IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received March 21, 2019, accepted April 11, 2019, date of publication April 18, 2019, date of current version May 6, 2019.

Digital Object Identifier 10.1109/ACCESS.2019.2912006

Location Correlated Differential Privacy Protection Based on Mobile Feature Analysis

ZHENLONG PENG[1,3,4,5], JIAN AN[2], (Member, IEEE), XIAOLIN GUI[3,5], ZHENXING WANG[3,5], WENDONG ZHANG[3,5], RUOWEI GUI[3,5], AND JINGXIAN XU[1,4]

[1]TSL Business School, Quanzhou Normal University, Quanzhou 362000, China
[2]Xi'an Jiaotong University Shenzhen Research School, Shenzhen 518057, China
[3]School of Electronics and Information Engineering, Xi'an Jiaotong University, Xi'an 710049, China
[4]High Educational Engineering Research Center of Fujian Province for E-Commerce Intelligent Based on Cloud Computing and Internet of Things, Quanzhou Normal University, Quanzhou 362000, China
[5]Shaanxi Province Key Laboratory of Computer Network, Xi'an Jiaotong University, Xi'an 710049, China

Corresponding authors: Jian An (anjian@mail.xjtu.edu.cn) and Xiaolin Gui (xlgui@mail.xjtu.edu.cn)

This work was supported partially by the NSFC (no. 61472316) and Science and Science and Technology Program of Shenzhen under grant JCYJ20170816100939373 and partly by the Key Projects of Science and Technology Plan in Fujian Province under grant 2016H0029, the Technology Project of Shaanxi Province (no. 2017ZDXM-GY-011), Fujian Social Science Project (no. FJ2018B022), and Key Projects of Quanzhou Technology Bureau no. 2018C004.

ABSTRACT Recently, with the popularity of smartphones and other GPS embedded devices, location-based service applications are being rapidly developed. In addition, individual privacy protection is also receiving increasing attention. Currently, most studies assume that individual location records are independent. However, the records are mostly interrelated in the real world. If the information is protected without considering the location-correlated information between users, an attacker can use a background knowledge attack to obtain the user's private information. Therefore, this paper proposes a method to protect multiuser location-correlated information under a strict privacy budget. First, a method for group movement analysis based on adaptive time segmentation is proposed in this paper. In addition, based on the time dimension, time-continuous hotspot areas are constructed by adaptively segmenting and merging the stay areas, which are established for subsequent location-correlated privacy protection. Second, a data publishing mechanism is proposed to resist inferred attacks and to adaptively protect user-correlated location information. In addition, this paper also proposes the individual user correlation sensitivity concept and extends differential privacy by building an individual sensitivity matrix to correct noise. The experiments on real datasets show that under the same conditions, compared with the existing methods, the heat value of the hotspot areas formed by the method is increased by 10.11% under the same time slice length. In addition, the method reduces the similarity of 26.98% of group users.

INDEX TERMS Correlated differential privacy, location protection, mobile feature analysis, time division.

I. INTRODUCTION

Recently, with the popularity of GPS mobile devices, location-based service applications have been widely used in social and commercial fields. However, users generate a large amount of spatiotemporal data when using these services, which leads to leaking sensitive personal information. Therefore, it is vital to protect users' spatiotemporal data [1].

Taking a wide view of current studies on the privacy protection of user location information, most of the methods assume that the dataset is independent and the internal data records are also independent. However, in real life, people have their social circles, and they often engage in certain social activities [2]. For example, some colleagues in a company may arrive at a designated place within a specified time in the morning and have breakfast together. Therefore, combining location data and time can reflect some of the user's social attributes [3], such as their home address, eating

The associate editor coordinating the review of this manuscript and approving it for publication was Kaigui Bian.

habits, entertainment preferences, and company address [4]. If the personal privacy of the user is protected but their related attributes are ignored, the attacker is likely to use background knowledge to carry out an inferred attack, resulting in a high probability that their personal information is leaked. Thus, protecting the location correlated information between users has become a crucial issue that requires an urgent solution.

However, for location protection, few studies consider correlated information and time factors [5], [6]. Therefore, to solve the problem of location correlated privacy protection, this paper proposes a personalized correlation location data publishing mechanism, which utilizes time-correlated location features to extract the user's location information. At the same time, we use the Jaccard similarity coefficient to extract position correlation information among multiple users and constructs a location-sensitive candidate set, which is the basis of personalized differential privacy protection data publishing. Then, a user correlation matrix is constructed, and we propose the concept of individual user correlation sensitivity, which makes full use of different correlation sensitivity to correct the noise, personalize the user protection and greatly reduce the introduction of noise. Finally, this paper proposes a personalized location publishing mechanism that uses different mechanisms to protect sensitive and non-sensitive candidate sets. More specifically, the main idea is summarized in the following three points in this article.

1) Feature extraction is performed on the location data by a movement feature analysis. Currently, most of the analysis of hotspot areas do not consider the time factors resulting in the clustering results that cannot sufficiently reflect relevant information from group users. To solve this problem, this paper proposes a time correlated group user hotspot areas extraction method to extract the user's relevance based on features such as location, speed, and time.

2) A correlation matrix is proposed, and the personal correlation sensitivity of the user is defined on the basis of user location feature extraction. By the premise of data privacy protection, the introduction of noise in differential privacy is minimized.

3) A personalized privacy location publishing mechanism is designed. Personalized privacy protection is a complement for user-sensitive candidate sets and non-sensitive candidate sets, based on user affinity, which ensures that the dataset has a high availability on the basis of protecting the location information of the user and the associated information.

II. RELATED WORK

Currently, there are few studies on the privacy protection of correlated location data, and most of the research aims to protect the data information without considering the time factor [5], [7]. To anonymize the locations of the users, Hayashid et al. [8] generated a virtual object based on the estimated virtual trajectory and sent its location to the LBS provider along with the actual location of the user, thereby anonymizing the user's location. Based on the k-anonymity location privacy preserving model, Zheng et al. [9] proposed a clustering algorithm aiming at eliminating outliers, which balances the conflict between privacy protection security and the query quality caused by the accuracy of location information. Besides, based on user density and the non-uniform user distribution, Yang et al. [10] proposed a data release mechanism for crowdsensing techniques, which satisfies differential privacy and provides rigorous protection for location information. However, these studies on privacy protection for location information start from user's personal location data. The implicit assumption in these research process is that the data of individual users are independent, and the internal records of the users are also independent. While tending to achieve good metrics in some areas, these privacy protection methods for individual users neglect the relevance of the records and are vulnerable to the associated attack.

Besides, Gu et al. [11] designed a trajectory data privacy protection scheme based on Laplace's differential privacy mechanism, in which the noise is added to the polygon centroid by the Laplace's differential privacy method, and then the new polygon centroid is used to replace the protected points, and finally the algorithm constructs and issues the new trajectory data. However, these methods are not optimal because it introduces a large amount of noise into the dataset and affects the availability of the data. Wang and Xu [12] proposed an effective correlated time-series data publication solution based on differential privacy by enforcing Series-Indistinguishability and designing a correlated Laplace mechanism, while the algorithm is too complex and expensive to apply in real-world scenarios.

