Wireless Networks (2020) 26:5551–5560
https://doi.org/10.1007/s11276-019-02058-8

Privacy preserving semantic trajectory data publishing for mobile location-based services

Rong Tan[1] · Yuan Tao[2] · Wen Si[1] · Yuan-Yuan Zhang[3]

Published online: 15 June 2019

Abstract
The development of wireless technologies and the popularity of mobile devices is responsible for generating large amounts of trajectory data for moving objects. Trajectory datasets have spatiotemporal features and are a rich information source. The mining of trajectory data can reveal interesting patterns of human activities and behaviors. However, trajectory data can also be exploited to disclose users' privacy information, e.g., the places they live and work, which could be abused by a malicious user. Therefore, it is very important to protect the users' privacy before publishing any trajectory data. While most previous research on this subject has only considered the privacy protection of stay points, this paper distinguishes itself by modeling and processing semantic trajectories, which not only contain spatiotemporal data but also involve POI information and the users' motion modes such as walking, running, driving, etc. Accordingly, in this research, semantic trajectory anonymizing based on the k-anonymity model is proposed that can form sensitive areas that contain $k-1$ POI points that are similar to the sensitive points. Then, trajectory ambiguity is executed based on the motion modes, road network topologies and road weights in the sensitive area. Finally, a similarity comparison is performed to obtain the recordable and releasable anonymity trajectory sets. Experimental results show that this method performs efficiently and provides high privacy levels.

Keywords Mobile services · Location-based services · Semantic trajectory · Trajectory privacy protection

1 Introduction

Due to the development of mobile devices and positioning technologies, various kinds of mobile positioning devices, such as car navigation systems, GPS-enabled mobile phones, mobile wearable devices, tablet computers and position sensors, have been made available to consumers in recent years [1–5]. The popularity of mobile positioning devices has spawned numerous location-based services (LBSs) [6–8] and has generated large amounts of locational data as well [9–11]. According to statistics, each moving object in LBSs transmits its current location every 15 s on average, which indicates that more than 100 million pieces of location information are transmitted per second. And the data are extensively applied in everyday life, thereby constantly influencing people's lifestyles, working habits and thinking modes. By making observations of a person's personal life, it is possible to provide a person with convenient location-based services by speculating on where he or she lives or where he or she goes every day. For instance, it is feasible to design ideal travel routes for a person in accordance with the person's available quantity of motion trajectory data [12–16].

Location data have created both benefits and problems, and of the problems, privacy disclosure is the most

✉ Yuan-Yuan Zhang
yyzhangzcmu@163.com

Rong Tan
tanrong529@gmail.com

Yuan Tao
taoyuan@shu.edu.cn

Wen Si
cs6401outlook@yeah.net

[1] College of Information and Computer Science, Shanghai Business School, Shanghai 201400, China

[2] Computing Center, Shanghai University, Shanghai 200444, China

[3] College of Information Technology, Zhejiang Chinese Medical University, Hangzhou 310053, China

prominent issue [17–23]. In fact, the abuse of location data may lead to the disclosure of a user's most important personal information such as their personal interests, social relationships and living habits. For instance, potential attackers can not only identify the locations visited by a mobile user but also discover their home address and job location by analyzing their spatiotemporal trajectories. They can even derive private information such as a user's behavioral patterns from their daily motion trajectories, thereby posing a great threat to the user's safety and property. There have been cases when exposure of trajectory data caused damaging privacy disclosures and threatened a user's personal safety.

Researchers have proposed multiple solutions for solving privacy disclosure problems caused by LBSs. The existing privacy protection methods for LBSs mainly include data encryption, pseudoaddresses, space conversions and anonymity areas [24–29]. These methods are mostly focused on the location data, without delving into the relationships between the location data and the users, or the privacy implications of the location data. It is difficult to capture the significance of real-time human activities. Therefore, a growing number of researchers have studied location privacy protection based on semantics, with a view toward achieving a deeper level of protection. The protection of semantics-based moving object trajectories has also become a focus of more research [30–32].

With the increasing awareness of semantic information in trajectories, trajectory protection methods have gradually developed into methods based on semantics. Monreale et al. classified locations in order to generate generalized user access addresses, which enabled the creation of anonymity trajectory datasets that ensured that the probability of identifying user IDs and accessing sensitive locations was lower than a given threshold [33]. Lee et al. also imposed a threshold on the information obtainable by adversaries [34]. They suggested exploring location semantics by observing users' length of stay. Moreover, the ratio of suppressed frequent sequences is a direct indication of anonymized data quality for trajectory pattern mining [35, 36]. This paper regards the length of stay as a semantic feature extractable for LBSs and considers it a location semantics metric that can protect users' privacy. It should be noted that the above methods merely involve semantic privacy protection for stay points. In fact, semantic trajectories can contain more information (i.e., motion modes of moving objects such as walking, running, cycling and driving) due to new developments in mobile technology. Therefore, it is necessary to adopt different trajectory privacy protection strategies for different motion modes. More importantly, increases in the semantic information contained in trajectories have posed greater challenges for user privacy protection.

This paper presents a semantic-based trajectory anonymity protection method. The semantic trajectory is modeled based on the data it obtains including longitude, latitude, timestamp, POI yellow page information, velocity and motion mode. Subsequently, users' sensitive points (stay points) are identified and combined with the different motion modes as inputs for a pruning process. The pruning processing is carried out in the geographical space that covers the sensitive points. Finally, similarity comparisons are performed to obtain recordable and releasable anonymity trajectory datasets.

