[1,*]Yu Qiao

[2]Hao Ji

A Privacy Protection Method Based on Trajectory Location Point Association

Abstract: - Location-based services (LBS) technology provides personalized services for users. However, private location information faces the risk of being leaked while users are enjoying the LBS services. In order to prevent this from occurring, this study proposes an anonymous privacy protection method based on trajectory location association, abbreviated as AS-AP, which considers the association between the user's current location and the background information. The association between locations is mined from time and query preferences, and then fake locations and queries are generated as part of the anonymous interference requests. This conforms to the k-anonymity feature, which prevents attackers from inferring the user's real intent from the network information, thereby enhancing the credibility of network technology, and protecting data privacy. The experimental results indicate that AS-AP can provide more personalized privacy protection services while ensuring that users' sensitive data is protected.

Keywords: Trajectory Data, Privacy Protection, Feature Correlation, Location Service Quality.

I. INTRODUCTION

As a result of the development of the Internet and the popularity of intelligent terminals, social networking has become a part of daily life. LBS (Location-based services, LBS) fulfill personalized requests [1-2], such as finding a nearby bank or navigating a route home. The process for accomplishing such tasks is as follows: 1) the user submits their current position and points of interest to the LBS provider, 2) the LBS provider processes and analyzes the information, and 3) the final result is sent back to the user. This allows the user to easily obtain relevant information about points of interest. In this process, certain aspects of the location information are regarded by users as irrelevant data and are sent to the service provider (SP) for free in exchange for relevant services. If such information is maliciously disclosed or used, the security of the user's personal information is jeopardized [3]. Therefore, protecting the user's location privacy while providing LBS has become an important research topic.

II. RELATED WORK

In the context of the widespread application of LBS technology, the protection of the user's private trajectory data has become an important issue [4]. In the early stages of the development of trajectory data protection scenarios, anonymity technology was employed. The most common type of this technology is k-anonymity, which prevents observers from identifying real information using confusion methods. For example, by sending a user's real intent and k-1 interference information to the SP, the SP cannot determine which of the k requests it received is real.

Based on LBS service architecture, anonymous technology can be divided into two types: service center and mobile terminals. Using an anonymous service as the center involves anonymizing data between the SP and the terminal and sending the anonymized requests to the SP together with the real requests. This design can reduce the computational complexity and storage overhead required for the user, but it also requires the anonymous center to process a large number of requests, resulting in a significant reduction in processing efficiency. Furthermore, if the anonymous center is maliciously hacked, a large amount of private data will be leaked [5-6].

The continuous development of intelligent terminal technology can solve these problems by providing better computing and storage capabilities. Therefore, the functions of an anonymous center can be adapted to mobile terminals. For example, the CacheCloak scheme [7] caches service information requested by users to avoid multiple interactions with the SP, which protects the sensitive information by reducing the number of interactions. The scheme designed and implemented by Niu [8] considers the historical query probability of each location in the background information. Simultaneously, in order to avoid low-quality privacy protection, the distance between anonymous locations is calculated to ensure that the selected locations are as distant from each other as possible.

[1] Nanjing Tech University Pujiang Institute, Nanjing, China

[2] Pinduoduo Inc, Shanghai, China

*Corresponding author: Yu Qiao

The privacy information in the track can be divided into the user's sensitive positions and private query content. For the privacy data contained in the user's query content, the l-diversity scheme [9] and t-progressive scheme [10] have evolved to support privacy protection algorithms based on anonymity. The l-diversity scheme [9] is an optimized privacy access method, which utilized the path and random model of users to analyze their spatial diversity and achieved a high level of privacy protection. In addition, to protect sensitive attributes, Chen [11] analyzed the limitations of l-diversity scheme and constrained the distribution distance of the sensitive attributes to be less than the preset threshold value.

Existing privacy protection schemes have the following two difficulties. First, they require an efficient sampling mechanism to collect user trajectory data. Second, after obtaining the real trajectory data of the user, a perturbation mechanism is required to resist external attacks. Therefore, it is necessary to improve the statistical accuracy and availability of the released data, but there is no scheme for solving these problems simultaneously.