Cao [13] and Cao et al. [14] used time interval correlation analysis to identify the correlation of records and analyzed the correlation information by intrinsic behavior function modeling. Furthermore, Li et al. [15], without considering the privacy protection of multiuser, proposed a multi-instance learning algorithm mapping association records to undirected graphs. Zhu et al. [16] proposed the use of correlation coefficients to extract the correlation between data, while the algorithm has high complexity, and the association properties between users are not easily extracted.

Compared with previous approaches to location privacy protection, the main contributions of this paper are summarized as follows:

- Under the premise of considering time, the method of the paper can not only protect the user's location information, but also protect the multi-user's correlation information.
- In order to reduce the noise in differential privacy protection, the paper propose the concept of individual user correlation sensitivity, which makes full use of different correlations to personalize the user protection, and reduce the introduction of noise.
- the method of the paper has a relatively small time complexity. The time complexity of hot spot extraction algorithm adopted by this paper is $O(m^2)$, and at the same time, the obtained hotspot areas are time-correlated. In addition, the process of constructing the user's

correlation sensitivity is translated into the construction of matrix, which also further improves the efficiency of the algorithm.

III. PROBLEM DEFINITION

A. RELATED DEFINITIONS

The spatio-temporal sequence data records a series of information that the continuous moving object appears in the corresponding position at a certain moment, which is a representation of the spatio-temporal attributes in its motion state [17], [18]. The trajectory is a formal expression of spatio-temporal sequence data, where the mathematical definition of trajectory data, set of stay points, and set of stay areas is given.

Definition 1 (Trajectory Data T): A point moves in time and space, so the trajectory data $\boldsymbol{T}$ represents the path formed by the moving point, and we set:

$$\boldsymbol{T} = \{T_{i1}, T_{i2}, \ldots, T_{\boldsymbol{in}}\} \tag{1}$$

where $|T|$ is the physical length of the current trajectory, and T_{in} is the n-th location point of user i, including longitude, dimension, elevation, and other information. The meaning of each trajectory is the historical data of the movement of a person/object. In attempting to ease of description, this article assumes that the sampling instants are the same for all trajectories.

Definition 2 (Stay Point P_{ik}): The individual stay point $\boldsymbol{P_{ik}}$ represents a ~~stay~~ set of individual location point T_{ij} satisfying the speed and distance requirements for the user M_i and we set:

$$\boldsymbol{P_{ik}} = \{T_{\boldsymbol{ij}}, T_{\boldsymbol{ij}+1}, \ldots, T_{\boldsymbol{im}}\} \tag{2}$$

If two adjacent stay points satisfy these requirements, the new user stay point can be obtained by merging the two stay points into one stay point.

Definition 3 (Individual Stay Area A_{ik}): Individual stay area $\boldsymbol{A_{ik}}$ represents a series of stay points that meet distance requirement.

$$\boldsymbol{A_{ik}} = \{\boldsymbol{P_{ij}}, \boldsymbol{P_{ij+1}}, \ldots, \boldsymbol{P_{im}}\} \tag{3}$$

When merging a stay point to an existing individual stay area, if the distance of longitude or latitude between the above stay point and a point in the existing stay area is less than a certain threshold, the stay point will be merged into the stay area.

Definition 4 (Group Stay Area GA_k): Group stay area $\boldsymbol{GA_k}$ denotes a set of individual stay area $\boldsymbol{A_{ij}}$ from different user.

$$\boldsymbol{GA_k} = \{\boldsymbol{A_{ij}}, \ldots, \boldsymbol{A_{nm}}\} \tag{4}$$

Definition 5 (Time Correlated Hotspot HA_k): Considering time continuity, the time correlated hotspot area is

$$\boldsymbol{HA_k} = \{\boldsymbol{GA_i^j}, \boldsymbol{GA_i^{j+1}}, \ldots, \boldsymbol{GA_i^m}\} \tag{5}$$

where $\boldsymbol{GA_i^j}$ represents the i-th group stay area at time j.

TABLE 1. Symbol definition.

Symbol	Definition
T_{in}	The n-th location point of user M_i
M_i	Identity identifier of user M_i
P_{ik}	The k-th stay point of user M_i
A_{ik}	The k-th stay area of user M_i
GA_k	The k-th group stay area
HA_k	The k-th hotspot area
L_{Pik}	The coordinates of the position of the stay point P_{ik}
L_{Aik}	The coordinates of the position of the individual stay area A_{ik}

Definition 6 (Heat Value): The heat value of each hotspot area is expressed in the form of the number of location points of group users per unit time. The specific definition presented as follows:

$$H_i = \frac{\|HA_i\|}{AT_i} \tag{6}$$

In which, $\|HA_i\|$ represents the number of location points in the ith hotspot area and AT_i represents the time length of the ith hotspot area. With other parameters remaining the same, the higher heat value indicates the larger number of people in the hot spot per unit time, i.e. which means stronger time-correlation.

B. SYMBOL DEFINITION

IV. LOCATION-CORRELATED PRIVACY PROTECTION MODEL

This section will explain the details of the framework in this paper based on the problems and challenges in the current study. The problem to be solved in this paper is to protect the location information of multiple users on the basis of ensuring the privacy security of the user location and proposes a data publishing algorithm that supports the protection of user location information and other relevant information. To meet the algorithm's requirements, the challenges previously mentioned are solved separately. As shown in Fig.1, the model design and detailed algorithm proposed in this article are introduced below.

Firstly, according to these factors such as location, speed, and time, feature extraction of location data can be carried out by movement feature analysis. The beginning of the feature extraction in this paper is data pre-processing. There are quite a few factors that must be taken into account here. Because there is no immediate information in the location data to indicate which users are correlated with other users, and some data have different social attribute, the relationship between users can be determined through the degree of interaction with each other to obtain the relevant correlation information between users. However, traditional position clustering does not take the time factor into account; if the clustering is performed without considering the time factor,

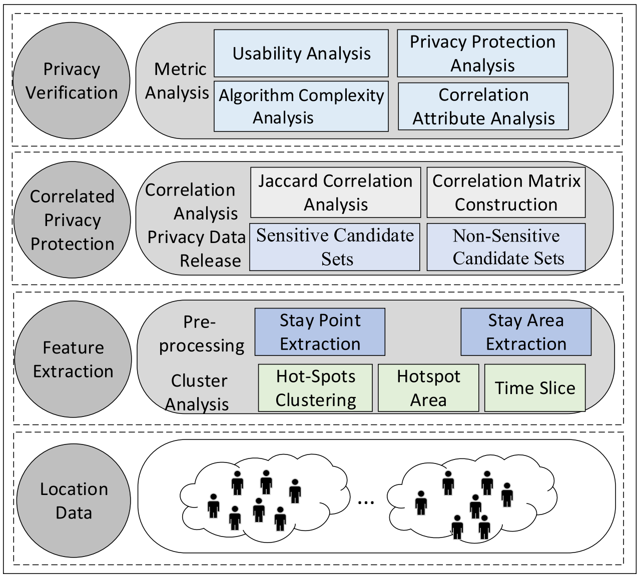


FIGURE 1. Diagram of the associated privacy protection model based on mobile feature analysis.