2 Semantic trajectory anonymity protection algorithm

In this section, an algorithm is proposed, and it consists of four main steps, as follows:

Step 1 Semantic trajectory modeling: The algorithm preprocesses the raw data and extracts spatiotemporal sequences, important spatial points (starting points, end points and stay points), velocities and motion modes. In other words, the raw data acquired are transformed into semantic trajectories as defined in Definition 1.
Step 2 Sensitive area construction: The sensitive point is processed based on the k-anonymity model, eventually forming a coverage area that contains $k-1$ POI points of a similar type to the sensitive point. The coverage area is referred to as the sensitive area.
Step 3 Trajectory ambiguity: Trajectory ambiguity is performed according to the users' motion modes, the road network topologies and the road weights in the sensitive area. The targets of ambiguity mainly include the start–end points and the stay points. The ambiguity methods can be divided into two types, trajectory segment pruning and trajectory segment addition.
Step 4 K-anonymity set construction: A similarity comparison is performed to form an anonymity set that contains the other $k-1$ trajectories with the highest similarity.

Step 2 can effectively prevent semantic location attacks and reduce the attack probability to 1/K. Step 3 can effectively prevent maximum velocity attacks. It prunes the existing trajectory segments or constructs new trajectory segments, thereby preventing the attackers from effectively calculating the users' range of motion. Finally, the privacy protection effects can be significantly improved by releasing the trajectory k-anonymity datasets.

2.1 Semantic trajectory modeling

Semantic information such as velocity, timestamp and motion mode is all directly obtainable from the client. The sampling locations merely contain the latitude and longitude and contain no actual semantic information. The acquisition of useful semantic location information depends on the client and the GIS server. This section mainly describes how to extract the start–end and the stay points from the original location data. The start–end point refers to the starting and the ending points of a trajectory, while the stay point refers to the locations visited by mobile users. Both contain important semantic information on the moving objects and are regarded as sensitive points that need special protection. Therefore, semantic trajectories can be defined as follows:

Definition 1 The semantic trajectory model is expressed as $ST = \langle (x_0,y_0,z_0,p_0,s_0,w_0),\ldots,(x_n,y_n,z_n,p_n,s_n,w_n) \rangle$, where x_i, y_i, z_i, p_i, s_i and w_i represent the longitude, latitude, timestamp, POI yellow page information, velocity and motion mode, respectively.

There are mainly two methods for extracting the stay points. One method is to extract stay points based on the length of stay, which is also the simplest method. It is necessary to set a time threshold, tth, when this method is adopted. A stay occurs when the time interval between two consecutive sampling locations is greater than tth and the distance between the two locations is smaller than the displacement threshold dth. Another method is to extract stay points based on the sampling density, which is essentially a supplement to the first method. Users tend to move at a low velocity when they stop at a certain outdoor location. Therefore, the actual stay points of users can be obtained by clustering low-velocity sampling points (the velocity is close to 0), as shown in Fig. 1.

In practice, the two methods are usually combined, thereby obtaining important location information.

2.2 Sensitive area construction

The entire geographic space is divided into several grid areas before sensitive area construction and detonated as $SG_{m\times n} = \{G(i,j)|1 \le i \le m, 1 \le j \le n\}$. Based on the actual conditions of the city where the objects are located and the roads shared by users, the unit length Δl of each grid area $G(i,j)$ can range from 0.02 to 0.05 latitude and longitude coordinate intervals. Here, we select the intervals based on latitude and longitude coordinates mainly because the road network is generally stored in the spatial database in the latitude and longitude format.

Second, each grid area G is further divided into $k \times k$ subgrids $G_{k\times k} = \{g(i,j)|1 \le i \le k, 1 \le j \le k\}$. The unit length of each subgrid is a 0.006 latitude and longitude coordinate interval, corresponding to an actual length of approximately 1 km. This unit of length not only achieves high computational accuracy but also reduces computational labor.

Sensitive area construction can be performed after dividing the areas and obtaining the semantic trajectories. This project adopts the k-anonymity model for sensitive area construction, that is, the area must contain at least $k-1$ location points of a similar type.

The sensitive area constructed based on the k-anonymity model can be quickly obtained through a k-nearest neighbor query of the GIS database. This project adopts the PostGIS spatial database – a database that can be obtained through the following query statements:

```
SELECT g1.gid g2.gid FROM points as g1, polygons g2
  WHERE g1.gid <> g2.gid AND g1.type = g2.type
  ORDER BY g1.gid, ST_Distance(g1.the_geom,g2.the_geom)
LIMIT k;
```

On the other hand, a MBB that satisfies the k value may be too large, eventually reducing the availability of the semantic trajectory. Therefore, it is necessary to determine the maximum value for a MBB. This paper considers the subgrid area that covers the sensitive point as the largest MBB possible.

2.3 Trajectory ambiguity

The trajectory ambiguity refers to the ambiguity processing of trajectories based on sensitive areas and other semantic information. The targets of ambiguity mainly include start–end points and stay points. The ambiguity method can be divided into two types, namely, trajectory segment pruning and trajectory segment addition. The pruning method, as the name suggests, removes trajectory segments that contain sensitive points. These trajectory segments tend to exist in the vicinity of sensitive points, and users tend to move at a low velocity in these areas. The addition method involves constructing new trajectory segments and combining them with real trajectory segments to form new

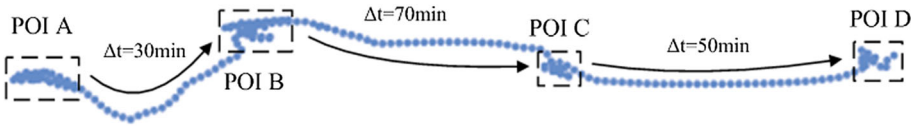

Fig. 1 The sampling density-based method

trajectories. The two methods can be combined together to form new trajectories, thereby achieving the goal of user privacy protection.

