Research on Location Privacy Protection Technology in Wireless Sensor Networks Based on Big Data

Hong Zhang and Pei Li*

Department of Information Engineering, Shanxi Conservancy Technical Institute, Yuncheng 044004, China
E-mail: sxsylp126126@126.com
**Corresponding Author*

Received 25 April 2023; Accepted 22 June 2023;
Publication 17 November 2023

Abstract

The digital footprint in wireless sensor networks can bring great academic and commercial value, but it will also bring the risk of privacy disclosure to users. This study discusses the location privacy protection methods in location based service under multiple scenarios. In the experiment, a false location filtering algorithm for real-time location request scenarios and a false path generation model for offline location release scenarios are proposed. The false position filtering algorithm is implemented based on the idea of a large top heap selection query. The algorithm can construct false position candidate sets and filter false positions. The false track generation model combines the false position technology and the generative adversarial networks model, which mainly protects the user's track data by synthesizing tracks. In the attacker's recognition experiment of a real location. The minimum distance between the false locations generated by the algorithm proposed in the study is above 400 m and the generation time does not exceed 5 ms, generating a better set of false locations in terms of both

Journal of Cyber Security and Mobility, Vol. 12_6, 845–868.
doi: 10.13052/jcsm2245-1439.1262

effectiveness and efficiency. Compared with several commonly used privacy-preserving algorithms, the proposed algorithm has the lowest probability of being identified with real locations, with no more than 21% overall, and is almost independent of the k value. the recognition accuracy of the trajectory user link task decreases from over 90% to about 34%, indicating that the proposed fake trajectory generation model can effectively protect users' data privacy. The experimental results demonstrate that the algorithm and model proposed in the study can quickly generate physically dispersed and semantically diverse sets of fake locations and effectively protect users' trajectory privacy, which is important for users' digital footprint privacy protection.

Keywords: Digital footprint, false position technology, generate confrontation network, synthetic track, TUL task.

1 Introduction

The emergence and popularization of sensor equipment and the Global Positioning System (GPS) enable the digital world to collect the user's daily behavior track and the dynamic changes of the physical world on a large scale. "Digital footprint" refers to all kinds of data generated by the interaction between people and media or things, which are scattered on the Internet, various information systems, and social media [1]. Information such as human behavior, group characteristics, and urban development trends mined from the digital footprints left by users on the Internet can be widely applied to social management, resource management, environmental protection, and public security [2, 3]. These massive digital footprints have brought fantastic academic and commercial value to Internet service providers. However, it also increases the risk of personal privacy data disclosure [4]. Illegal elements can use big data analysis technology to dig out the user's basic identity information, home address, living habits, etc. from the user's digital footprint. Such behavior will cause certain adverse consequences and losses to individuals and society [5, 6]. Currently, the more popular Location Based Service (LBS) makes people more accustomed to exchanging network services with personal information. For example, the telephone address is used for takeout ordering and online shopping, real-time location of taxi travel, etc. [7]. Recently, as the harm caused by privacy disclosure has been gradually taken seriously, people began to hesitate and distrust the use of LBS. This distrust may in turn affect the future development of LBS [8]. Therefore, in the era of big data, building an effective information security management system and

protecting the privacy of digital footprints is an essential part of social development. Regarding the privacy issue based on location services, most existing research focuses on protecting users' location privacy through methods of location perturbation and obfuscation, which typically use privacy metrics such as k-anonymity. However, these schemes do not take into account the Semantic information of the user's location, and cannot guarantee the semantic diversity of constructing false location sets. Currently, research on trajectory privacy protection mainly focuses on three aspects: trajectory generalization, false trajectories, and trajectory suppression. Most of these methods group or mix trajectories of different users, thereby transforming personal trajectory data recognition into k-anonymity problems. However, most of these methods do not consider other information beyond the spatial attributes of trajectory point data, making it difficult to ensure the availability of generated position data. This study discusses location privacy protection and tracks privacy protection, aiming to provide users with a more secure and effective digital footprint privacy protection scheme.

Considering the above, the main contributions of this work include:

- To address the problem that most current fake location privacy protection schemes do not fully consider the background knowledge possessed by attackers, the study innovatively proposes a location privacy protection scheme based on multivariate data by considering the query probability, semantic information and physical distribution of locations.
- For the massive trajectory data generated by the increasing number of LBS services in social networks and the shortcomings of traditional trajectory protection methods. The study proposes a dummy trajectory generation model based on multidimensional feature fusion, which combines dummy location techniques and Generative Adversarial Networks (GAN) models to generate privacy-preserving synthetic trajectory data.

The rest of this paper is structured as follows. Section 2 introduces the research background of this paper as well as an overview and summary of relevant studies at home and abroad, and presents the significance and content of the research in view of the shortcomings of existing studies. Section 3 presents a privacy-preserving technique for digital footprints in wireless sensor networks based on big data, including two parts: a fake location screening algorithm and a fake trajectory generation model. Section 4 experimentally validates the effectiveness of the fake location filtering algorithm to generate fake location sets and the trajectory privacy preserving performance of the fake trajectory generation model. Section 5 provides a concluding description

of the article and presents the shortcomings of the study, giving ideas for improvement and directions for subsequent research.

2 Related Work

Recently, scholars have carried out a lot of research on location privacy protection and put forward many location privacy protection methods that are applicable to different fields.

Wu Z et al. [9] proposed a framework for location privacy protection system and introduced a location privacy model to describe the constraint conditions for constructing masking ranges, achieving more efficient and secure location privacy protection.

Then, Hassan et al. [10] investigated the implementation performance of differential privacy technology in transportation systems and industrial Internet of Things and other applications. They also described the problems, difficulties, and future research directions of differential privacy technology. This research provided a theoretical basis for the solution of physical information system problems and data privacy protection.

Next, Wu et al. [11] proposed an LBS user privacy protection framework, which mainly covered the query location and attributes of users by constructing pseudo-query sequences. Finally, the experiment verified the effectiveness and feasibility of the method from both theoretical analysis and experimental evaluation.

In Zhang [12] and other scholars proposed a user privacy scheme combining cache and spatial K anonymity. The multi-level cache in this scheme could effectively reduce the exposure risk of user information and thus achieved efficient privacy protection with small LBS server overhead.