the group's relevant information cannot be extracted at all, and yet considering the time factor, the algorithm [19], [20] will have difficulty in practical application. Therefore, based on the above correlation characteristics, this paper proposes a position correlation mobile feature extraction algorithm. The basic process is that after extracting the individual stay points and stay areas, the group stay areas are constructed. After this, the cluster analysis is performed. According to group stay area obtained, the dynamic time alignment algorithm is used to merge and divide the time slice and ultimately the time correlated hotspot area is obtained.

Based on the hotspot areas obtained, we implement the correlated privacy protection. First, the correlation analysis is conducted, and on the basis of the analysis above, the release of private data is performed. The concrete details are as follows: At first, we construct the correlation degree matrix with Jaccard correlation coefficient. In view of the Jaccard correlation coefficient, the relevance between each user is calculated and the user correlation degree matrix is constructed to formalize the correlation information between users. Moreover, the correlation can be used to infer the common interests between users. The location correlated attributes of the users, therefore, need to be strictly protected.

Then we construct the concept of correlation sensitivity with the correlation degree matrix, which is the basis for the data release of sensitive candidate sets. After the correlation degree matrix is acquired, the correlation information between users can be extracted, and then the problem to be solved is how to extend differential privacy to protect the user's location correlated attributes. However, for traditional differential privacy, when using traditional global sensitivity for independent data record protection, there is not the problem. However, if the number of correlation records is directly multiplied by the global sensitivity, considerable redundant noise will be introduced resulting in a serious decrease in data availability [21]. Therefore, this article, in view of overcoming the problem of location correlated data protection, proposes a user correlation sensitivity so that the privacy protection of the correlation attributes between users satisfies the requirements of each sensitive set and introduces less noise.

Besides, we define sensitive candidate set for data release, where the candidate sets are the hotspot areas obtained above. These candidate sets can reflect a variety of social attributes of the user, such as residential addresses, interests, hobbies, and work addresses. At the same time, aiming at nonsensitive candidate sets, if pure Laplacian noise with differential privacy is utilized for the original data, the protected user's trajectory will be uneven, which seriously decreases the usability of the data. Furthermore, attackers can use filtering attacks to restore the data.

Finally, after finishing the location correlated differential privacy protection, we perform privacy verification including analysis of data availability, analysis of algorithm complexity, analysis of privacy protection and analysis of correlated attribute.

V. LOCATION CORRELATED DIFFERENTIAL PRIVACY PROTECTION ALGORITHM

A. POSITION CORRELATION MOBILE FEATURE EXTRACTION ALGORITHM

The construction of hotspots, in view of the current spatiotemporal data clustering analysis, simply clusters spatial data. The hotspots thus obtained are not time-consecutive, which creates biased planning in the data analysis. Aiming to this problem, we propose an efficient time-constructed population hotspot algorithm (dynamic time merge, DTM), in which the main idea is to obtain personal stay points, personal stay area and group stay area according to distance and speed factors for the original location point. Moreover, according to group stay area, the dynamic time alignment algorithm is used to merge and divide, and ultimately obtain a hotspot area with continuous time. The definition of the user's stay point is:

$$\boldsymbol{P_{ik}} = \{\boldsymbol{T_{ij}}, \boldsymbol{T_{ij+1}}, \ldots, \boldsymbol{T_{im}}\}$$

The details are as follows: All location points are numbered in order of 1-n, $\boldsymbol{t_j}$ represents j time, and $\boldsymbol{\delta_v}$ is speed threshold (50% of the normal person's moving speed). Besides, the coordinate of location point T_{ij} is defined as L_{ij} and the concrete content of $\boldsymbol{P_{ik}}$ is:

$$\boldsymbol{P_{ik}} = \{T_{ij} | \delta_v < \frac{|L_{ij+1} - L_{ij}|}{|t_{j+1} - t_j|}, \quad j = m, m+1, \ldots, H, 1 \leq m, H \leq n-1\} \quad (7)$$

We can conclude that traversing all the user's location points can be used to find the stay point of the user and these location points meet certain speed requirements and are continuous. Besides, if two adjacent stay points satisfy the speed requirements, the set of the user stay points can be obtained by merging these stay points into one stay point.

In addition, the average coordinates of stay point P_{ik} is:

$$\boldsymbol{L}\mathbf{p_{ik}} = \frac{L_{ij} + L_{ij+1} + \ldots L_{im}}{m - j + 1} \tag{8}$$

That is, for each stay point P_{ik}, the coordinates of the all location points in the stay points P_{ik} are averaged to obtain the average coordinates L_{Pik} corresponding to the stay point P_{ik}.

In order to construct group stay area, we will divide the process into two parts, when obtaining user's personal stay point. All stay points are numbered in order of 1-n, and $\delta_{\boldsymbol{D}}$ is distance threshold. The individual stay area is

$$\begin{aligned}\boldsymbol{A_{ik}} = \{&\boldsymbol{P}_{\mathrm{ij}} | \delta_{\boldsymbol{D}} < |\boldsymbol{Lp_{ij+1}} - \boldsymbol{Lp_{ij}}|,\\ &\mathrm{j = m, m+1, \ldots, H, 1 \le m, H \le n-1}\}\end{aligned} \tag{9}$$

Which means that the single stay points of user $\boldsymbol{M_i}$ which meet the distance requirement are combined into the user's individual stay area and these stay points in user $\boldsymbol{A_{ik}}$ are continuous.