2.3.1 Start–end point ambiguity

The main steps for accomplishing start–end point ambiguity are as follows:

- The first step is to calculate the sensitive area.
- A trajectory can be directly pruned when the trajectory contains the start–end point, involves the motion mode of walking and satisfies the following two conditions:
  - There are trajectory segments that contain different motion modes in the sensitive area.
  - The sensitive point is at least 300 m (200–500 m) away from the end point of the trajectory section that contains the starting point of the trajectory, or the starting point of the trajectory section that contains the end point of the trajectory. The remaining trajectory segments will form new semantic trajectories.
- When the conditions are not satisfied, it is necessary to recalculate the weights of roads in the sensitive area and select a point in the road with the lowest weight as the new start–end point (the start–end point should be at least 300 m away from a sensitive point). The point will be combined with the starting point or ending point of the original trajectory segment in the sensitive area to form a new trajectory segment, eventually forming a new semantic trajectory.

For instance, the semantic trajectory in Fig. 2 can be considered to be $ST = \langle ST_{s1}, ST_{s2} \rangle$. ST_{s1} mainly involves the motion mode of walking and contains semantic information of the starting point (home), while ST_{s2} mainly involves the motion mode of driving. Suppose that the starting point (home) is set as a sensitive point. The first thing to do is calculate the sensitive area (red rectangle in the figure). Since the end point of ST_{s1} is less than 300 m from a sensitive point (home), the trajectory cannot be simply pruned. It is necessary to recalculate the weights of the roads in the sensitive area and select the road with the lowest weight to construct a new trajectory segment. In the figure, the blue road indicates an arterial road and has the lowest weights. Therefore, the red point is selected as the new starting point and combined with the black point to form the shortest path. Consequently, a new trajectory segment set $ST = \langle ST_{s3}, ST_{s2} \rangle$ is formed.

2.3.2 Stay point ambiguity

In contrast to start–end point ambiguity, stay point ambiguity can directly prune a trajectory segment that contains a stay point. The remaining trajectory segments can be processed according to the length of stay.

- The length of stay exceeds the threshold Δt.

When the user stays at a location for a long time, a recombination of the remaining trajectory segments will lead attackers to search for abnormal semantic information and obtain privacy information due to the rich semantic information contained in the semantic trajectory. To address this problem, this paper directly splits the remaining set of trajectory segments and recombines the trajectory segments by using start–end point ambiguity.

- The length of stay does not exceed the threshold Δt.

In this case, this paper performs ambiguity processing on the semantic information of other trajectory segment datasets in the sensitive area, to achieve the goal of sensitive point protection. The ambiguity method mainly includes velocity ambiguity (random average velocity) and

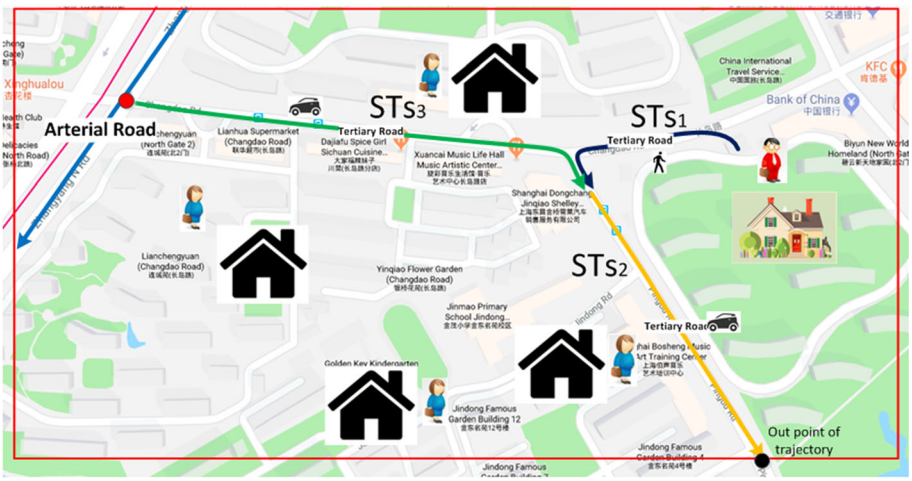

Fig. 2 Start-end point ambiguity

timestamp ambiguity (random average time). The ambiguity of velocity and time prevents the attackers from directly obtaining abnormal semantic information, thereby preventing them from obtaining information about the sensitive point.

For instance, the semantic trajectory in Fig. 4 can be considered to be $ST = \langle ST_{s1}, ST_{s2}, ST_{s3} \rangle$. ST_{s1} and ST_{s3} mainly involve the motion mode of driving, while ST_{s2} mainly involves the motion mode of walking. In addition, ST_{s2} contains semantic information for the hospital. Suppose that the hospital is set as a sensitive point. The first step is to construct a sensitive area (red rectangle in the figure). Subsequently, ST_{st2} can be directly pruned and the time interval between ST_{st1} and ST_{st3} can be evaluated.