In view of this situation, this study proposes an anonymous privacy protection method based on the association relationship between track positions. This method uses track data generated by the onboard system as background, generates interference positions and queries by mining the potential relationship between the positions, and sends them to the SP together with the real request. This meets the user’s need for personalized location services while protecting the privacy of their sensitive data. The contributions of this study are as follows:

(1) While considering the relevance of background knowledge and the personal information, we further mined the relationships between locations from the dimensions of time and query scope.

(2) In this scheme, the user's location information and query contents are personalized simultaneously, considering the existence of the background information.

(3) The results of experiments showed that the service quality of AS-AP relative to Ba-2PS[12] and spati[13] improved by 34.9% and 19.2% on average.

III. Theories

A. LBS Architecture Model

The basic architecture of an LBS primarily consists of four parts: GPS satellites, mobile terminals, communication base stations, and SPs (as shown in Figure 1). The entire process followed in each request is explained as follows. First, the current geographic location of the intelligent terminal is determined by the GPS satellites (step 1). In the process of acquiring these data, the privacy protection algorithm does not need to be considered, and it is safe by default. After receiving the location data (step 2), together with the query content and other information, they are transmitted to the LBS server as a user request through the communication base station (steps 3 and 4). The LBS server then processes the data and sends the results back to the personal terminal (steps 5 and 6). The communication base station only forwards the user's requests and receives responses, but it makes no changes to the requests.

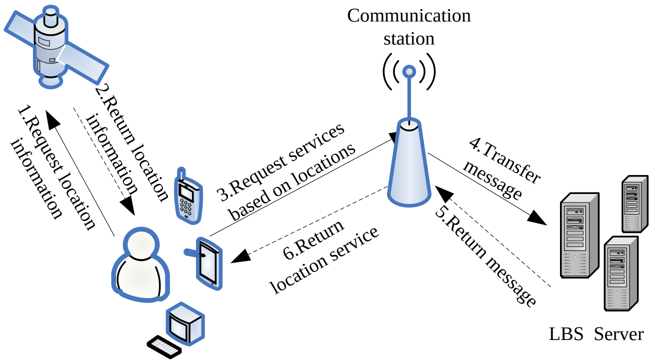
Communication station
2.Return location information
1.Request location information
3.Request services based on locations
6.Return location service
4.Transfer message
5.Return message
LBS Server

Figure 1: LBS Architecture and Process Followed for Each User Request

B. Definitions

Definition 1 (point of interest, or POI): a location with characteristics that distinguishes it from coordinate information [12]. For example, if a user requests the locations of banks surrounding the parking lot near a scenic area, the parking lot is the POI, the coordinates are the location information, and the locations of the banks form the query content.

Definition 2 (query probability): the probability of each position in the map that has been requested in the history. In this study, the query probability was confined to the region that we obtained through the Google Maps application programming interface (API), and the background information was stored in the intelligent terminal.

Definition 3 (background knowledge): owing to the development of map positioning and navigation technology, any entity with calculation and storage functions can obtain historical queries pertaining to a location, including the records of requests that have been submitted or are currently pending, the query probability of the area, and the POI.

Definition 4 (k-anonymity mechanism): when a user requests an LBS, it sends a real location and a query to a trusted anonymous server. To protect the user's privacy, the anonymous service removes the user's personal tag, generates an anonymous area that contains at least k-1 other users, and then sends this information to the LBS provider together with the real request [13].

Definition 5 (privacy measurement based on the *k*-anonymity mechanism): if the selected anonymous location area is $R=<loc_1, loc_2, \ldots, loc_r>$, that is, $|R|=r$, then the value of k is the number of location units in the area ($k=|R|=r$). Therefore, the probability of the actual request being disclosed is P(disclosure)$=1/k=1/r$.