Subsequently, there is an even higher level of abstraction that shifts the concept of the individual stay area ~~zone~~ towards an abstract layer becoming a group stay area $\boldsymbol{GA_k}$.

$$\begin{aligned}\boldsymbol{GA'_k} &= \{\boldsymbol{A}_{\mathrm{il}},\ \boldsymbol{A}_{\mathrm{jm}} | \delta'_D < |\boldsymbol{A}_{\mathrm{il}}\text{-}\boldsymbol{A}_{\mathrm{jm}}|, \mathrm{i,j,l,m} = 0,1,2\ldots\mathrm{n}\}\\ \boldsymbol{GA_k} &= \{Ga |\ \ \boldsymbol{GA'_k}\ \mathrm{size()} > \delta_{num}\ , Ga \in \boldsymbol{GA'_k}\ \}\end{aligned} \tag{10}$$

where δ'_D denotes the distance threshold and δ_{num} denotes the number threshold. The group stay area $\boldsymbol{GA_k}$ denotes a set of individual stay area $\boldsymbol{A_i}$ from different user, and these individual stay areas in the group stay area meet the requirements of distance and quantity. $\boldsymbol{GA'_k}$ is the transitional set. When $\boldsymbol{GA'_k}$ meets the requirements of δ_{num}, it can become a group stay area. However, the stay area does not possess the concept of time and sometime of the stay area may cross. The algorithm is as follows:

The first step is to scan the user's stay point set P_{ib} of the user M_i. In addition, P_{ik} and $P_{i(k+1)}$ are merged into a user stay area A_{im} if $\|\mathrm{L}_{\mathrm{P_{ik}}}\mathrm{L}_{\mathrm{P_{i(k+1)}}}\| \le \mathrm{P_{i(k+1)}}\delta_{\mathrm{D}}$ $\left\|L_{P_{ik}} - L_{P_{i(k+1)}}\right\| \le P_{i(k+1)}\delta_D$. Then, for each set of the stay areas $A_{ib,}$ average coordinates L_{Pik} corresponding to all the stay points P_{ik} corresponding to all individual stay areas A_{im} in the set are averaged to obtain the corresponding average coordinates L_{Aim}.

If each personal stay area A_{im} satisfies $\left\|L_{A_{im}} - L_{A_{qm}}\right\| \le \delta'_D$, $\|\mathrm{L}_{\mathrm{A_{im}}}\mathrm{L}_{\mathrm{A_{qm}}}\| \le \delta'_{\mathrm{D}}$ and the number of A_{qm} satisfying the requirement is greater than δ_{num}, then A_{im} and $A_{qm}\mathrm{A_{qm}}$ are combined into a group stay area GA_k, and A_{qm} meeting the requirement is searched in turn. If the requirement is not satisfied, this stay area is skipped until all the residences of all users are found. The algorithm is shown in algorithm 1.

Algorithm 1 Group Stay Area Extraction Algorithm

```
Input:
  $\boldsymbol{p} = \{\boldsymbol{p_{i1}}, \boldsymbol{p_{i2}}, \ldots, \boldsymbol{p_{ib}}\}$: collection of user i individual stay point
  $\delta_{\boldsymbol{D}}$: Distance threshold of individual stay area
  $\delta'_{\boldsymbol{D}}$: Distance threshold of group stay area
  $\delta_{\boldsymbol{num}}$: number threshold
Output:
  $\boldsymbol{GA} = \{\boldsymbol{GA_1}, \boldsymbol{GA_2}, \ldots, \boldsymbol{GA_0}\}$: User group stay are collection
Step:
m = 1
$\boldsymbol{Lp_{ij}}$ = ave($\boldsymbol{p_{ij}}$): average coordinate of all points in the
              stay point $\boldsymbol{p_{ij}}$
$\boldsymbol{A} = \{\boldsymbol{A_{11}}, \ldots, \boldsymbol{A_{nm}}\}$: collection of stay area
for i = 1:n
    for: k = 1:b-1
        distance = $\|\boldsymbol{Lp}_{\mathrm{ik}} - Lp_{\mathrm{i(k+1)}}\|$
        if distance $\le \delta_{\boldsymbol{D}}$
            merge $\boldsymbol{p}_{\mathrm{ik}}$ and $\boldsymbol{p}_{\mathrm{i(k+1)}}$ into $\boldsymbol{A}_{\mathrm{im:}}$
        else
            m = m + 1
        end if
    end for
end for
for i = 1:n and k = 1:m
    $\boldsymbol{LA}_{\mathrm{ik}}$ = ave($\boldsymbol{A}_{\mathrm{ik}}$): average coordinate of all stay points in the stay area $\boldsymbol{A}_{\mathrm{ik}}$
    end for
for $\boldsymbol{A}_{\mathrm{im}}$ in $\boldsymbol{A}$
    for $\boldsymbol{A}_{\mathrm{iq}}$ in $\boldsymbol{A}$
        distance = $\|\ \boldsymbol{L}\,\boldsymbol{A}_{\mathrm{im}} - \boldsymbol{L}\,\boldsymbol{A}_{\mathrm{qm}}\|$
        if distance $\le \delta'_{\boldsymbol{D}}$
            merge $\boldsymbol{A}_{\mathrm{im}}$, $\boldsymbol{A}_{\mathrm{qm}}$ into $\boldsymbol{GA_K}$
        end if
    end for
    K++;
    if $\boldsymbol{GA_k}$.size()< $\delta_{\boldsymbol{num}}$
        clear(GA$_{\mathrm{k}}$)
end for
```

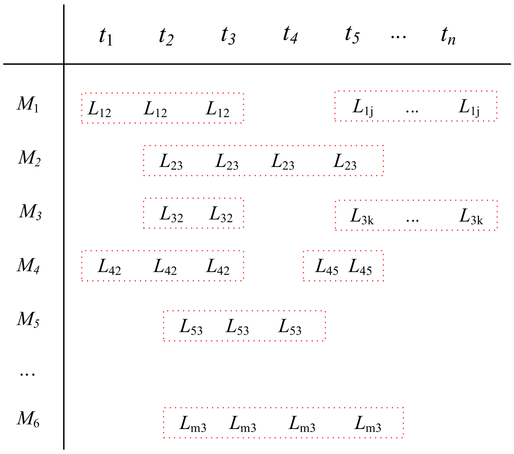
Finally, we will construct the concrete content of a hotspot area $\boldsymbol{HA_k}$. At first, all time in a group area are numbered in order of 1-n, and δ_M is the threshold of number of people and δ_T is continuous time interval threshold in a hotspot area. $\boldsymbol{HA_k}$ is expressed as

$$\begin{aligned}\boldsymbol{HA'_k} &= \{GA_i^j | GA_i^j(num) > \delta_M,\\ &\qquad j = m, m+1, \ldots, H, 1 \le m, H \le n\}\\ \boldsymbol{HA_k} &= (Ha|\boldsymbol{HA'_k}.size() > \delta_T, Ha \in \boldsymbol{HA'_k})\end{aligned} \tag{11}$$

where GA_i^j represents the ith group stay area at time j and $GA_i^j(num)$ represents the number of people in the stay area GA_i of time j. In addition, $\boldsymbol{HA'_k}$*.size()* is denoted as the number of group area in different time. $\boldsymbol{HA'_k}$ is the transitional set. When $\boldsymbol{HA'_k}$ meets the requirements of δ_T, it can become a hotspot area.