If the time interval is less than the threshold Δt, it is necessary to perform velocity and timestamp ambiguity on ST_{st1} and ST_{st3} and reset the corresponding semantic information. For instance, the average velocity is set to: $(dist(ST_{s1}) + dist(ST_{s3}))/(time(ST_{s1}) + time(ST_{s2}) + time(ST_{s3}))$.

If the time interval exceeds the threshold Δt, it is necessary to split the remaining trajectory segment set and recombine the trajectory segments by using start–end point ambiguity. For instance, $ST = \langle ST_{s1}, ST_{s2}, ST_{s3} \rangle$ in Fig. 3 is split into $ST_1 = \langle ST_{s1} \rangle$ and $ST_2 = \langle ST_{s3} \rangle$. Suppose the red point is selected as a new start–end point; then, the new trajectory sets are $ST_1' = \langle ST_{s1}' \rangle$ and $ST_2' = \langle ST_{s2}', ST_{s3} \rangle$, respectively.

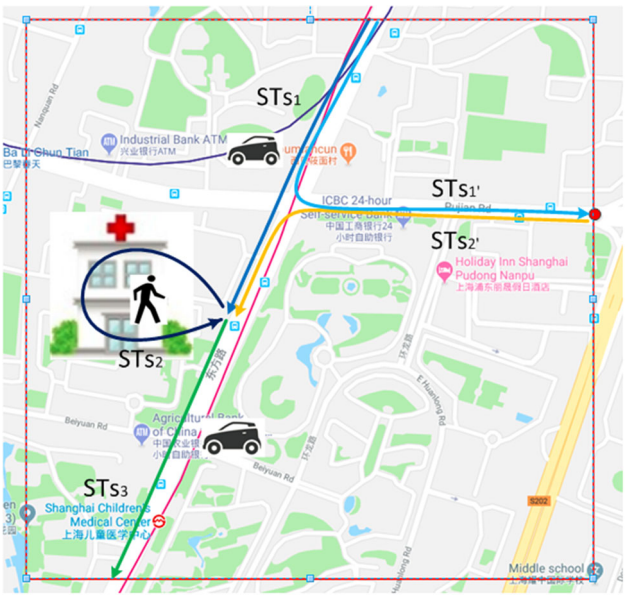

Fig. 3 Stay point ambiguity

3 Trajectory set construction based on the K-anonymity model

Trajectory sets can be constructed on the basis of the k-anonymity model after semantic trajectory ambiguity is accomplished. The construction of anonymity sets mainly depends on two factors, namely, spatiotemporal similarity and semantic similarity. Spatiotemporal similarity mainly refers to the similarity of two trajectories in geospatial and temporal dimensions, while semantic similarity mainly refers to the semantic similarity of two trajectories for stay points and motion modes.

3.1 Spatial distance measurement

In terms of spatial similarity, this paper adopts a similarity algorithm based on the Hausdorff distance (HD). The HD is a measure of the degree of similarity and is a defined form of the distance between two sets of points. HD can effectively calculate the distance between images without establishing a corresponding relationship between the templates and the sample pixels, and thereby, it is widely used in the field of mode recognition.

Definition 2 Hausdorff distance (HD)

Given two point sets $A = \{a_1, \ldots, a_p\}$ and $B = \{b_1, \ldots, b_q\}$, the HD between the two point sets can be calculated as follows:

$$H(A,B) = max(h(A,B), h(B,A)) \tag{1}$$

where

$$h(A,B) = \max_{a_i \in A} \left(\min_{b_j \in B} \|a_i - b_j\| \right) \tag{2}$$

$$h(B,A) = \max_{b_j \in B} \left(\min_{a_i \in A} \|b_j - a_i\| \right) \tag{3}$$

$\|\cdot\|$ is the distance paradigm between the two point sets A and B.

Since the HD is highly sensitive to outliers such as noise points, even a few noise points can significantly affect the distance values. To address this problem, some scholars have proposed the modified Hausdorff distance (MHD). The MHD is defined as follows:

Definition 3 Modified Hausdorff distance (MHD)

$$H(A,B) = \frac{1}{m_a} \sum_{a_i \in A} \min_{b_j \in B} \|a_i - b_j\| \tag{4}$$

where m_a is the number of objects in point set A.

The spatial distance between trajectories can be calculated by using the following equation.

Definition 4 Given two point sets $A = \{a_1,\ldots,a_p\}$ and $B = \{b_1,\ldots,b_q\}$, the spatial distance between the two point sets can be calculated as follows:

$$H_{spatial}(A,B) = \max(h_{spatial}(A,B), h_{spatial}(B,A)) \quad (5)$$

where

$$h_{spatial} = \frac{1}{l_A}\sum_{a_i \in A}\left(\frac{\|a_{i-1} - a_i\| + \|a_i - a_{i+1}\|}{2} \times \min_{b_j \in B}\|a_i - b_j\|\right) \quad (6)$$

where l_A is the total spatial length of trajectory A.

3.2 Temporal distance measurement

The temporal attributes (i.e., the timestamp) of trajectories are generated along with the spatial sampling. It is meaningless to discuss the temporal distance of moving objects without considering the specific forms of the spatial trajectories. Therefore, MHD can be also used to measure the temporal distance between trajectories. The definition is as follows:

Definition 5 Given two point sets $A = \{a_1,\ldots,a_p\}$ and $B = \{b_1,\ldots,b_q\}$, the temporal distance between the two point sets can be calculated as follows:

$$H_{temporal}(A,B) = \max(h_{temporal}(A,B), h_{temporal}(B,A)) \quad (7)$$

where

$$h_{temporal} = \frac{1}{t_A}\sum_{a_i \in A}\left(\frac{|t_{a_{i-1}} - t_{a_i}| + |t_{a_i} - t_{a_{i+1}}|}{2} \times |t_{a_i} - t_{b_j}|\right) \quad (8)$$

where t_A is the total temporal length of trajectory A.