Definition 6 (privacy measurement based on information entropy): if the real location of the user is denoted as *u*, and *k-1* anonymous locations are obtained through the *k*-anonymity algorithm, the information entropy *H* can be expressed as

$$H=-\sum_{i}^{k} P_i \log P_i \,, \tag{1}$$

where P_i represents the probability of selecting *i* as the anonymous area according to the user's real location *u*. When the probability of the event occurring is the same, the information entropy reaches its maximum value and the uncertainty of the probability event is the highest. Thus, the degree to which the user's privacy is protected is the highest.

IV. Privacy Protection Method Based on Trajectory Location Association: AS-AP

A. *Description of AS-AP*

The existing privacy protection schemes consider too much about historical data, without paying attention to the query probability of a specific time. According to the definition of background knowledge in Section 3.2, it includes not only historical data, but also the query probability of current locations, and even the information of interest points used to distinguish locations. The specific steps are as follows:

(1) User definition: users define their own sensitive locations according to the actual situation, including target locations, and error range that they can accept.

(2) Preprocessing of the track data: the navigation route is divided on the map, the probability $P(loc_u)$ that the current location loc_u may be queried according to the background information is obtained, and the locations that satisfy $|P(loc_u)-P(loc_i)|<\delta$ ($\delta \geq 0$)are filtered out and placed in the collection θ.

(3) Generation of interference locations: secondary filtering is conducted on the basis of step (1) in order to select location units that are as distant as possible from each other, which ensures that the generated interference positions are decentralized. This step uses a quadtree structure for data storage.

(4) Generation of anonymous query content: the potential relationship between locations and queries are analyzed from the perspective of time, and then the users' query preferences are mined at different times to determine the anonymous query content.

(5) Completion of interference requests: the query radius corresponding to each anonymous request is redefined to prevent attackers from using background information when speculating that the query radius of the anonymous location units is too large or too small. This strategy improves the effective number of anonymous locations and ensures that the user’s sensitive location information is not disclosed.

B. *Preprocessing of Trajectory Data*

First, the original data in the trajectory dataset are preprocessed such that the corresponding position distribution is drawn according to the coordinate information in the source file. Then the map area is initialized by dividing it into n location units, which can be described as:

$$\mathrm{m}ap=\{loc_1, loc_2, loc_3, \cdots, loc_i, \cdots, loc_\mathrm{n}\} \,, \tag{2}$$

where n is the granularity of the map area, loc_i represent the i^{th} location unit, and i and $n \in N$. Each position in the map is considered as the basic unit for subsequent processing.

The location of user u at time t_u is denoted as loc_u, the user's request radius is r_u, and $query_u$ represents the user’s request, which can be expressed as $Req=<u, t_u, loc_u, r_u, query_u>$. The probability $P(loc_u)$ that the loc_u may be queried according to the background information is obtained, other locations in the map are traversed, and the loc_i

with a query probability similar to that of loc_u is chosen. That is, the location points satisfy $|P(loc_u)-P(loc_i)|<\delta$ ($\delta \geq 0$), and they are stored in the set θ. The specific steps are described in Algorithm 1.

```
Algorithm 1. Algorithm for preprocessing trajectory data.
Input: original coordinate data and the user's current position $loc_u$.
Output: set $\theta$.
Step 1: draw the corresponding distribution points in the map area according to the coordinate information in the original dataset.
Step 2: divide the drawn area into $n$ position units.
Step 3: query the background information and determine the $loc_i$ with a query probability similar to that of the current location $loc_u$.
Step 4: if $loc_i$ satisfies the condition $|P(loc_u)-P(loc_i)|<\delta$, execute step 5; otherwise, $loc_i$ is not processed.
Step 5: store the selected location points $loc_i$ into set $\theta$.
```

C. Method for Generating Interference Locations

If the location in set θ is in close proximity to the user's real position (e.g., both inside a building), the user’s privacy cannot be adequately protected. Therefore, it is necessary to further filter the appropriate anonymous locations.

Suppose that the user caches the query records during the period ΔT, which the records are denoted as $query_u$=<u, t_i, loc_i, r_i, $query_i$>, where $|t_i - t_u|<\Delta T$ $(\Delta T > 0)$, and are stored in set γ. Then, the intersection between sets γ and θ is determined. Next, the elements from the intersection are deleted to form a new set θ_1. After these operations, the number of locations in θ_1 will not exceed the original set θ.