Liu et al. [13] aimed at the problem that k anonymous regions were prone to leak user information when choosing the construction location, thus they proposed a completely pseudo-k anonymous algorithm that generated multiple virtual addresses through location offset. This algorithm achieved a higher tracking success rate without increasing communication overhead.

Currently, the protection of massive track data generated by LBS service was mainly divided into three aspects: track generalization, false track, and track suppression.

Zhang et al. [14] proposed a double-K mechanism. This mechanism implemented k-anonymity by inserting multiple anonymous machine for receiving query locations between users and location service providers. Meanwhile, the experiment was based on a dynamic pseudonym and location

selection mechanism, which increased the difficulty of the anonymous machine and location service provider to obtain track.

Then, Qu Y et al. [15] introduced a differential privacy identifier in GAN to balance the highly approximate spatiotemporal trajectory of purification data generation and data privacy protection. The evaluation results on real data sets verified the superior performance of this method in terms of protection efficiency and optimization.

Next, Xiong et al. [16] designed two models of image and video privacy protection for automatic driving using GAN. These two models hid the real position of users by hiding the edge information and generated privacy protection output according to the category of sensitive objects. This model realized effective defense against attackers' location inference in offline applications.

Huang et al. [17] suggested a semantic-oriented antagonism network that introduced an attention mechanism and rollover module to synthesize trajectory. In the transmission simulation experiment of COVID-19 under the three prevention measures, obtained 91%–98% of the determination measurement coefficient.

Finally, Chen S et al. [18] proposed a differential privacy scheme for big data publishing based on attribute correlation confusion. The experiment innovatively constructed an identifier based on sensitive attributes and privacy ratio to evaluate the vulnerability of data sets. The experiment finally achieved the balance between flexibility and privacy.

In the existing location privacy protection methods, the semantic information of the user's location is rarely considered or too dependent on trusted third parties. That is why, the research performed in this paper comprehensively considers the semantic information and physical distribution of the location to realize the fast construction of the false location set. Apart from that, the experiment combines false location technology with GAN and hopes to achieve efficient privacy protection by synthesizing trajectory data.

3 Digital Footprint Privacy Protection Algorithms in Different Scenarios

With the rapid advancement of the Internet and social networking services, the emergence of a large number of devices equipped with various sensors, the increasing popularity of GPS positioning in the daily use of the public, and the large-scale deployment of various static sensing devices in urban

management, the digital world is capturing the trajectory of daily human behavior on an unprecedented scale. These collected digital footprints can be vividly called "digital footprints". How to build an information security management system, protect digital footprints and protect users' personal privacy in the era of big data is a challenge that must be faced in the process of future social development. Two common elements of digital footprint privacy protection are user real location protection and track privacy protection.

3.1 False Location Filtering Algorithm for the Real-time Location Request

Wireless sensors form self-organized networks through a large number of micro-sensor nodes in the sensing area and then conduct sensing, transmission, and information collaborative processing. They are usually in a development environment when processing data and are susceptible to damage and tampering. Therefore, it is extremely required to protect the privacy of wireless sensor networks. The location privacy protection methods, when users use LBS, include area anonymity, location disturbance and confusion, and false location. LBS is the use of various types of positioning technology to obtain the current location of the positioning device, and provide information resources and basic services to the positioning device through the mobile Internet, which has the advantages of powerful, simple and direct use, but also brings the risk of leaking user privacy. The false location technology relies on submitting the user's real location and several false locations to the LBS server so that the attacker cannot distinguish between the real location and the false location. Figure 1 demonstrates the system architecture.

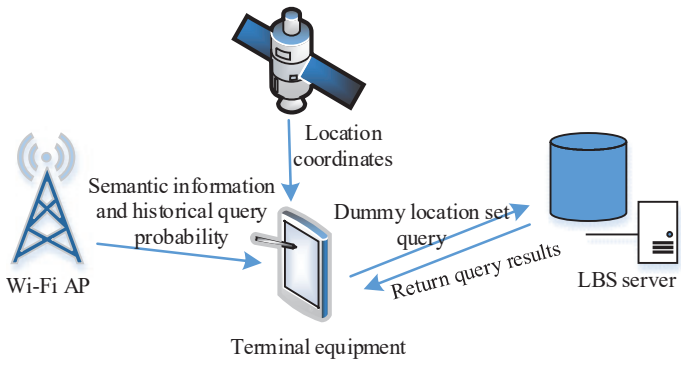


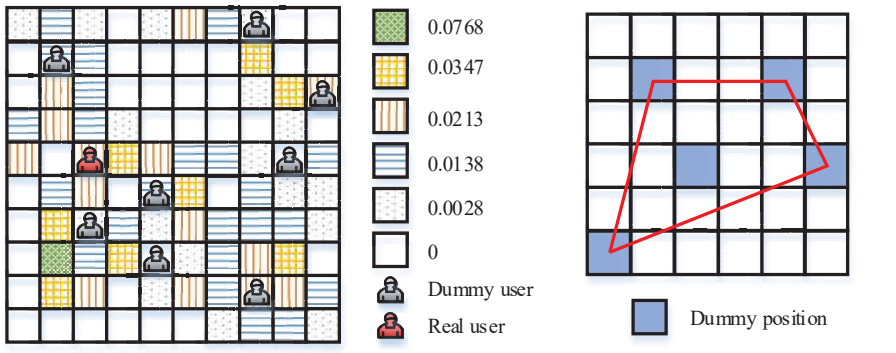
Figure 1 System architecture diagram.

The user's position information in the network is represented by longitude and latitude coordinates (x, y). The LBS request includes (x, y), the requested query content, and the date and time of submission of the request. The false location filtering method based on location query probability and location semantics is studied. Screening and judgment are conducted simultaneously. First of all, based on the large top heap selection query, the location candidate set is generated which is close to the user's real location query probability. Then, the physical distance and semantic distance between the candidate location and the real location are compared respectively, and the final qualified false location is selected. Finally, a false location set P_{dum} containing the user's real location and the size k are generated. In addition to the geographic information of the selected region and the user's location, the algorithm also needs to define the parameters related to privacy requirements in advance: anonymity k and proportion coefficient r. The degree of anonymity k represents the size of P_{dum}. The higher the k value, the more difficult it is for an attacker to steal the user's real location. The proportion coefficient r indicates the weight of semantic distance in the comprehensive distance. Define the rectangular area composed of ($n \times n$ A) uniform grids selected in the experiment as $R_L = \{n \times n, S_{RL}\}$. Wherein, S_{RL} is the dataset of all location points in the region. Define the user's current real location as $Local_R$. The difference between the historical query probability of the real location is d_q. The historical query probability of access location l_i is usually expressed by location query probability P_i. In this research we use the proportion of location access time in $Local_R$ to represent P_i, as shown in formula (1).