3.3 Spatial–temporal distance measurement

The method of measuring the spatiotemporal similarity between trajectories can be derived from the spatial and temporal distance measurement methods.

Definition 6 Given two point sets $A = \{a_1,\ldots,a_p\}$ and $B = \{b_1,\ldots,b_q\}$, the spatiotemporal distance between the two point sets can be calculated as follows:

$$H_{spatial-temporal}(A,B) = (H_{spatial}(A,B), H_{temporal}(A,B)) \quad (9)$$

3.4 Semantic distance measurement

Cosine similarity is a measure of the similarity between two nonzero vectors of an inner product space that measures the cosine of the angle between them. Since cosine similarity can be applied to a comparison of vectors of any dimension, it is widely used in similarity measurements, especially in text similarity measurements. This paper adopts the cosine similarity method to measure the similarity between the semantic trajectories in stay points and motion modes.

Cosine similarity is defined as follows:

Definition 7 Suppose that the semantic values of two locations vectors are *sem(A)* and *sem(B)*. The similarity between the two semantic values can be expressed as follows:

$$S^i_{sim}(A,B) = \frac{sem(A)sem(B)}{\|sem(A)\|\|sem(B)\|} \quad (10)$$

where i indicates the semantic contents compared.

The cosine value is limited to a range of [0,1]. The higher the semantic similarity between the two locations is, the closer the cosine value is to 1. The lower the semantic similarity is, the closer the value is to 0.

3.5 Semantic trajectory similarity measurement

Definition 8 The similarity between two semantic trajectories ST_a and ST_b can be defined as follows:

$$Sim(ST_a, ST_b) = \left(H_{spatial-temproal}(A,B), \frac{1}{m_{poi}}\sum S^{poi}_{sim}(A,B), \frac{1}{n_{way}}\sum S^{way}_{sim}(A,B)\right) \quad (11)$$

The specific calculation steps are described below:

Step 1 Perform noise reduction on the two trajectories A and B (the trajectory segments merely include the start–end point, velocity outlier, stay point and road network node).

Step 2 Interpolate the various points in trajectories A and B into a third trajectory, to eliminate the impacts of different sampling sizes, reference locations and strategies.

Step 3 Calculate the spatial distance between the two trajectories by using Eq. 5.

Step 4 Calculate the temporal distance between the two trajectories by using Eq. 7.
Step 5 Calculate the spatiotemporal similarity between the two trajectories by using Eq. 9.
Step 6 Calculate the semantic similarity between the two trajectories for the start–end and stay points.
Step 7 Calculate the semantic similarity between the two trajectories for motion modes.
Step 8 Obtain the similarity between the two trajectories.

In summary, the pseudocode of the k-anonymity model-based trajectory set construction algorithm is as follows:

Algorithm **Semantic Trajectory Anonymizing based on K-anonymity Model**

Input:
$ST = \{ST_0,...,ST_n\}$: semantic trajectory dataset; ST_i : target semantic trajectory
Output:
AS(ST): anonymized set of ST ;
Algorithm:

```
1:  begin
2:  if|AS| < kthen
3:  foreach ST_j in ST ( ST_j ≠ ST_i ) do
4:  computeSimilairty( ST_i , ST_j ); //fomular 5-12
5:  S=(k-1)NN( ST_i );
6:  AS.put(S);
7:  returnAS;
```

4 Evaluation

The algorithm was written in Java and implemented on a DELL Optiplex host (CPU Core: 2 Duo 2 GHz; RAM: 4096 MB). The relational database system and the GIS database were Postgre9.1 and Postgis1.5, respectively. All the road datasets in the experiment came from national road datasets provided by OpenStreetMap. All the trajectory information came from the MyMap App. The experiment involved a total of 352,234 road data records and 36,825 trajectory records.

4.1 Overall performance

First, the overall performance of the algorithm was evaluated. We adopted the default grid division method to divide the grids into subgrids with unit lengths of 0.006 of the latitude and longitude coordinate interval (corresponding to an actual length of about 1 km). We considered POI settings such as home, school, hospital, bank and restaurant as the sensitive points and set the k value of the sensitive area(sak) to 2, 3, 4 and 5. For the semantic trajectory anonymity protection algorithm, we measured the algorithm efficiency in the cases of k = 3, 5, 8 and 10, and dataset amount ranges from 10 to 30 k separately and took the average. The execution time of the algorithm is shown in Fig. 4.

It can be seen from the figure that the average execution time is lengthy, which indicates that the algorithm does not have good efficiency. First, the algorithm needs to calculate sensitive areas and sort the road weights in the trajectory ambiguity process. Second, Dijkstra's shortest-path algorithm is adopted to construct new trajectory segments in sensitive areas and regenerate new semantic trajectories. For both of these reasons, the algorithm is time-consuming. All the road network data are prestored in the GIS database when the grid division method is adopted for sensitive area construction. The construction of new trajectory segments will be accelerated in that case. In other words, the overall performance of the algorithm is improved by taking these optimization measures. In general, the anonymity sets are released when the database is offline; therefore, it has no impact on the actual users.