The secondary screening ensures that the generated interference locations are as dispersed as possible, the locations with the same query probability and a sufficient distance between them are selected. To implement these operations, we used the quadtree structure to store the data. When iteratively dividing the map area according to the background information, each section of the map is divided into quarters and stored accordingly. Figure 2 shows the $map=\{loc_1,loc_2,loc_3,\ldots, loc_n\}$, where $n=4a$ and $a \in N^+$.

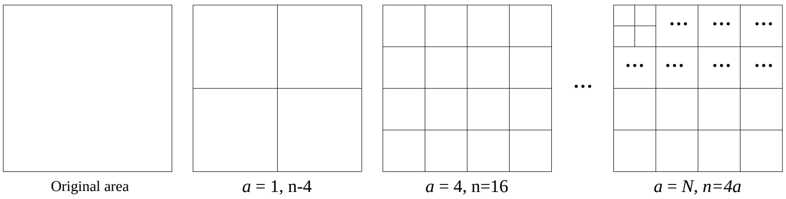

Figure 2: Schematic Diagram of Regional Divisions

```
Algorithm 2. Algorithm for generating jamming locations.
Input: set $\theta$ of location units, query record collection $\gamma$, and value of $k_{\max}$
Output: set $\theta_1$
Step 1: initialize $i$=0 and start traversing.
Step 2: if $\theta \cap \gamma \neq \varnothing$, that is, if the existing $loc_i \in \theta$ && $loc_i \in \gamma$, delete $loc_i$ from set $\theta$.
Step 3: execute $i$=$i$+1;
Step 4: if $\theta \cap \gamma = \varnothing$, that is, if no $loc_i$ satisfies the requirement $loc_i \in \theta$ && $loc_i \in \gamma$, go to step 3.
Step 5: obtain the parent node of the current position of user $u$, and select $node_1$, $node_2$, and $node_3$ with the same depth as $node_u$.
Step 6: when the value of $k$-1 is not 0, proceed to step 7; otherwise, execute step 9.
Step 7: traversal collection $\theta$; set $loc_i$ as an anonymous location unit and add it to $\theta_1$ if $loc_i$ is a child node among $node_1$, $node_2$ and $node_3$; then execute $k=k-1$.
Step 8: once $k \leq 1$, execute step 9; otherwise, repeat step 7.
Step 9: output $\theta_1$.
```

To ensure that the selected locations are as decentralized as possible, the selected anonymous locations all originate from different parent nodes, and the depth traversal of the branches to which they belong satisfies $Dep(loc_i) \geq \mu$. Finally, the k-1 positions are considered the interference locations. Considering the complexity of the implementation of the algorithm, three nodes with the same depth were selected in Algorithm 2.

According to Algorithm 2, the selected k-1 anonymous locations have the following characteristics: (1) they are generated based on the user's background information; (2) the query probability is similar to the current location;

(3) the selected locations are scattered in the map and are not adjacent to each other; and (4) they do not belong to the set of records that the user has queried during the period ΔT. To protect the user's query privacy, the requested information sent to the LBS cannot be the same as the user's real content. It is also necessary to avoid sending requests that are impossible to fulfill at a specific point in time (such as querying a nearby vegetable market at 1:00 am).

Algorithms 1 and 2 describe the process of generating anonymous locations that only need to consider the relationship between the background information and user information. To improve the effect of interference on anonymous information, the AS-AP method further considers the impact of a user's query content and query radius on generating anonymous requests; the latter are denoted as $Req_{anonymous}=<UID, t, loc_{anonymous}, r_u, query_{anonymous}>$.

Next, the relationship between locations and query content from the perspective of time are analyzed, and then completely anonymous query content is generated. The query content requested by user u at loc_u is recorded as $query_u$, and there is an anonymous location with the same query probability in set θ_1, which is denoted $loc_{anonymous}$. According to the background information, the anonymous $\mathrm{query}_{\mathrm{anonymous}}$ with a high query probability and the probability of the content being queried satisfies $P_{loci}(query_{anonymous}) > \mu$, $0 < \mu < 1$. The details are as follows.