	t_1	t_2	t_3	t_4	t_5	…	t_n
M_1	L_{11}	L_{12}	L_{13}	L_{14}	L_{15}	…	L_{1n}
M_2	L_{21}	L_{22}	L_{23}	L_{24}	L_{25}	…	L_{2n}
M_3	L_{31}	L_{32}	L_{33}	L_{34}	L_{35}	…	L_{3n}
M_4	L_{41}	L_{42}	L_{43}	L_{44}	L_{45}	…	L_{4n}
M_5	L_{51}	L_{52}	L_{53}	L_{54}	L_{55}	…	L_{5n}
…	…	…	…	…	…	…	…
M_6	L_{m1}	L_{m2}	L_{m3}	L_{m4}	L_{m5}	…	L_{mn}

FIGURE 2. Raw data.



	t_1	t_2	t_3	t_4	t_5	…	t_n
M_1	L_{12}	L_{12}	L_{12}		L_{1j}	…	L_{1j}
M_2		L_{23}	L_{23}	L_{23}	L_{23}		
M_3		L_{32}	L_{32}		L_{3k}	…	L_{3k}
M_4	L_{42}	L_{42}	L_{42}	L_{45}	L_{45}		
M_5		L_{53}	L_{53}	L_{53}			
…							
M_6		L_{m3}	L_{m3}	L_{m3}	L_{m3}		

FIGURE 3. Individual stay point extraction.

Here we will introduce location correlated mobile feature extraction algorithm in detail which is shown in algorithm 2. In this work, first, the number C_i is calculated in time t_{ii} in each group stay area GA_o. If $C_i \geq \delta_M \mathrm{C_i} \geq \delta_\mathrm{M}$ where δ_M represents the number threshold, the current mark, which is responsible for recording the maximum length of the continuous time range, can be rewritten as $max_i = max_{i-1}+1$. In addition, the mark $begin_i$ can be written as $begin_i = begin_{i-1}$. If $C_i < \delta_M$, $max_i = 0$, $begin_i = i$. After scanning, time i is traversed from back to front, and if $max_i > \delta_T$ where δ_T denotes the continuous time interval threshold, the stay area of the time interval $[begin_i, i]$ is merged into a time-continuous hotspot stay area HA_k, $i = begin_i - -1$. If $max_i < \delta_T$, the current time does not satisfy the time continuity requirement, and then $i = i\ -1$. After the traverse is completed, HA_k is the population hotspot area satisfying the time continuity.

By dynamically merging and dividing the time slices, the hot spots of the time continuous groups are obtained. In addition, as illustrated in Figs. 2, 3, 4 and 5, there is a schematic diagram of the overall processing of a total data flow. Fig. 5 shows that the obtained time is a continuous hotspot area map.

Algorithm 2 Time-Related Hotspot Area Generation Algorithm

```
Input:
  $GA = \{GA_1, GA_2, \ldots, GA_0\}$: User group stay area collection
  $\delta_M$: Threshold of number of people
  $\delta_T$ :Continuous time interval threshold
Output:
  $HA = \{HA_1\ HA_2, \ldots, HA_k\}$: Group hotspot staying area satisfing time continuity
Step:
$k = 1$
for $j = 1$:o
  for $i = 1$:n
    //Calculate the number of people $C_i$ in the stay area at time $t_i$
    if $C_i \geq \delta_M$
      $max_i = max_{i-1} + 1$
    // Record the maximum length of the continuous time interval
      $begin_i = begin_{i-1}$
    // Record the mark of the initial position of the continuous time
    else
      $max_i = 0$
      $begin_i = i$
    end if
  end for
  while $i > 0$
    if $max_i > \delta_T$
      merge $GA$ into $HA_k$
      // Combine GA $GA$ when its time is included in $[begin_i, i]$
      $i = begin_i - 1$
    else
      $i = i - 1$
    $k = k + 1$
    end if
  end while
end for
```

B. LOCATION-ASSOCIATED PRIVACY PROTECTION DATA PUBLISHING ALGORITHM

In real life, the stronger the degree of correlation is between the two users, the more likely mapping to a specific location dataset is, which means that the number of times that the two users appear in the same location at the same time quantum is more, and the degree of relevance of their attributes is also greater. Therefore, it is necessary to find the degree of similarity between users to protect the location correlated information between group users. The chapter mainly introduces the location-associated privacy protection data

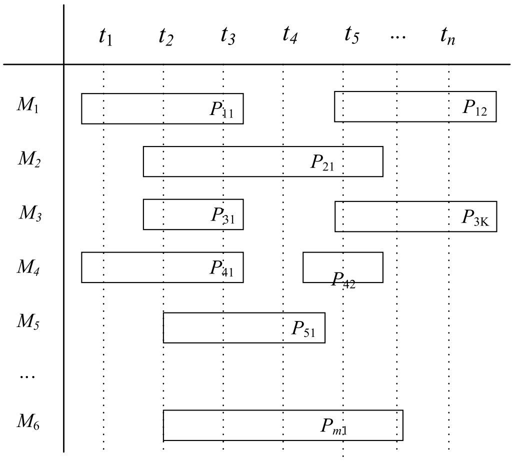


FIGURE 4. Individual stay area extraction.

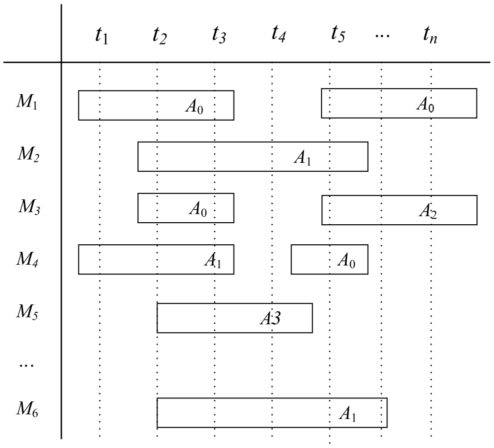


FIGURE 5. Group stay area.

Algorithm 3 Correlation Matrix Construction Algorithm

```
Input:
    $P$: User distribution matrix
    $\Delta$: User correlation matrix
    $m$: The number of users in a group
    $\Delta C$: Correlation sensitivity matrix
Output:
for i = 1:n do
    for j = 1:m do
        $\Delta C_{ij} = 0$
        if $A_{ij} = 1$
            for k = 1:m do
                $\Delta C_{ij} = P_{ik} \times \Delta_{kj} + \Delta C_{ij}$
            end for
        end if
    end for
end for
return $\Delta C$
```

```
Input:
    $C = \{C_1, C_2, \ldots, C_i, \ldots, C_n\}$: Sensitive Candidate Set
    $\Delta C$ : Correlation sensitivity matrix
    $\varepsilon` = \{\varepsilon_1, \varepsilon_2, \ldots, \varepsilon_i, \ldots, \varepsilon_n\}$: Privacy Budget Set
Output:
    $C_t$: Privacy protection location set for sensitive candidate sets
for i = 1:n do
    foreach $(x_i, y_i) \in C_i$
        select $w_i$ from Laplace $(\Delta C_{ij}/\varepsilon_i)$
        $(x_j', y_j') = (x_j, y_j) + w_i$
    end for
return $C_t$
```

publishing algorithm based on related difference algorithm. Then we evaluate the degree of similarity between the users, and use Jaccard correlation coefficient to construct the group similarity matrix, and use similarity matrix and user distribution matrix to construct the user individual correlation degree matrix, which is the basis for the data release of sensitive candidate sets.