$$P_i = \frac{t_i}{\sum_{i=1}^{n^2} t_i} \tag{1}$$

where, t_i represents the time of access at the location. The sum of the query probabilities of all locations in the region is equal to 1. d_p and d_s indicate the physical distance and semantic distance between locations respectively. The semantic distance d_s is calculated by Jaro-Winkler similarity between two positions. Jaro-Winkler similarity is based on the number and order of common characters between two strings s_1 and s_2. The calculation of this similarity Sim_J is performed using formula (2).

$$Sim_J = \begin{cases} 0, & m = 0 \\ \frac{1}{3}\left(\frac{m}{|s_1|} + \frac{m}{|s_2|} + \frac{m-b}{m}\right), & m > 0 \end{cases} \tag{2}$$



(a) Select false location based on anonymity (b) Convex hull area of false position set

Figure 2 Schematic diagram of false position selection and convex hull area of false position set.

In formula (2), $|s_1|$ and $|s_2|$ are the lengths of two strings; m is the number of characters matching two strings; b replaces the digits for the character. Jaro-Winkler similarity Sim_{J-W}, based on Sim_J, further considers the impact of the length of the common prefix of a string on the semantics of the string. Its definition is shown in formula (3).

$$Sim_{J-W} = Sim_J + l_{pre} \times p \times (1 - Sim_J) \tag{3}$$

where, l_{pre} denotes the length of the common prefix of two strings, and the maximum value is 4; p is a constant factor, with a maximum of 0.25 and a default of 0.1. The degree of anonymity of false position sets can be measured using entropy theory. Figure 2 demonstrates the schematic diagram of the convex hull area of the false position and the false position selected according to the anonymity.

The uncertainty of false position set P_{dum} is measured by entropy, which indicates the degree of chaos and uncertainty within the system, as shown in formula (4).

$$E = -\sum_{i=1}^{k} q_i \log_2 q_i \tag{4}$$

In formula (4), when the query probabilities of k locations in the pseudo location set P_{dum} are equal, the information entropy $E_{\max}$ reaches the maximum value $\log_2 k$ at this time, and the anonymity effect is also the best at this time. q_i denotes the query probability of different locations in the false

location set, and its definition is shown in formula (5).

$$q_i = \frac{P_i}{\sum_{i=1}^{k} P_i} \tag{5}$$

The anonymous area D is defined as the convex hull area enclosed by the outermost point of position k points in the false position set, which can be calculated according to the shoelace formula shown in formula (6).

$$D = \frac{1}{2}\left|\sum_{i=1}^{C}(x_i y_{i+1} - x_{i+1} y_i)\right| \tag{6}$$

In formula (6), C is the number of outermost position points in P_{dum}. The semantic difference of false position set is measured by θ_safety, whose definition is shown in formula (7).

$$\theta\_safety = 1 - \frac{|SEM|}{C_k^2} \tag{7}$$

where, $SEM = \{d_s | d_s(l_i, l_j) \leq u\}$; l_i and l_j are any two positions in P_{dum}; u is the set semantic difference threshold. The size of θ_safety value is proportional to the semantic difference of the generated false position set.

3.2 False Track Generation Model for Offline Position Release

The existing track privacy protection methods usually blur the track position by adding uncertainty, but it is tough to balance the effectiveness of privacy protection and the practicality of data using these methods. Therefore, based on the machine learning model, this paper proposes a false track generation model that combines the Long Short-Term Memory (LSTM) network and the Generation Adversary Networks (GAN). The model is applied to generate synthetic trajectory data that cannot be recognized by the trajectory discriminator. The model consists of three important steps: track coding, track generation and track identification. Figure 3 illustrates the overall workflow.

The track coding model used in the model includes track point coding and track filling. Track point $p = (x, y, t, A)$ is GPS coordinate point with time stamp collected by LBS application. It includes longitude and latitude coordinates (x, y), time stamps t, which can be described by r attributes in $A = \{a_1, a_2, \ldots, a_r\}$, such as point of interest (POI) type, access period, weather conditions, etc. Track is composed of a set of track points arranged in sequence. Wherein $p_i (i \in \{1, 2, \ldots, k\})$ denotes the i

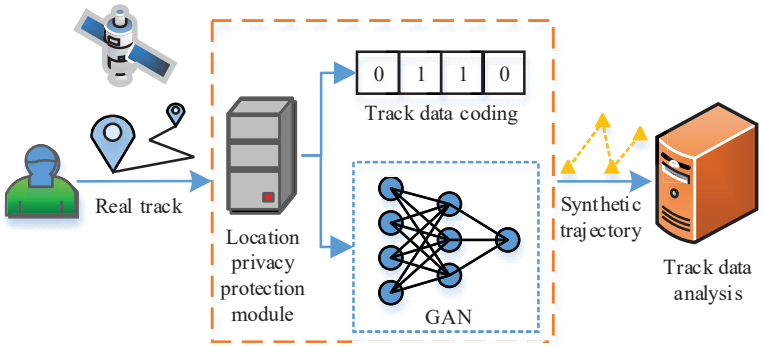


Figure 3 Track data release process for privacy protection based on GAN.