Algorithm 3. Algorithm for generating anonymous query content.

```
Input: set $\theta_1$, background information
Output: set $\alpha$

Step 1: initialize $i$=1, begin traversing the background information.
Step 2: if $P_{loci}(query_{anonymous}) > \mu$ ($0 < \mu < 1$, where $\mu$ is the mean value of the query probability of the current location point), then insert $query_{anonymous}$ into the collection of anonymous query contents $\alpha$; otherwise, continue.
Step 3: execute $i=i+1$; if $i \le k-1$, repeat Step 2; otherwise, end.
Step 4: output the collection of anonymous query contents $\alpha$.
```

D. Implementation of Algorithm for Generating Anonymous Requests

In Section 4.3, the map was stored using a quadtree structure (Figure 3). Assuming k=4, three anonymous locations are selected and marked as loc_1, loc_2, and loc_3. The depths of these locations are marked as $Dep(loc_1)$, $Dep(loc_2)$, $Dep(loc_3)$, and $Dep\ (loc_u)$. Then $Dep(loc_i)$ is compared to μ. If $Dep(loc_i) \geq \mu$, the query probability of this region is relatively high, and thus the query radius is redefined as $r_i=r_i+\beta$, where $\beta<0$. Conversely, $r_i=r_i+\beta$, where $\beta\geq 0$.

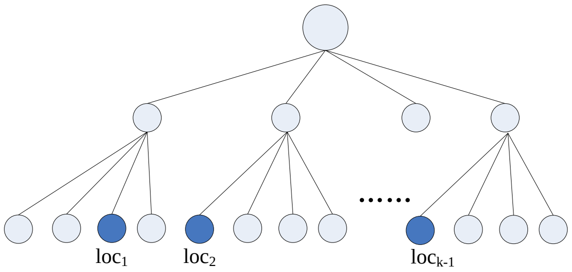

Figure 3: Quadtree Structure

Finally, the obtained query radius, corresponding anonymous location, and query content are reconstituted into an anonymous request, $Req_{anonymous}=<UID, t, loc_{anonymous}, r_u, query_{anonymous}>$, and sent to the LBS together with the user's real request.

V. Experiments

A. *Data Set and Experimental Environment*

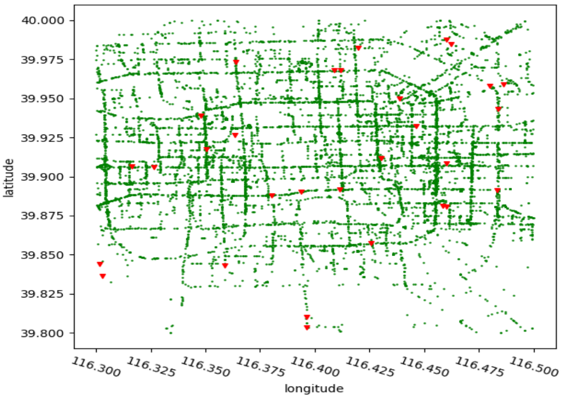

Figure 4: Example of User Trajectory Distributions

To verify the effectiveness of the privacy protection of the AS-AP method proposed in this study, we conducted an experimental simulation on data extracted from the T-Drive dataset in a Python environment. The dataset contains some GPS track data of 10357 taxis in Beijing in 2008 and is commonly used for research work related to track data privacy protection. There are about 15 million locations, sampling every 170s (the average distance is about 620m), and the data set records various outdoor activities of users, including shopping, sports, going to work and going home. Each track is marked by a series of coordinate points, such as the user's ID, timestamp, latitude and longitude. Figure 4 shows the example of user trajectory distributions, which longitude ranges from 116.3 to 116.5 and latitude ranges from 39.8 to 40.