The previous study focused only on personal attributes and assume that the dataset is independent and the internal data records are also independent. There are problems in practical application. Because, in real life, people have their social circles, and they often engage in certain social activities. If the personal privacy of the user is protected but their associated attributes are ignored, users can be attacked, which can lead to the exposure of personal information.

This paper first extracts the correlations between group users and then protects the location correlated information between users. As shown in Fig. 6, the distribution of users in each hotspot area is counted and the results are stored in the user's distribution matrix. Then, the users are used as the main body to count the hotspot areas where each user has ever appeared and to obtain sets of hotspot area where each user has stayed. As a result, the similarity between two users according to the Jaccard coefficient can be calculated. The Jaccard similarity coefficient definition is defined below:

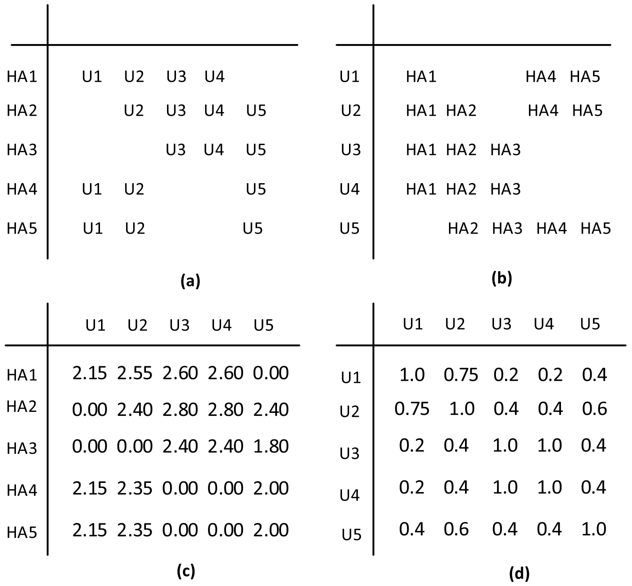
Definition 7 (Jaccard Similarity Coefficient): The ratio of the intersection elements of two sets X and Y in the union of X and Y is called the Jaccard similarity coefficient of two sets which can be expressed by J(X,Y) as below:

$$J(X, Y) = \frac{|X \cap Y|}{|X \cup Y|} \tag{12}$$

Definition 8 (User Correlation Degree): This article assumes that two users M_i and M_j are correlated, in which the relationship is expressed as the degree of correlation $\delta_{ij} \in [0, 1]$.

$$\delta_{ij} = \frac{\sum_1^m HA_i \cap \sum_1^n HA_j}{\sum_1^m HA_i \cup \sum_1^n HA_j} \tag{13}$$

where $\sum_1^m HAi$ is the sum of the hotspot area of user M_i and $\sum_1^n HAj$ is the sum of the hotspot area of user M_j.



HA1	U1	U2	U3	U4	
HA2		U2	U3	U4	U5
HA3			U3	U4	U5
HA4	U1	U2			U5
HA5	U1	U2			U5

(a)

U1	HA1			HA4	HA5
U2	HA1	HA2		HA4	HA5
U3	HA1	HA2	HA3		
U4	HA1	HA2	HA3		
U5		HA2	HA3	HA4	HA5

(b)

	U1	U2	U3	U4	U5
HA1	2.15	2.55	2.60	2.60	0.00
HA2	0.00	2.40	2.80	2.80	2.40
HA3	0.00	0.00	2.40	2.40	1.80
HA4	2.15	2.35	0.00	0.00	2.00
HA5	2.15	2.35	0.00	0.00	2.00

(c)

	U1	U2	U3	U4	U5
U1	1.0	0.75	0.2	0.2	0.4
U2	0.75	1.0	0.4	0.4	0.6
U3	0.2	0.4	1.0	1.0	0.4
U4	0.2	0.4	1.0	1.0	0.4
U5	0.4	0.6	0.4	0.4	1.0

(d)

FIGURE 6. **User correlation extraction.**

Definition 9 (Correlation Degree Matrix): The degree of correlation between group users is used to constitute the correlation degree matrix, which is expressed as follows:

$$\Delta = \begin{bmatrix} \delta_{11} & \delta_{12} & \cdots & \delta_{1n} \\ \delta_{21} & \delta_{22} & \cdots & \delta_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ \delta_{n1} & \delta_{n2} & \cdots & \delta_{nn} \end{bmatrix} \tag{14}$$

Definition 10 (ε-Differential Privacy): A mechanism A gives ε-differential privacy for any pair of D and D ' and for every set of outcomes S, the randomized mechanism A satisfies:

$$\Pr[A(\mathrm{D}) \in \mathrm{S}] \leq \exp(\varepsilon)\times\Pr[A(\mathrm{D}') \in \mathrm{S}] \tag{15}$$

Definition 11 (Global Sensitivity): For $f : \mathrm{D} \rightarrow \mathrm{R}$, the global sensitivity of Q is defined as

$$GS = \max_{\mathrm{D},\mathrm{D}'} \|f(\mathrm{D})\text{-}f(\mathrm{D}')\|_1 \tag{16}$$

For correlated data, a common solution is to use global sensitivity, which inevitably introduces a lot of noise. Therefore the paper proposed an important concept of correlated degree, whose relatively intuitive explanation is that most records are partially linked with others.

Definition 12 (Correlation Sensitivity): For the given Δ and query function Q, the individual correlation sensitivity of user M_i is:

$$CS_i^Q = \sum\nolimits_{j=0}^{n} \left|\delta_{ij}\right|(\left\|Q(D^j) - Q(D^{-j})\right\|_1) \tag{17}$$

Individual correlation sensitivity denotes the influence on the other records when deleting the records of user M_i, where $\delta_{ij} \in \Delta$ denotes the degree of user correlation between the records. If the dataset is an independent dataset, CS_i^Q is equal to the global sensitivity.

After the individual correlation sensitivity is defined, the data protection method is defined as:

$$\hat{Q}(D) = Q(D) + Laplace(CS_i^Q \big/ \varepsilon) \tag{18}$$

Definition 13 (Correlation Sensitivity Matrix CS): The relation of correlation sensitivity between group data is expressed by correlation sensitivity matrix CS, where the element CS_i^Q represents the correlation sensitivity of user i in query Q, and the number of rows of the correlation sensitivity matrix is the number of hotspot areas and the number of columns is the number of users. The algorithm is as follows:

Definition 14 (Sensitive Candidate Set): The stay area which can reflect the user's various interests, is regarded as a sensitive candidate set, where candidate sets are the hotspot areas obtained above.