4.2 Information loss rate

Information loss refers to the loss of the original trajectory information caused by trajectory anonymization. It is calculated by using the following equation:

$$IL = M_{poi}/N_{poi} \tag{12}$$

where k is the number of trajectories in the anonymity sets; M_{poi} is the number of sensitive points after ambiguity processing; and N_{poi} is the number of sensitive points in the original trajectory sets.

Figure 5 illustrates the information loss caused by an anonymity set release. It can be seen from the figure that the information loss rate gradually increases with the increase in the k value. It is recommended that both k values of sensitive area and semantic trajectory anonymizing are set to a small threshold. Then, the privacy level and information loss can both be acceptable.

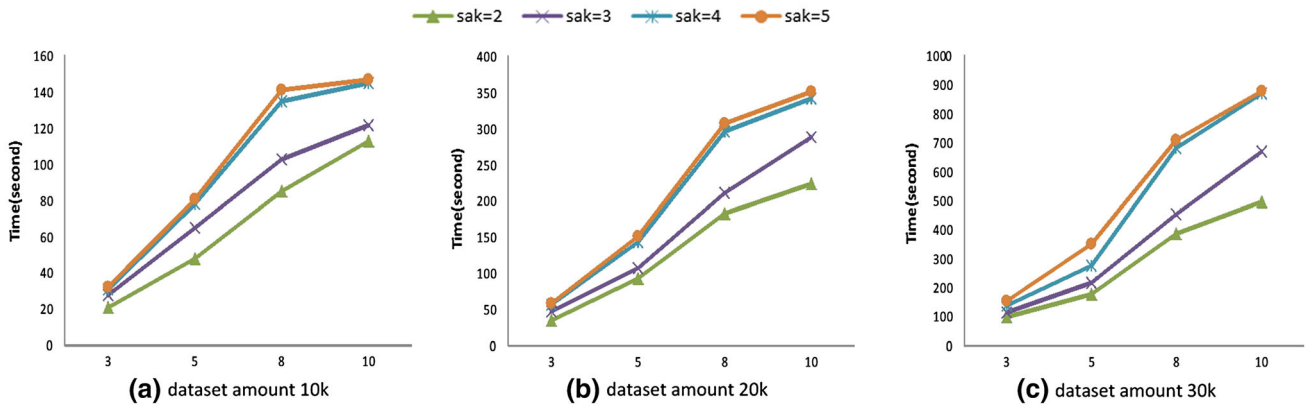

Fig. 4 Overall performance of the algorithm

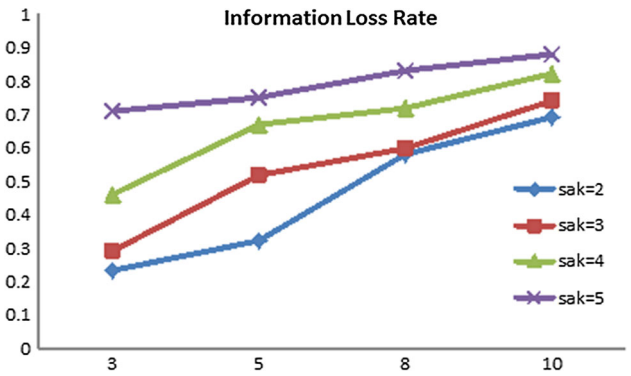

Fig. 5 Information loss rate

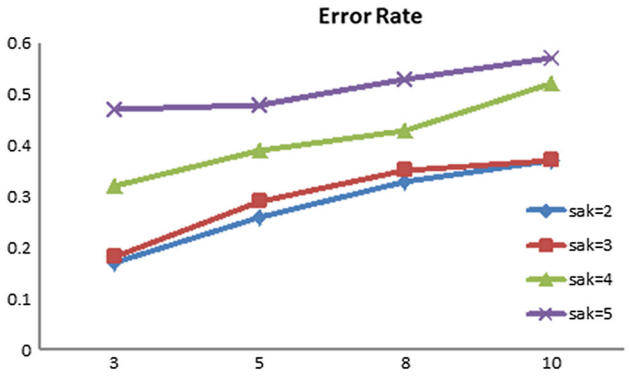

Fig. 6 Query error rate

4.3 Query error rate

The error rate of the spatial range query is also an important measure of information loss. The so-called spatial range count query means querying the number of moving objects in a certain spatial area within a certain period of time. It will inevitably produce a certain error after the semantic trajectories are anonymously processed. The error is represented by *error* and is obtained by calculating Eq. 13.

$$error = \frac{\min(Q(D), D(D^*))}{\max(Q(D), D(D^*))} \tag{13}$$

where $Q(D)$ is the value obtained by performing a spatial range count query on the original trajectory data and $Q(D^*)$ is the value obtained by performing a spatial range count query on the data after privacy protection processing. The query error rate is shown in Fig. 6.

It can be seen from the figure that the query error rate is less than 20% in the case of k = 3 and sak is 2 or 3. In addition, the error rate increases with the increase in the k value. Using the k-anonymity model generally protects semantic trajectory privacy. Considering the computational efficiency and query accuracy, the k value usually ranges between 3 and 5.