B. *Experimental Verification*

1) *Relationship between parameter k, parameter δ, and execution time*

The parameter *k* in the experiment refers to the number of requests sent to the SP based on the *k*-anonymous method, and it has a range of [3, 50]. The parameter *δ* is the probability threshold used to filter anonymous locations. Figure 5 shows the execution time of the AS-AP method as a function of the parameter *k* for *δ=0.05* and *δ=0.15*, which proves that the larger the value of *δ*, the longer the entire algorithm will take to execute because *δ* determines the deviation between the query probability of the selected anonymous interference location and that of the real location. The larger the value of *δ*, the higher the number of location units that satisfy $|P(loc_u) - P(loc_i)| < \delta$, and thus the longer the algorithm takes to execute.

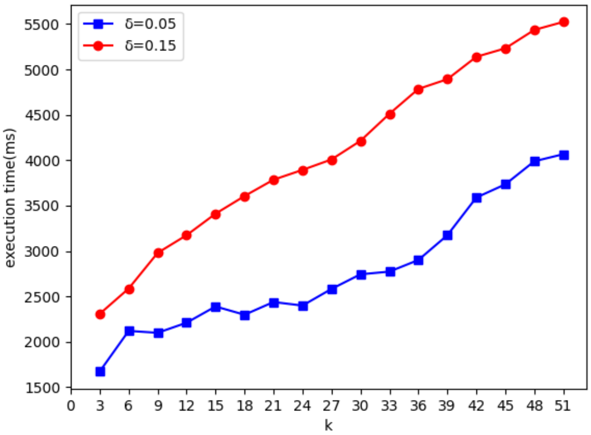

Figure 5: Effect of Parameter K on Execution Time for δ=0.05 (Blue) and δ=0.15 (Red).

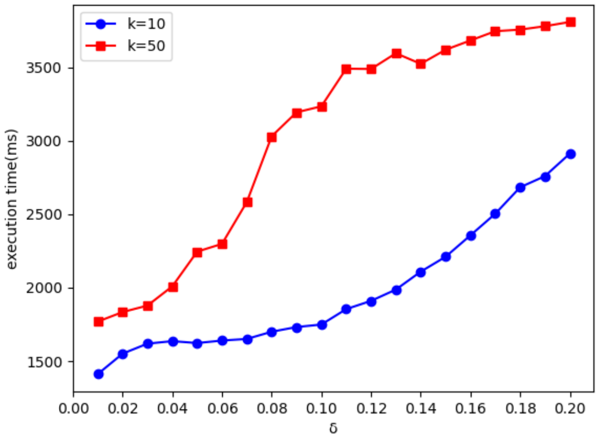

Figure 6: Effect of Parameter δ on Execution Time for k=10 (Blue) and k=50 (Red).

The influence of the value of δ on the execution time is shown in Figure 6, where k=10 and k=50 was selected to represent the time that elapsed between sending the request and receiving the feedback information. The size of the request depends on the value of k, which has no direct relationship with δ. However, δ does affect the number of elements in θ; the higher the value of δ, the higher the number of elements in set θ, and the longer it takes to iterate the algorithm.

2) Degree of privacy protection

The effectiveness of the privacy protection of the AS-AP algorithm was evaluated from two perspectives: the privacy leakage probability and the degree of privacy protection. Additionally, it was compared to Ba-2PS[12] and spati[13] to verify the security of the algorithm.

a) Privacy leakage probability

The abovementioned experimental results indicate that the larger the value of k, the more request information is sent to the SP, and the lower the probability that the attacker can infer the true information. Therefore, in this part of the experiment, a random value was selected as the baseline; the theoretical optimal value, Ba-2PS, spati, and AS-AP methods were chosen as the verification object; and the leakage probability was evaluated as a function of k, as shown in Figure 7. The experimental results show that the larger the value of k is, the more data requests are sent to the LBS, and the lower the probability that the attacker can speculate real information.