Assuming that $(\mathrm{x_j}, \mathrm{y_j}) = \mathrm{Q}(\mathrm{C_i})$, according to the formula above, differential privacy for sensitive candidate set C_i is calibrated by the following equation:

$$(\mathrm{x}_\mathrm{j}', \mathrm{y}_\mathrm{j}') = (\mathrm{x_j}, \mathrm{y_j}) + \mathrm{Laplace}(\frac{CSij}{\varepsilon i}) \tag{19}$$

The privacy protection algorithm is as follows:

C. SECURITY ANALYSIS

The section mainly introduces, after the data set is processed by algorithm, the security analysis of the data set, and the protection effect of related attributes, which is strictly verified theoretically. The discussion on the security issues of position-associated correlation is analyzed mainly from the following two perspectives. First, the work involves privacy protection for individuals, and the second is the protection of associated privacy between group users.

First, privacy protection of the individual user's work location in that the true trajectory of the individual user is processed and published through an intuitive representation, which protects the user's real-track. Formally, the published individual data of the user actually satisfies the definition of ε-difference privacy; namely, it meets:

$$\frac{\Pr[A(D) = O]}{\Pr[A(D') = O]} \leq e^{\varepsilon} \tag{20}$$

where D and D' represent a brother dataset which has only the difference of one data record, A denotes a mapping function, and ε denotes a privacy budget.

Below we will prove that this algorithm satisfies ε-differential privacy.

Proof: trajectory data in this paper $T = \{(x_1, y_1, t_1, \ldots, p_1), (x_2, y_2, t_2, \ldots, p_2), \ldots, (x_n, y_n, t_n, \ldots, p_n)\}$ is an n-dimensional dataset in which the attributes of the tuples $(x_n, y_n, t_n, \ldots, p_n)$ are independent of each other, so this article only discusses attribute x_n and other attributes prove to be similar. Assuming that Laplacian noise is added to $f(D) = (x_1, x_2, \ldots, x_n)^T$, we can acquire the mapping function $A(D)$. Therefore, $A(D) = f(D) + \left(Lap_1(\Delta f \big/ \varepsilon), Lap_2(\Delta f \big/ \varepsilon), \ldots, Lap_n(\Delta f \big/ \varepsilon)\right)^T$, where $\Delta f = \max\limits_{D,D'} \left\|f(D) - f(D')\right\|_p$, in which p is equal to 1. Then, we assume $f(D') = (x'_1, x'_2, \ldots, x'_n)^T = (x_1 + \Delta x_1, x_2 + \Delta x_2, \ldots, x_n + \Delta x_n)^T$, and obtain $\Delta f = \max(\sum_{i=1}^{n} \left|x_i - x'_i\right|) = \max(\sum_{i=1}^{n} |\Delta x_i|)$.

Assume that the output vector is $O = (y_1, y_2, \ldots, y_n)^T$. Because *lap*(λ) is the Laplace distribution which satisfies the probability distribution $\Pr(x|\lambda) = {}^{1}\!/_{2\lambda} \exp(-{}^{|x|}\!/_{\lambda})$, $\Pr(A(D') = O) = \prod_{i=1}^{n} {}^{\varepsilon}\!/_{2\Delta f} \exp(-\frac{\varepsilon}{\Delta f} |x_i + \Delta x_i - y_i|)$ is obtained as well as $\Pr(A(D') = O) = \prod_{i=1}^{n} {}^{\varepsilon}\!/_{2\Delta f} \exp(-\frac{\varepsilon}{\Delta f} |x_i + \Delta x_i - y_i|)$. Therefore, the following formula is obtained:

$$\frac{\Pr[A(D) = O]}{\Pr[A(D') = O]} = \exp(-{}^{\varepsilon}\!/_{\Delta f} \sum_{i=1}^{n} (|x_i - y_i| - |x_i + \Delta x_i - y_i|)) \quad (21)$$

Therefore, to prove formula (5), we need only prove:

$$e^{-{}^{\varepsilon}\!/_{\Delta f} \sum_{i=1}^{n} (|x_i - y_i| - |x_i + \Delta x_i - y_i|)} \leq e^{\varepsilon}$$

Namely to prove:

$$\sum_{i=1}^{n} (|x_i - y_i| - |x_i + \Delta x_i - y_i|) \leq \Delta f \quad (22)$$

Because $\Delta x_i > 0$, $|x_i + \Delta x_i - -y_i| - |x_i - -y_i| > 0$. Then, we use the absolute value inequality and $|x_i + \Delta x_i - -y_i| - |x_i - -y_i| \leq |x_i + \Delta x_i - -y_i - -(x_i - -y_i)| = |\Delta x_i|$. So $\sum_{i=1}^{n} (|x_i - y_i| - |x_i + \Delta x_i - y_i|) \leq \sum_{i=1}^{n} |\Delta x_i| \leq \max(\sum_{i=1}^{n} |\Delta x_i|) = \max(\sum_{i=1}^{n} |x_i - x'_i|) = \Delta f$.

So, formula (5) is proven to be correct.

Therefore, the protection algorithm for this property satisfies ε-differential privacy. Moreover, because the proof of the other properties is similar to the process, the algorithm satisfying ε-differential privacy can adequately protect the privacy of users. Below we analyze the location-correlation protection of group users. Intuitively, position correlation information is hidden due to the protected correlation properties between users, and formally, this article uses conditional probability to describe location-correlated protection. Therefore, this article, to achieve location-correlated protection, needs not only the location publishing mechanism but also relies on the prior probability of the attacker. Hence, it is usually necessary to analyze the feature of the probability distribution of the position trajectories. In particular, the attacker may infer the position of the two users according to the prior probability knowledge of the observed information. For example, two users with the same attributes may all be students. Then, the attacker analyses the sensitive attributes of the two users by observing that the two users appear at the same point at the same time. Under this circumstance, the goal of this paper is not only to protect the location attributes of the two users but also to limit the correlation impact for the users after the trajectory is released. Specifically, we let the prior probability distribution before the trajectory release be similar to the posterior probability distribution after the release. From the above argument, it can be seen that the probability distribution after protection meets the requirements of the user attack model in Definition 1.

TABLE 2. Experiment parameter.

Parameter	Original value
δ_M	10
δ_T	50 minutes$_i$
δ_V	3 km/h
δ_D	200m
δ_{num}	10
ε	0.1

VI. EXPERIMENT AND RESULT ANALYSIS

The experiment is carried out on a machine with Intel i5-2400 3.1Ghz with 8GB RAM and Windows 7 Ultimate. And the algorithm is implemented using Matlab R2012a. The environment for these experiments is a simulation environment configured in Java. Moreover, the experimental parameters involved in the experiment are shown in table 2.

A. TIME COMPLEXITY AND HEAT VALUE XPERIMENT

Because of the characteristics of the high dimension of location and the need to consider the time factor to construct the correlation information, the process of constructing the group hot spot area is complex, making it the first problem that needs to be solved. Therefore we first analyzed the time complexity and the heat value of the proposed algorithm when constructing the hotspot area.