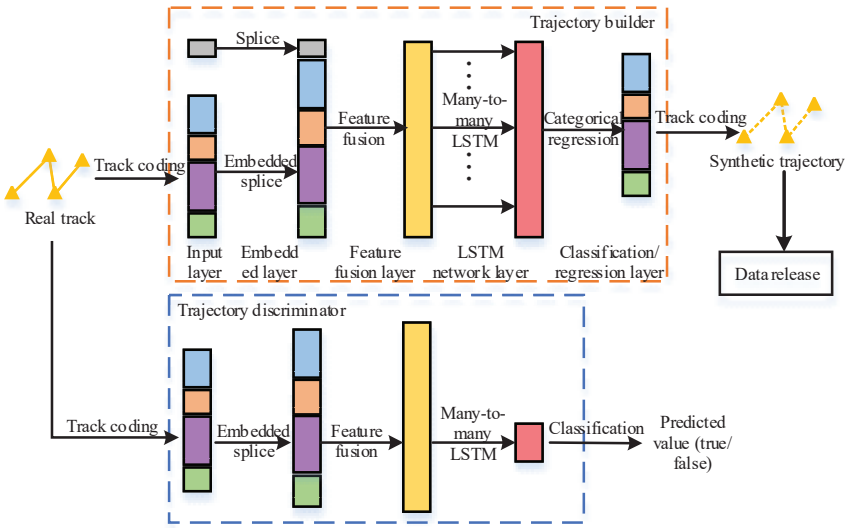


Figure 4 Neural network structure of trajectory generation model.

track point in the track. The track data set composed of n tracks is represented as $\Gamma = \{T_1, T_2, \ldots, T_n\}$. GeoHash coding is applied to represent the spatial attributes of track points. One-hot coding is used to encode other temporal attributes and classification attributes into binary vectors. GeoHash code converts two-dimensional longitude and latitude into one-dimensional string. The higher the prefix matching degree of GeoHash code in two positions, the closer the distance between the two positions is. Figure 4 illustrates the neural network structure of the trajectory generation model.

In Figure 4, the model uses GAN to generate the track most similar to the existing track. GAN is composed of two neural networks, generator D and discriminator G. The track generator first takes real track data and random noise $z \sim p(z)$ as input. Wherein, $p(z)$ means that the noise obeys normal distribution. The discriminator D takes the output sample $G(z)$ of generator G as the input, and the output sample is the probability of training data. The goal of training generator G is to maximize the probability of misjudgement of discriminator D, that is, the generator and discriminator play a minimax game against the objective function shown in formula (8).

$$V(D,G) = \min_{G} \max_{D} \mathrm{E}_{x \sim p_{data}}[\log D(x)] + \mathrm{E}_{z \sim p(z)}[\log(1 - D(G(z)))] \tag{8}$$

In formula (8), the training goal of discriminator D is to maximize the probability of judging true data $x \sim p_{data}$ as true $D(x)$ and minimize the probability of judging false data as true. Therefore, the training goal of generator G is to maximize $D(G(z))$. To obtain the sequence characteristics of trajectory data, LSTM is applied in D network and G network. The generator first converts the input data into a fixed-length vector through the multi-layer neural network embedding process, as shown in formula (9).

$$\begin{cases} e_i^g = \phi^g(v_i^g; W_{es}) \\ e_i^w = \phi^w(v_i^w; W_{ew}) \\ e_i^h = \phi^h(v_i^h; W_{eh}) \\ e_i^p = \phi^p(v_i^p; W_{ep}) \end{cases} \tag{9}$$

where, v_i^s is the binary vector obtained from the coordinates of the i-th track point after GeoHash coding. v_i^d, v_i^h, and v_i^p represent the one-hour codes of the week, hour and POI type of the track point respectively. ϕ^s, ϕ^w, ϕ^h and ϕ^p indicate the neural networks in the embedded layer respectively; W_{es}, W_{ew}, W_{eh} and W_{ep} are the weight matrices of these networks. e_i^s, e_i^w, e_i^h and e_i^p are the embedding vectors of the four outputs. In this work. The application of many-to-many LSTM structure in LSTM network layer is studied. The LSTM network layer accepts the track sequence of fusion features to generate the same size track sequence H, as shown in formula (10).

$$H = LSTM(F; W_{LSTM}) \tag{10}$$

In formula (10), F is the fusion feature of all track points on the same track. W_{LSTM} is the weight matrix of LSTM network. Each feature vector

h_i in H contains the spatiotemporal attribute and semantic feature of the composite track point. Finally, the synthetic track data is decoded from the output of the LSTM network layer, as shown in formula (11).

$$\begin{cases} v_i^s = D^s(h_i^s; W_{ds}) \\ v_i^w = D^w(h_i^w; W_{dw}) \\ v_i^h = D^h(h_i^h; W_{dh}) \\ v_i^p = D^p(h_i^p; W_{dp}) \end{cases} \tag{11}$$

where, D^s, D^w, D^h, and D^p represent the use of fully connected layers with excitation function sigmoid or softmax, respectively. h_i is the vector output by the output LSTM network layer; W_{ds}, W_{dw}, W_{dh}, and W_{dp} represent the weight matrix of the full connection layer. In general, the discriminator D and generator G in GAN can set the binary cross entropy as the loss function L_{BCE} during training. This function is often used in binary classification problems, as shown in formula (12).

$$L_{BCE} = -\frac{1}{N}\sum_{i=1}^{N} Q_i \cdot \log(p(Q_i)) + (1 - Q_i) \cdot \log(1 - p(Q_i)) \tag{12}$$

In formula (12), the value of Q is 1 or 0, which is a binary label of real data. $p(Q)$ indicates the probability that the output belongs to Q. The proposed model needs to input real trajectory data. Therefore, the experiment needs to consider the similar loss of real track data and synthetic track data in space-time and POI category dimensions, and use the updated loss function to train the generator. The updated loss function $L^*(Q_r, Q_p, T_r, T_p)$ is shown in formula (13).

$$\begin{aligned} L^*(Q_r, Q_p, T_r, T_p) = \alpha L_{BCE}(Q_r, Q_p) + \beta S_s(T_r, T_s) \\ + \gamma S_t(T_r, T_s) + \lambda S_c(T_r, T_s) \end{aligned} \tag{13}$$

Q_r and Q_p represent the real label of the trajectory and the prediction label of the discriminator for the trajectory; T_r and T_p are real tracks and synthetic tracks respectively; S_s, S_t, and S_c are the spatial similarity loss, temporal similarity loss and POI category similarity loss of the two tracks respectively; β, γ and λ represent their weights. The study used ACC@1, ACC@5 and Macro-F1 to evaluate the accuracy of the Trajectory User Link (TUL) task for the model. ACC@1 and ACC@5 indicate the maximum predicted classification probability and the probability of having a correct

classification in the top five, respectively. Macro-F1 is equal to the average value of F1 of all categories, as shown in formula (14).

$$\begin{cases} Macro\text{-}F1 = \dfrac{1}{v}\sum_{i=1}^{v} F1_i \\ F1 = \dfrac{2PR}{P+R} \end{cases} \tag{14}$$

In formula (14), v is the number of classifications; $F1$ is the average of precision P and recall R, which can reflect the overall performance of the model. The spatial and temporal characteristics of synthetic trajectories are evaluated by using the Hausdorff distance measurement index and the access probability distribution index of different time periods and without POI type. The Hausdorff distance is calculated as shown in formula (15).

$$\begin{cases} H(T_A, T_B) = \max[h(T_A, T_B), h(T_B, T_A)] \\ h(T_A, T_B) = \max\limits_{p_a \in T_A} \min\limits_{p_b \in T_B} \|p_a - p_b\| \\ h(T_B, T_A) = \max\limits_{p_b \in T_B} \min\limits_{p_a \in T_A} \|p_b - p_a\| \end{cases} \tag{15}$$

where, $h(T_A, T_B)$ and $h(T_B, T_A)$ are the one-way Hausdorff distance between the two point sets T_A and T_B respectively.