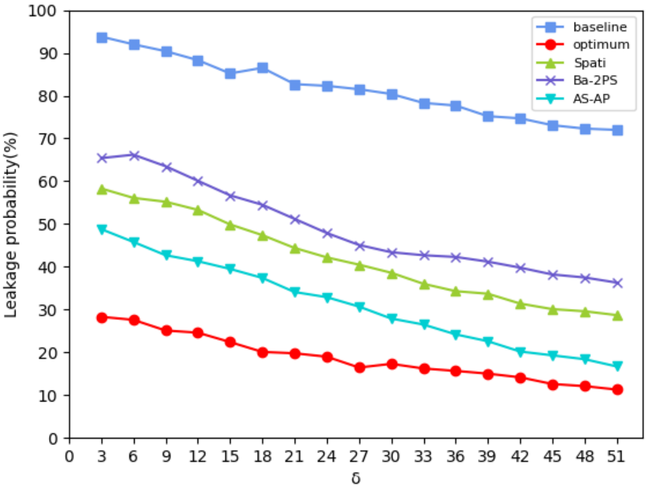

Figure 7: Leakage Probability as a Function of Parameter k.

b) Degree of privacy protection

The AS-AP method proposed in this study personalizes the user's request range when constructing anonymous interferences, reducing the possibility that attackers can use the query range and location unit to discover real information. Therefore, the AS-AP method improves the uncertainty in the user’s location privacy, and thus effectively protects sensitive information. In AS-AP, information entropy is used as a measure of user location and query content privacy. The specific formula is shown in Eq. (1), the greater the entropy value, the better the privacy protection.

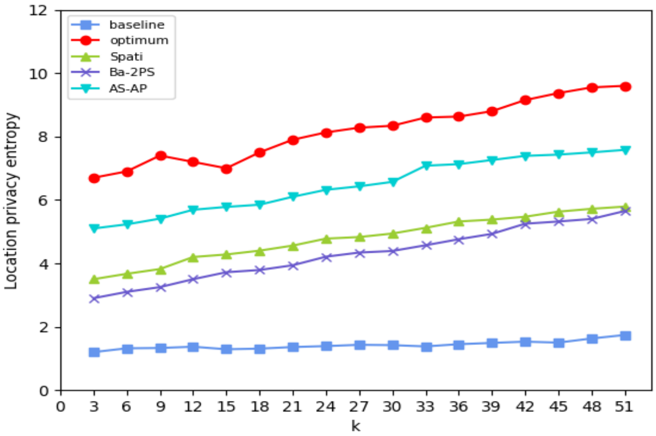

Figure 8: Location Privacy Performance of Different Algorithms

The experimental results are shown in Figure 8 and 9. The spati algorithm considers the correlation between time and space, which increases the difficultly for attackers to infer the user's real information. Therefore, the level of location privacy protection of this scheme, as demonstrated in Figure 8, is better than that of k with random values. Compared to the spati and Ba-2PS algorithm, the AS-AP method exhibited an average improvement in location privacy of 34.9%. The spati algorithm does not involve the protection of query privacy; as a result, its performance (as shown in Figure 9) was the worst. Compared to the Ba-2PS algorithm, the query privacy protection performance of AS-AP improved by 19.2% on average. This is because the AS-AP method considers the time factor when selecting anonymous location units while considering the request scope, thereby avoiding the possibility of using spatiotemporal correlation factors to infer the user's real location and query content.

