network trust and security, and social network applications.

Dr. Wu is a Distinguished Speaker of the China Computer Federation (CCF). He serves on several editorial boards, including that of the IEEE TRANSACTIONS ON COMPUTERS, the IEEE TRANSACTIONS ON SERVICE COMPUTING, and the *Journal of Parallel and Distributed Computing*. He served as the General Chair for the Association for Computing Machinery (ACM) International Symposium on Mobile Ad Hoc Networking and Computing (MobiHoc) 2014. He was the General Cochair/Chair for the 2006 IEEE International Conference on Mobile Ad hoc and Sensor Systems (MASS) and the 2008 IEEE International Parallel and Distributed Processing Symposium, the Program Cochair for the 2011 IEEE International Conference on Computer Communications (INFOCOM), the General Chair for the 2013 IEEE International Conference on Distributed Computing Systems, and the Program Chair for the 2013 CCF China National Computer Congress. He has been an IEEE Computer Society Distinguished Visitor, an ACM Distinguished Speaker, and the Chair for the IEEE Technical Committee on Distributed Processing. He received the 2011 CCF Overseas Outstanding Achievement Award.

Qinli Kou received the B.S. and M.S. degrees from Beijing University of Posts and Telecommunications (BUPT), Beijing, China, in 2013 and 2016, respectively.

She is currently with the State Key Laboratory of Networking and Switching Technology, BUPT. Her research interests include participatory sensing and social network analysis.

Zheng Song received the Ph.D. degree from Beijing University of Posts and Telecommunications (BUPT), Beijing, China.

Prior to studying at BUPT, he was with the Research and Development Department, Sina Corporation. He is currently with the State Key Laboratory of Networking and Switching Technology, BUPT. He is the holder of ten U.S. and China patents. His research interests include participatory sensing, indoor localization, and Internet of Things.

Edith C.-H. Ngai (S'02–M'07) received the Ph.D. degree from The Chinese University of Hong Kong (CUHK), Hong Kong, in 2007.

During 2007–2008, she was a Postdoctoral Fellow with the Department of Electrical and Electronic Engineering, Imperial College London, London, U.K. She is currently an Associate Professor with the Department of Information Technology, Uppsala University, Uppsala, Sweden. Previously, she conducted research with the VIEW Laboratory, CUHK; the Multimedia and Wireless Networking Group, Simon Fraser University, Vancouver, BC, Canada; the Tsinghua National Laboratory for Information Science and Technology, Tsinghua University, Beijing, China; the Intelligent Systems and Networks Group, Imperial College London; and the Networked and Embedded Systems Laboratory, University of California, Los Angeles, CA, USA. She was also a VINNMER Fellow (2009), awarded by VINNOVA, Sweden. Her research interests include wireless sensor and mobile networks, information-centric networking, Internet of Things and cloud, network security and privacy, and e-health and smart city applications.