4 Application of Privacy Protection Technology for Digital Footprints in Wireless Sensor Networks Based on Big Data

The study addresses location privacy protection methods in different scenarios in location-based services, and optimizes and improves the shortcomings of current privacy protection techniques by combining existing privacy protection techniques from two scenarios: real-time single query location and trajectory publishing in location sharing. The current metrics for evaluating location privacy protection techniques need to consider various aspects such as the operation speed, security, stability and performance overhead of the method, so the study evaluates the proposed fake location screening algorithm in five aspects, namely the efficiency of fake location set generation, physical dispersion, semantic diversity, uncertainty and recognition rate, and analyzes the effectiveness of trajectory protection by suppressing the accuracy of the TUL algorithm.

4.1 Performance Evaluation Index and Experimental Parameter Setting

Two algorithms are proposed in this study, namely, the false location filtering algorithm for real-time location requests and the false path generation model for offline location publishing. The effectiveness of the two performances is verified by experiments. The first experiment is to validiate sthe false position filtering algorithm. This experiment applies the Geolife [19] track data set from Microsoft Asia Research Institute to simulate the user's historical query probability in various regions of the map, and supplements the semantic information obtained through Baidu Map API. The Geolife trajectory dataset contains 17621 GPS trajectories from 182 users from 2007 to 2012. The coordinate range of the experimental area selected in the data set is 39.95°–40.00° N and 116.30°–116.35° E. The experiment divides the area into 200 × 200 grid; The position of each grid is represented by the longitude and latitude of the grid center.

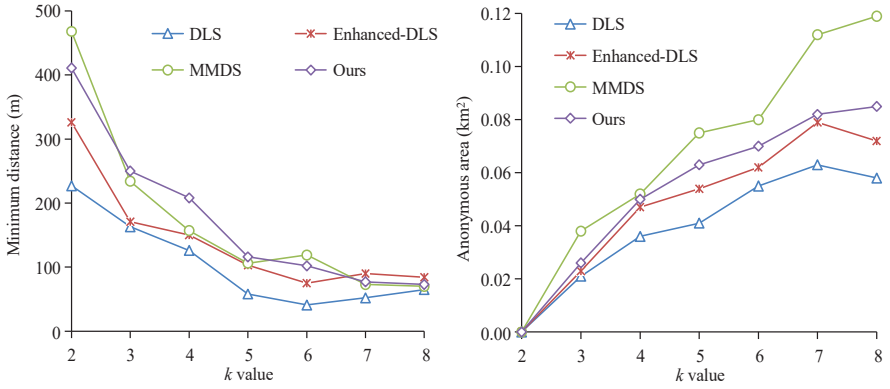
The second experiment is to evaluate the false track generation model for offline position release. The experiment uses the Foursquare weekly track data set extracted from the Foursquare NYC Check-Ins data set. The attribute data is screened in the experiment. 2/3 of the data in the data set is selected as the training data set of the model. The remaining 1/3 serves as the test data set. The parameter settings of Experiments 1 and 2 are shown in Table 1.

4.2 Experimental Results of False Location Filtering Algorithm for Real-time Location Request

In this experiment we use five metrics to evaluate the algorithm: physical dispersion, false position set generation efficiency, semantic diversity, uncertainty and recognition rate. In addition, three other algorithms are selected for performance comparison: Dummy Location Selection (DLS) algorithm [20],

Table 1 Experimental parameter setting

Experiment 1		Experiment 2	
Parameter	Value	Parameter	Value
Anonymity k	≥ 2	Batch size	256
Scale factor r	9	Parameter saving cycle/epoch	100
Semantic difference threshold u	0.15	Number of longest tracks	144
Number of grids	20 × 20	Loss function weight ($\alpha, \beta, \gamma, \lambda$)	1,5,1,1
Number of selectable areas	100	Geohash encoding length	8
Wi-Fi AP coverage area/m	700	Geohash binary output threshold	0.5



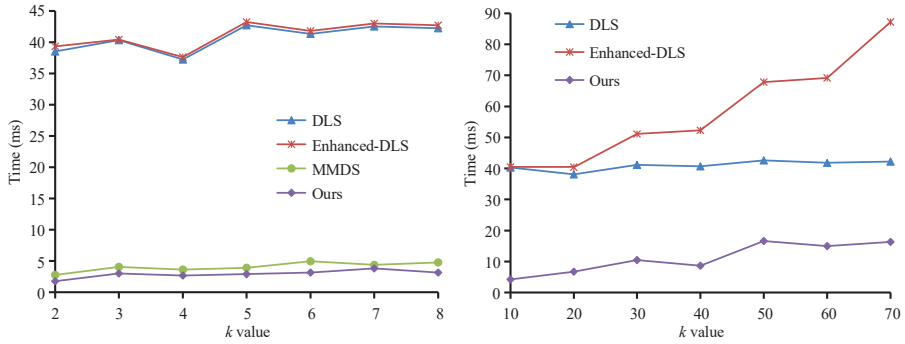
(a) The relationship between the minimum distance between false positions and k value

(b) The relationship between the area of anonymous area and k value

Figure 5 Relationship between k value and minimum distance between false position.

Enhanced-DLS algorithm [21] and Maximum and Minimum Dummy Selection (MMDS) algorithm [22]. The physical dispersion of false location sets can be measured from the minimum distance between false locations and the area of anonymous regions. The larger the value of both, the more evenly distributed the false positions generated. The relationship between the k value and the minimum distance between the false positions and the area of the anonymous area covering the false positions is shown in Figures 5(a) and 5(b), respectively.