5 Conclusion

With regard to publishing trajectory data, this paper proposes a privacy preserving method that adopts semantic trajectory anonymizing based on the k-anonymity model. In contrast to traditional trajectory data models, which only contain the spatiotemporal attributes, our semantic trajectory model incorporates semantic information from sensitive points and users' motion modes. The algorithm first preprocesses the raw data and extracts spatiotemporal sequences, important spatial points, velocities and motion modes. Sensitive points are processed based on the k-anonymity model, eventually forming a coverage area that contains $k-1$ POI points of a similar type to the sensitive points that form a sensitive area. Trajectory ambiguity is accomplished based on the motion modes, road network topologies and road weights in the sensitive area. Finally, a similarity comparison is performed to form an anonymity set that contains the other $k-1$ trajectories with the highest similarity. The experimental results show

that the method performs efficiently and provides an outstanding level of privacy.

Acknowledgements The authors would like to thank the reviewers for their invaluable comments and suggestions, which greatly helped to improve the presentation of this paper.

Funding This research was sponsored by the Shanghai Natural Science Fund (Nos. 14ZR1429800, 15ZR1430000), the Research Fund of the National 12th Five-Year Education Plan (No. EIA140412), and the Zhejiang Province medical and health science and technology platform Project No. 2017KY497.

Compliance with ethical standards

Conflict of interest The authors declare that they have no conflicts of interest.

Ethical approval This article does not contain any studies performed with human participants or animals by any of the authors.

Informed consent Informed consent was obtained from all individual participants included in the study.

References

1. Gao, H. H., Huang, W. Q., Yang, X. X., Duan, Y. C., & Yin, Y. Y. (2018). Towards service selection for workflow reconfiguration: An interface-based computing. *Future Generation Computer Systems, 28,* 298–311.
2. Cranshaw, J., Toch, E., Hong, J., Kittur, A., & Sadeh, N. (2010) Bridging the gap between physical location and online social networks. In *12th International conference on ubiquitous computing* (pp. 119–128).
3. Gao, H. H., Duan, Y. C., Miao, H. K., & Yin, Y. Y. (2017). An approach to data consistency checking for the dynamic replacement of service process. *IEEE Access, 5*(1), 11700–11711.
4. Yin, Y. Y., Chen, L., Xu, Y. S., & Wan, J. (2018). Location-aware service recommendation with enhanced probabilistic matrix factorization. *IEEE Access, 6,* 62815–62825.
5. Cho, E., Myers, S. A., & Leskovec, J. (2011). Friendship and mobility: User movement in location-based social networks. In *ACM SIGKDD international conference on knowledge discovery and data mining* (pp. 1082–1090).
6. Yin, Y. Y., Xu, Y. S., Xu, W. T., Gao, M., Yu, L. F., & Pei, Y. J. (2017). Collaborative Service selection via ensemble learning in mixed mobile network environments. *Entropy, 19*(7), 358.
7. Yin, Y. Y., Yu, F. Z., Xu, Y. S., Yu, L. F., & Mu, J. L. (2017). Network location-aware service recommendation with random walk in cyber-physical systems. *Sensors, 17*(9), 2059.
8. Yin, Y. Y., Song, A. H., Gao, M., Xu, Y. S., & Wang, S. P. (2016). QoS prediction for Web service recommendation with network location-aware neighbor selection. *International Journal of Software Engineering and Knowledge Engineering, 26*(4), 611–632.
9. Yoon, H., Zheng, Y., Xie, X., & Woo, W. (2012). Social itinerary recommendation from user-generated digital trails. *Personal and Ubiquitous Computing, 16*(5), 469–484.
10. Xu, J. J., Zheng, K., Chi, M. M., Zhu, Y. Y., Yu, X. H., & Zhou, X. F. (2015). Trajectory big data: Data, applications and techniques. *Journal on Communications, 36*(12), 97. https://doi.org/10.11959/j.issn.1000-436x.2015318.
11. Shang, S., Zheng, K., Jensen, C. S., Yang, B., Kalnis, P., Li, G. H., et al. (2015). Discovery of path nearby clusters in spatial networks. *IEEE Transactions on Knowledge and Data Engineering, 27,* 1505–1518. https://doi.org/10.1109/TKDE.2014.2382583.
12. Li, S., & Peter, R. S. (2014). Review of GPS travel survey and GPS data-processing methods. *Transport Reviews*. https://doi.org/10.1080/01441647.2014.903530.
13. Gao, P., Kupfer, J. A., Zhu, X., & Guo, D. (2016). Quantifying animal trajectories using spatial aggregation and sequence analysis—A case study of differentiating trajectories of multiple species. *Geographical Analysis, 48*(3), 275–291.
14. Gao, H. H., Zhang, K., Yang, J. H., Wu, F. G., & Liu, H. S. (2018). Applying improved particle swarm optimization for dynamic service composition focusing on quality of service evaluations under hybrid networks. *International Journal of Distributed Sensor Networks (IJDSN), 14*(2), 1–14.
15. Yuan, J., Zheng, Y., & Xie, X. (2013). Discovering regions of different functions in a city using human mobility and POIs. In *ACM SIGKDD international conference on knowledge discovery and data mining*. ACM186-194. https://doi.org/10.1145/2339530.2339561.
16. Zheng, Y. (2015). Trajectory data mining: An overview. *ACM Transactions on Intelligent Systems & Technology, 6*(3), 1–41. https://doi.org/10.1145/2743025.
17. Gao, H. H., Miao, H. K., Liu, L. L., Kai, J. Y., & Zhao, K. (2018). Automated quantitative verification for service-based system design: A visualization transform tool perspective. *International Journal of Software Engineering and Knowledge Engineering (IJSEKE), 28*(10), 1369–1397.
18. Gedik, B., & Liu, L. (2008). Protecting location privacy with personalized k-anonymity: Architecture and algorithms. *IEEE Transactions on Mobile Computing, 7*(1), 1–18.
19. Zhang, X. J., Gui, X. L., & Wu, Z. D. (2015). Privacy preservation for location-based services: A survey. *Journal of Software, 26*(9), 2373–2395. https://doi.org/10.13328/j.cnki.jos.004857.
20. Niu, B., Li, Q., Zhu X., et al. (2014). Achieving k-anonymity in privacy-aware location-based services. In *IEEE conference on computer* (pp. 754–762). IEEE. https://doi.org/10.1109/INFOCOM.2014.6848002.
21. Chow, C. Y., & Mokbel, M. F. (2011). Trajectory privacy in location-based services and data publication. *ACM SIGKDD Explorations Newsletter, 13*(1), 19–29.
22. Zheng, H., & Meng, X. F. (2011). A survey of trajectory privacy-preserving techniques. *Chinese Journal of Computers, 34,* 1820–1830. https://doi.org/10.3724/SP.J.1016.2011.01820. **(in Chinese with English abstract)**.
23. Gao, H. H., Mao, S. Y., Huang, W. Q., & Yang, X. X. (2018). Applying probabilistic model checking to financial production risk evaluation and control: A case study of Alibaba's Yu'e Bao. *IEEE Transactions on Computational Social Systems (TCSS), 5*(3), 785–795.
24. Liu, L. (2007). From data privacy to location privacy: Models and algorithm. In *Proceedings of the 33rd international conference on very large data bases* (pp. 1429–1430).