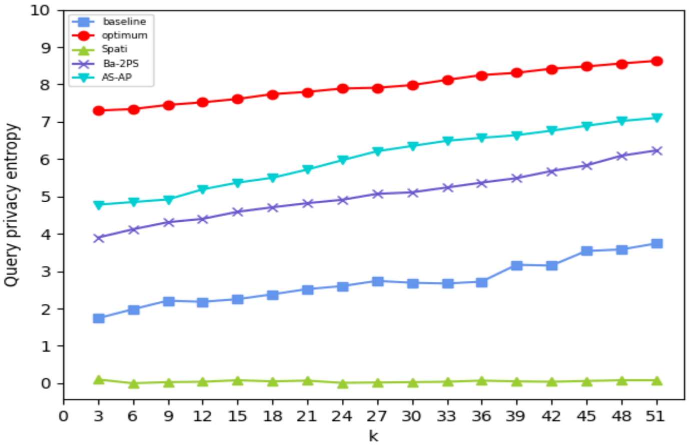

Figure 9: Query Privacy Performance of Different Algorithms

VI. Summary and Conclusions

In LBS, both users and LBS SPs store user-request information for a certain period of time. Therefore, an attacker can overcome anonymity by analyzing historical query information. This study presented the AS-AP method, a privacy protection method based on trajectory location point association. When selecting anonymous location units, AS-AP considers the association of the user's query history to prevent attackers from making inferential attacks. In addition to considering the user's basic information, AS-AP also considers the query probability and query scope, and further considers both the space and time coordinates where users are located to improve the protection of user location privacy and query privacy. Finally, the validity and security of the scheme were verified through a theoretical explanation and an experimental simulation.

The privacy protection algorithm in this paper still has certain limitations, such as not considering the impact of other attributes in the trajectory on privacy leakage, such as the acceleration and direction of movement; In addition, there is insufficient consideration in real-time protection. Next, we will further improve the location privacy protection model to provide better LBS services for daily life.

ACKNOWLEDGMENT

The authors appreciate the Natural Science Research of Jiangsu Higher Education Institutions (No.21KJB520028).

REFERENCES

[1] XU C, LUO L, DING Y, et al. Personalized Location Privacy Protection for Location-Based Services in Vehicular Networks. IEEE Wireless Communications Letters, 2020, 9(10): 1633-1637.

[2] TAN Z, WANG C, YAN C, et al. Protecting Privacy of Location-Based Services in Road Networks. IEEE Transactions on Intelligent Transportation Systems, 2021, 22(10): 6435-6448.

[3] CHEN J, HE K, YUAN Q, et al. Blind filtering at third parties: An efficient privacy-preserving framework for location-based services. IEEE Transportations on Mobile Computing, 2018,17(11): 2524-2535.

[4] WANG J, WANG C R, MA J F, et al. Dummy location selection algorithm based on location semantics and query probability. Journal on Communications, 2020, 41(3): 53-61.

[5] WANG J, LI Y, YANG D, et al. Achieving effective k-anonymity for query privacy in location-based services. IEEE Access, 2017, 5:24580-24592.

[6] LIU H, ZHANG S, LI M, et al. An Effective Location Privacy-Preserving K-anonymity Scheme in Location Based Services. 2021 IEEE International Conference on Electronic Technology, Communication and Information (ICETCI), Changchun, China, 2021: 24-29.

[7] MA C S, YAN Z S, CHEN C W. SSPA-LBS: Scalable and Social-Friendly Privacy-Aware Location-Based Services. IEEE Transactions on Multimedia, 2019, 21(8):2146-2156.

[8] NIU B, LI Q, ZHU X, et al. Achieving k-anonymity in privacy-aware location-based services. International Conference on Computer Communications. IEEE, 2014: 754-762.

[9] HE X, JIN R, DAH I. Leveraging spatial diversity for privacy-aware location-based services in mobile networks. IEEE Transactions on Information Forensics and Security, 2018,13(6):1524-1534.

[10] REN W, KAMBIZ G, LIAN X. KT-Safety: Graph Release via k-Anonymity and t-Closeness. IEEE Transactions on Knowledge and Data Engineering, 2022, pp.1-12.

[11] CHEN Z, HU X, JU X, et al. LISA: Location information ScrAmbler for privacy protection on smartphones. 2013 IEEE Conference on Communications and Network Security (CNS), National Harbor, MD, USA, 2013:296-304.

[12] LI W, CAO J, LI H. Privacy self-correlation privacy-preserving scheme in LBS. Journal of Communications, 2019, 40 (5): 57-66.

[13] YANG M X, WU Y T, CHEN Y L. A K-anonymity Optimization Algorithm Under Attack Model. 2022 IEEE International Conferences on Internet of Things (iThings) and IEEE Green Computing & Communications (GreenCom) and IEEE Cyber, Physical & Social Computing (CPSCom) and IEEE Smart Data (SmartData) and IEEE Congress on Cybermatics (Cybermatics), Espoo, Finland, 2022:357-362.