In Figure 5(a), the minimum distance between false positions generated by the four algorithms gradually decreases with the increase of k value. It tends to flatten after $k \geq 5$. When $k \leq 4$, the minimum distance of the proposed algorithm and MMDS algorithm is more than 400 meters, which is significantly greater than DLS and Enhanced-DLS algorithms. The two algorithms that perform well consider the physical distance and semantic distance between locations. The result shows that ensuring semantic diversity can make the generated false position distribution more dispersed. In Figure 5(b), when $k > 5$, the anonymous area of MMDS algorithm is much larger than that of other algorithms. It reached 0.12 square kilometers at $k = 8$. This is because when the other three algorithms generate false locations, the first consideration is the historical query probability of the location. However, the algorithm proposed in the study still has better physical dispersion than DLS and Enhanced-DLS algorithms. Figure 6 illustrates the influence of the k value on the time when different algorithms generate false positions.



(a) The time for different algorithms to generate false position sets when the value of k is small

(b) The time for different algorithms to generate false position sets when the value of k is large

Figure 6 The influence of k value on the time of generating false position by different algorithms.

In Figure 6, when the value of k is small, the time for MMDS algorithm and the algorithm proposed in the study to generate false position sets does not exceed 5 ms; The time of DLS and Enhanced-DLS algorithms is between 37 and 45 ms. Figure 6(b) shows the time-consuming change of each algorithm to generate false position set when the value of k continues to increase. Due to the limited number of POI categories, the probability of successful anonymity of MMDS algorithm after $k \geq 10$ is extremely low. Therefore, it is only required to compare the time consumption of the other three algorithms. The increase of the proposed algorithm is between Enhanced-DLS and DLS, and the time difference is not more than 15 ms. This method adopts the strategy of selecting while judging, which is the most efficient in generating false position sets among the three algorithms. The semantic diversity of false location is evaluated by θ_safety. The larger the value, the more difficult it is for attackers to determine the semantic information of the real location. The θ_safety of the false position set is shown in Figure 7.

In Figure 7, the θ_safety of the proposed algorithm is between 0.8 and 0.9, slightly lower than that of MMDS algorithm. The θ_safety of MMDD is above 0.9. However, with the increase of θ_safety value, the probability of anonymity failure of the algorithm will increase until anonymity is impossible. The θ_safety of Enhanced-DLS algorithm is slightly higher than that of DLS, and about 0.7 higher. The reason is that the former considers the factor of anonymous area. It can be seen that there is a positive correlation between physical dispersion and semantic diversity. Figure 8 illustrates the position entropy and recognition rate of the false position set.

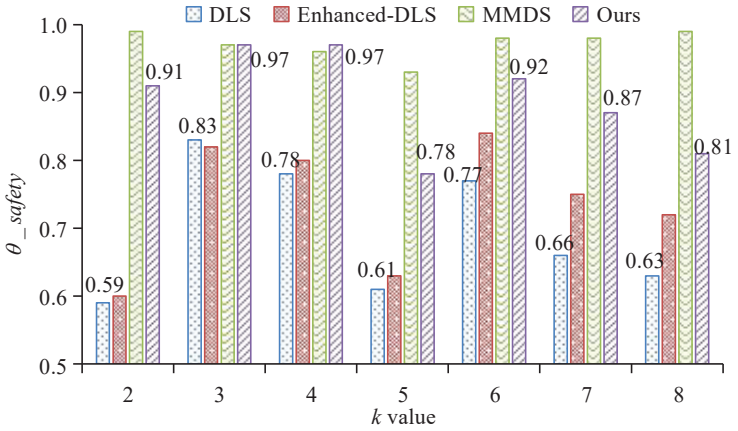


Figure 7 θ_safety of false position set.

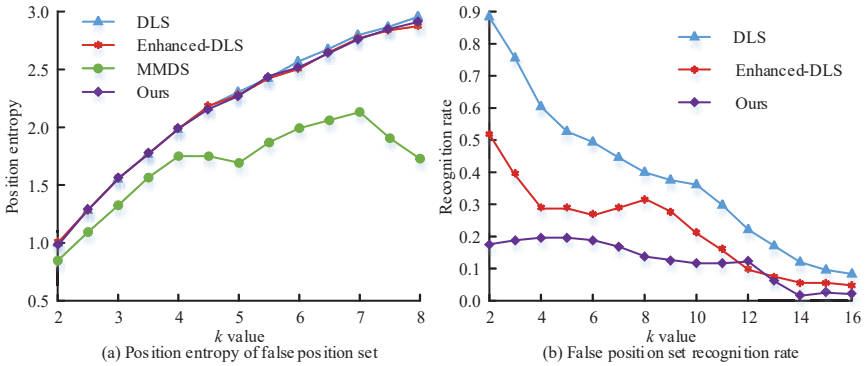


Figure 8 Position entropy and recognition rate of false position set.

In Figure 8, except for the MMDS algorithm, the position entropy of the other three algorithms gradually approaches the maximum value $\log_2 k$ as the value of k increases. The recognition rate in Figure 8(b) is the attacker's recognition rate of the user's real location. The higher the value, the lower the anonymity rate of the algorithm. The false location set generated by the three algorithms is most easily recognized by attackers, and its recognition rate is close to 90% at the highest. The algorithm proposed in the study has the lowest probability of being recognized, which does not exceed 21% as a whole, and is almost unaffected by the increase of k value. According to the experimental results of five metrics, it can be seen that the proposed false location filtering algorithm can quickly generate physically dispersed

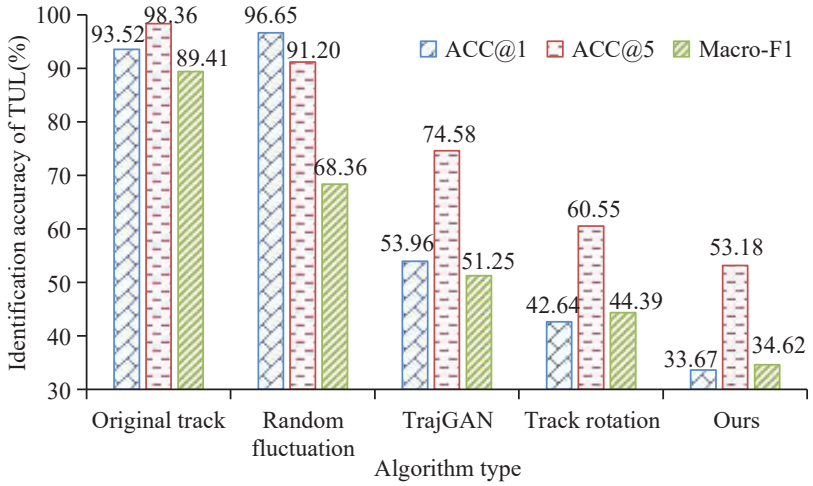


Figure 9 Accuracy of TUL task to identify users.

and semantically diverse false location sets. Therefore, the proposed method can effectively prevent the attack and identification of attackers, and plays an important role in protecting the privacy of users' digital footprints.