25. Nergiz, M. E., Atzori, M., Saygin, Y., & Güc, B. (2009). Towards trajectory anonymization: A generalization based approach. *Transactions on Data Privacy, 2*(1), 47–75.
26. Chris, Y. T. M., David, K. Y. Y., Nung, K. Y., & Nageswara, S. V. R. (2010). Privacy vulnerability of published anonymous mobility traces. In *Proceedings of the 16th annual international conference on mobile computing and networking* (pp. 185–196).
27. Terrovitis, M., & Mamoulis, N. (2008). Privacy preserving in the publication of trajectories. In *Proceedings of the 9th international conference on mobile data management* (pp. 65–72).
28. Zheng, H., Meng, X. F., Hu, H. B., & Yi, H. (2012). *You can walk alone: Trajectory privacy-preserving through significant stays protection, database systems for advanced applications* (pp. 351–366). Berlin Heidelberg: Springer. https://doi.org/10.1007/978-3-642-29038-1_26.
29. You, T. H., Peng, W. C., & Lee, W. C. (2007). Protecting moving trajectories with dummies. In *Proceedings of the international workshop on privacy-aware location based mobile services*.
30. Richter, K. F., Schmid, F., & Laube, P. (2012). Semantic trajectory compression: Representing urban movement in a nutshell. *Journal of Spatial Information Science, 4*(4), 3–30. https://doi.org/10.5311/josis.2012.4.62.
31. Ying, J. C., Lee, W. C., Weng, T. C., & Tseng, V. S. (2011). Semantic trajectory mining for location prediction. *ACM Sigspatial International Symposium on Advances in Geographic Information Systems*. https://doi.org/10.1145/2093973.2093980.
32. Elragal, A., & El-Gendy, N. (2013). Trajectory data mining: Integrating semantics. *Journal of Enterprise Information Management, 26*(5), 516–535. https://doi.org/10.1108/JEIM-07-2013-0038.
33. Monreale, A., Pinelli, F., Trasarti, R., & Giannotti. F. (2009). WhereNext: A location predictor on trajectory pattern mining. In *Proceedings of the 15th ACM SIGKDD international conference on knowledge discovery and data mining* (pp. 637–646). ACM.
34. Lee, W. C., & Krumm, J. (2011). *Trajectory preprocessing. Computing with spatial trajectories* (pp. 3–33). New York: Springer. https://doi.org/10.1007/978-1-4614-1629-6_1.
35. Gao, H. H., Chu, D. Q., & Duan, Y. C. (2017). The probabilistic model checking based service selection method for business process modeling. *Journal of Software Engineering and Knowledge Engineering, 27*(6), 897–923.
36. Giannotti, F., Nanni, M., Pedreschi, D., & Pinelli, F. (2007). Trajectory pattern mining. In *Proceedings of the 13th ACM SIGKDD international conference on knowledge discovery and data mining* (pp. 330–339).

Rong Tan is a Lecturer of IOT (Internet of things) engineering Dept. at Shanghai Business School. After receiving his Ph.D. from East China Normal University in 2013, he worked as a postdoctoral research associate at Beijing Tehua postdoctoral programme. He joined Shanghai Business School in 2013. His research area includes spatial–temporal database, NoSQL and blockchain.

Yuan Tao received the Ph.D. degree in Computer Science from Shanghai University, in 2011. She is currently an Assistant Professor with the Computing Center of Shanghai University. Her research interests are in the areas of intelligent information processing, artificial intelligent and pattern recognition.

Wen Si is Associate Professor of IOT (Internet of things) engineering Dept. at Shanghai Business School. After receiving his Ph.D. from Shanghai University in 2011, he worked as a post-doctoral research associate at Fudan University's Rehabilitation Medicine in Huashan Hospital. He joined Shanghai Business School in 2011. His research area includes Biomedical engineering and Internet of things technologies.

Yuan-Yuan Zhang is an Assistant Professor at College of Information Technology of Zhejiang Chinese Medical University. Her research includes Medical Middleware and Web Information Mining.