4.3 Experimental Results of False Track Generation Model for Offline Position Release

To verify the data privacy and practicability of the model, firstly, the effectiveness of the model for trajectory privacy protection is evaluated on TUL. Then the validity of the synthetic trajectory in data analysis is verified according to its spatiotemporal characteristics. The research applies traditional random perturbation [23], trajectory rotation [24] and Trajectory generation model based on GAN (TrajGAN) [25]. The core idea of TUL task is to identify users based on data mining. The higher the recognition accuracy, the worse the ability of track privacy protection. Figure 9 demonstrates the accuracy of TUL task for user identification of different algorithms.

In Figure 9, the model used in the study significantly suppresses the recognition accuracy of TUL tasks; In particular, the decline of ACC@1 and Macro-F1 metrics is the most obvious, falling from more than 90% to about 34%. ACC@5 also decreases from 98.36% to 53.18%. It shows that the false track generation model proposed in the study can effectively protect the user's data privacy. The Hausdorff distance of the track point set is shown in Figure 10.

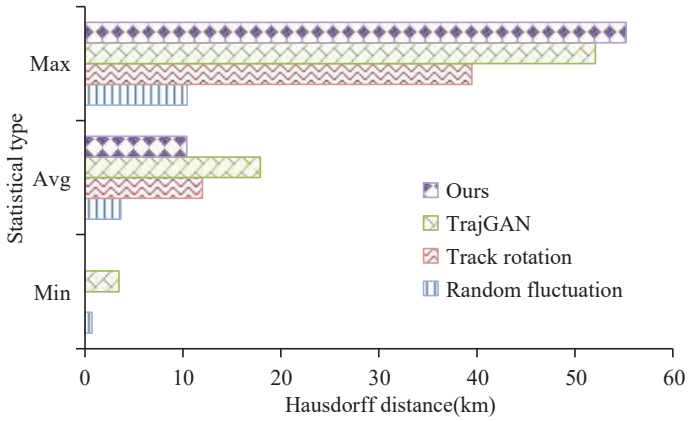


Figure 10 Hausdorff distance of locus point set.

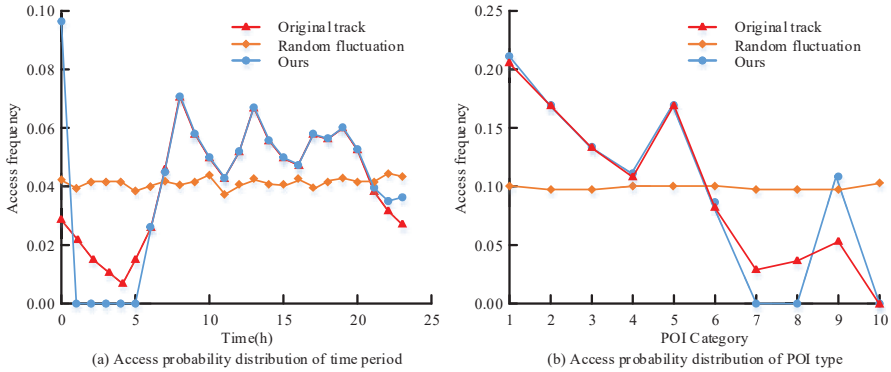


Figure 11 Period access probability distribution and POI type access probability distribution.

In Figure 10, the track data with the smallest average Hausdorff distance is random disturbance. But this method is to preserve spatial similarity at the expense of privacy. The trajectory generated by trajectory rotation has the maximum average Hausdorff distance, which is 16.8 kilometers, and also has the lowest spatial similarity. The time interval access probability distribution and POI type access probability distribution of the original track and the composite track are shown in Figure 11.

From Figure 11, the model proposed in the study can better fit the overall time access frequency distribution of the original data compared with the random disturbance. This indicates that the model better preserves the

time and POI category characteristics of the track data. Generally speaking, there is always a conflict between user privacy protection and data validity. However, the model based on deep learning can better monitor and quantify this relationship in the process of training and testing, so as to find the most balanced parameter setting. The performance of the proposed algorithm on several metrics is not necessarily optimal, while it is the most balanced. It can not only maintain the spatiotemporal characteristics of data, but also ensure a low TUL recognition rate.

5 Conclusion

The popularization and development of LBS have brought great convenience to social media based on mobile devices, satellite positioning, and geographic tagging to collect users' digital footprints in the cyberspace. However, how to avoid large-scale private information leakage caused by digital footprints is a crucial issue. For the protection of location privacy in different scenarios, the false location filtering algorithm of multiple data and the false path generation model combined with false location technology and GAN are proposed. Protecting user privacy data by constructing fake location candidate sets and synthesizing trajectories. The false position filtering algorithm performs well in five indicators. The time to generate false position set is within 5 ms. The index θ_safety of semantic diversity is between 0.8 and 0.9. Compared with the original trajectory, the three metrics of the model proposed in the study in the TUL task decreases by about 64%, 45.9% and 61.3% respectively. The results prove that the synthetic trajectory generated by the model can significantly inhibit the accuracy of TUL task prediction, and thus effectively protect the user's trajectory privacy. However, the model proposed in the study does not consider the variations in privacy needs of different users, so it is tough to carry out practical applications in complex scenarios. In the subsequent research, it is required to adjust the privacy protection level according to the actual situation, so as to avoid excessive data protection and further enhance the practicability of the model.

References

[1] Mou J. Extracting Network Patterns of Tourist Flows in an Urban Agglomeration Through Digital Footprints: The Case of Greater Bay Area. IEEE Access, 2022, 10: 16644–16654.

[2] Tucakovi'c L, Boji'c L. Computer-based personality judgments from digital footprints: theoretical considerations and practical implications in politics. Srpska politièka misao, 2022, 74(4): 235–253.
[3] Quach S, Thaichon P, Martin K D, Weaven S, Palmatier R W. Digital technologies: Tensions in privacy and data. Journal of the Academy of Marketing Science, 2022, 50(6): 1299–1323.
[4] Kolasa K, Ken Redekop W, Berler A, Zah V, Asche C V. Future of data analytics in the era of the general data protection regulation in Europe. PharmacoEconomics, 2020, 38(10): 1021–1029.
[5] Andrew J, Baker M. The general data protection regulation in the age of surveillance capitalism. Journal of Business Ethics, 2021, 168: 565–578.
[6] Wei K, Li J, Ding M, Ma C, Yang H H, Farokhi F, Jin S, Quek T Q S, Poor H V. Federated learning with differential privacy: Algorithms and performance analysis. IEEE Transactions on Information Forensics and Security, 2020, 15: 3454–3469.
[7] Yadav V K, Verma S, Venkatesan S. Linkable privacy-preserving scheme for location-based services. IEEE Transactions on Intelligent Transportation Systems, 2021, 23(7): 7998–8012.
[8] Farouk F, Alkady Y, Rizk R. Efficient privacy-preserving scheme for location based services in VANET system. IEEE Access, 2020, 8: 60101–60116.
[9] Wu Z, Wang R, Li Q, Lian X, Xu G, Chen E, Liu X. A location privacy-preserving system based on query range cover-up or location-based services. IEEE Transactions on Vehicular Technology, 2020, 69(5): 5244–5254.
[10] Hassan M U, Rehmani M H, Chen J. Differential privacy techniques for cyber physical systems: a survey. IEEE Communications Surveys & Tutorials, 2019, 22(1): 746–789.
[11] Wu Z, Li G, Shen S, Lian X, Chen E, Xu G. Constructing dummy query sequences to protect location privacy and query privacy in location-based services. World Wide Web, 2021, 24: 25–49.
[12] Zhang S, Li X, Tan Z, Peng T, Wang G. A caching and spatial K-anonymity driven privacy enhancement scheme in continuous location-based services. Future Generation Computer Systems, 2019, 94: 40–50.
[13] Liu J, Wang S. All-dummy k-anonymous privacy protection algorithm based on location offset. Computing, 2022, 104(8): 1739–1751.

[14] Zhang S, Mao X, Choo K K R, Peng T, Wang G. A trajectory privacy-preserving scheme based on a dual-K mechanism for continuous location-based services. Information Sciences, 2020, 527: 406–419.
[15] Qu Y, Yu S, Zhou W, Tian Y. Gan-driven personalized spatial-temporal private data sharing in cyber-physical social systems. IEEE Transactions on Network Science and Engineering, 2020, 7(4): 2576–2586.
[16] Xiong Z, Cai Z, Han Q, Alrawais A, Li W. ADGAN: Protect your location privacy in camera data of auto-driving vehicles. IEEE Transactions on Industrial Informatics, 2020, 17(9): 6200–6210.
[17] Huang C, Chen S, Zhang Y, Zhou W, Rodrigues J. A robust approach for privacy data protection: IoT security assurance using generative adversarial imitation learning. IEEE Internet of Things Journal, 2021, 9(18): 17089–17097.
[18] Chen S, Fu A, Yu S, Ke H, Su M. DP-QIC: A differential privacy scheme based on quasi-identifier classification for big data publication. Soft Computing, 2021, 25: 7325–7339.
[19] Masnabadi N, Hosseinali F, Bahramian Z. Developing a spatial and temporal density-based clustering algorithm to extract stop locations from the user's trajectory. Engineering Journal of Geospatial Information Technology, 2021, 9(2): 105–128.
[20] Sun G, Chang V, Ramachandran M, Sun Z, Li J, Yu H, Liao D. Efficient location privacy algorithm for Internet of Things (IoT) services and applications. Journal of Network and Computer Applications, 2017, 89: 3–13.
[21] Chung B, Ptasznik A, Wu D, Bonaci T. Privacy and location-based services. IEEE Potentials, 2022, 41(4): 31–37.
[22] Wu L, Wei X, Meng L, Zhao S, Wang H. Privacy-preserving location-based traffic density monitoring. Connection Science, 2022, 34(1): 874–894.
[23] Gao S, Rao J, Liu X, Kang Y, Haung Q, App J. Exploring the effectiveness of geomasking techniques for protecting the geoprivacy of Twitter users. Journal of Spatial Information Science, 2019, 19: 105–129.
[24] Wang T, Zeng J, Bhuiyan M Z A, Tian H, Cai Y, Chen Y, Zhong B. Trajectory privacy preservation based on a fog structure for cloud location services. IEEE Access, 2017, 5: 7692–7701.
[25] Chen X, Xu J, Zhou R, Chen W, Fang J, Liu C. TrajVAE: A Variational AutoEncoder model for trajectory generation. Neurocomputing, 2021, 428: 332–339.

[26] Parmar D, Rao U P. Towards privacy-preserving dummy generation in location-based services. Procedia Computer Science, 2020, 171: 1323–1326.
[27] Jiang J, Han G, Wang H, Guizani M. A survey on location privacy protection in wireless sensor networks. Journal of Network and Computer Applications, 2019, 125: 93–114.
[28] Huang Q, Du J, Yan G, Yang Y, Wei Q. Privacy-preserving spatio-temporal keyword search for outsourced location-based services. IEEE Transactions on Services Computing, 2021, 15(6): 3443–3456.
[29] Manju A B, Subramanian S. Fog-Assisted Privacy Preservation Scheme for Location-Based Services Based on Trust Relationship. International Journal of Grid and High Performance Computing (IJGHPC), 2020, 12(4): 48–62.
[30] Yang P, Xiong N, Ren J. Data security and privacy protection for cloud storage: A survey. IEEE Access, 2020, 8: 131723–131740.

Biographies

Hong Zhang, male, 1982.5. He received a master's degree in measurement technology and instruments from North University of China In 2013. He is currently a associate professor and senior engineer in the Information Engineering Department of Shanxi Conservancy Technical Institute. His research interests are mainly in the fields of network engineering and information security.

Pei Li, female, 1981.7. In 2010, she received a master's degree in measurement technology and instruments from North University of China. She is currently a lecturer in the Information Engineering Department of Shanxi Conservancy Technical Institute. Her research interests mainly focus on software engineering and water conservancy informatization.