In this paper, we refer to the idea of dynamic programming and utilize some tables to save the intermediate results, which can greatly cut down the running time of the algorithm. The time complexity of the hot spot extraction algorithm (DTM) proposed by this paper is analyzed below. Firstly, assuming that the number of users' location points is n, the construction of personal stay points only needs to traverse the user's trajectory to find the location points that meet the requirements and merge the required location points into individual user stay points, and the time complexity of the process is O(n). Then we extract of hotspot area. Assuming that the number of stay points is m, the time complexity of clustering hotspot areas is O(m^2). Therefore, the total time complexity of the algorithm in this paper is O(m^2), in which m is a number far less than n.

GENE and DTM are the two algorithms for comparison in this paper. This GENE is a single-dimensional clustering algorithm. After obtaining the hotspot area, the algorithm performs cluster on the time dimension to obtain the time-continuous hotspot areas, and the time complexity is O(n^3). This STHSRD algorithm maps all the data to the time axis first, and then forcibly intercepts the position data of the specified time slice. The time complexity of the algorithm is also O(m^2), but the resulting hotspot areas are not time-correlated. The experimental results are shown in figure 7. At the position of one million position points, the running time of the algorithm in this paper is shortened by 28.43%,

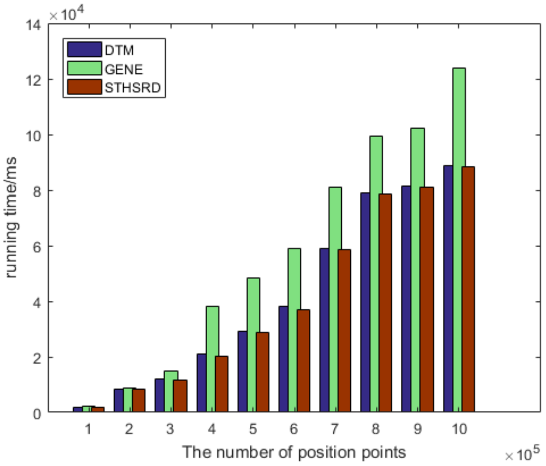


FIGURE 7. Running time chart.

compared with the algorithm (GENE). Although the running time of the STHSRD algorithm is roughly the same, the hot spots obtained by the algorithm in this paper are not time-correlated.

In addition, in the construction of time-continuous hotspot areas, the hotspot areas should accord with the user's actual stay time as much as possible so that the extraction of user association attributes can be carried out to the maximum extent. For measuring the degree of correlation between time and user hotspot areas, this article uses the concept of heat value as defined above. In this paper, three comparative algorithms DTM, GENE and STHSRD are performed in order to make a comparison among their generated the heat values. As shown in figure 8, the heat value of the three algorithms go up with the increase of the number of data sets. At the scale of one million position points, the heat value generated by the algorithm (DTM) in this paper is 93.04, only 3.29% lower than the heat value (96.23) generated by GENE algorithm. However, GENE has a greater time complexity. Meanwhile, for the STHSRD algorithm, as the algorithm is forced to intercept the time slice, the time correlation of the obtained hot spots is not very strong. At the scale of one million position points, the heat value generated by the algorithm (STHSRD) in this paper is only 82.93, which is 10.11% lower than the heat value generated by the algorithm (DTM).

B. TIME SLICE LENGTH ANALYSIS EXPERIMENT

This paper analyzes the correlated attributes between users by extracting the time correlated hotspot areas of group users. Because the time properties are considered, a good time slice length needs to be selected. As shown in figure 9, 300,000, 600,000 and 900,000 location points are selected for experiment in this paper. The abscissa is the length of the time slice, lasting from 10 minutes to 120 minutes, and the ordinate is the number of the hotspot areas. The experiment of the algorithm shows that the number of hotspot areas dropped rapidly from 10 minutes to 50 minutes, and then the after 50 minutes, the declining trend remains relatively stable.

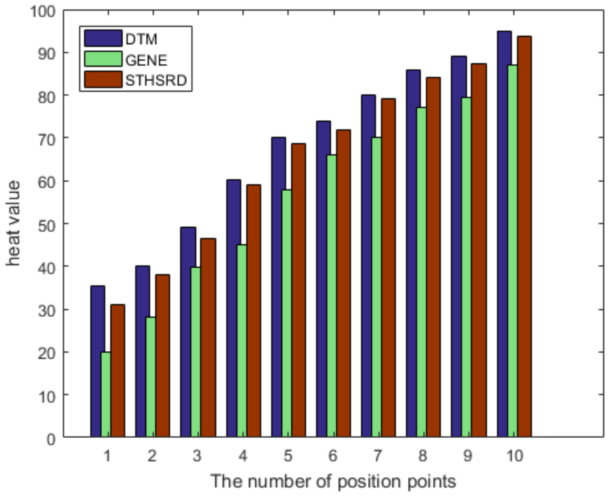


FIGURE 8. The heat value under different algorithms.

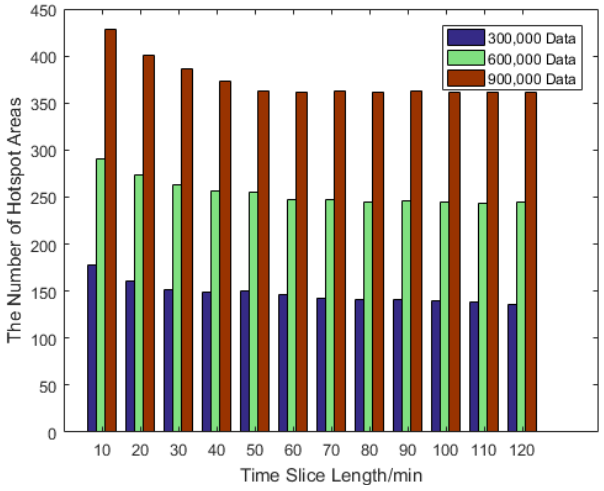


FIGURE 9. The graph of time period and hotspot area relationship.

Through the analysis of experiment, before 50 minutes, because the time slice is relatively small, the obtained hot spots are scattered, which leads to a rapid decline in the number of hot spots. However, after 50 minutes, the number of hot spots declines slightly, but the overall trend remains stable. Taking the data of 900,000 position points as an example, the number of user hot spots will decrease with the continuous increase of time slice length. When dropping to 355 after about 50 minutes, the number remained basically stable. Therefore, based on the above analysis, when the time slice is about 50 minutes, the obtained group hot spots will have a relatively good comprehensive attribute

In addition, because this paper first analyzes the time-correlated hotspot area of the group location data, the user similarity analysis based on the results is analyzed to construct the user similarity matrix. The choice of the length of the time slice is also an important problem in construction of user similarity matrix. Five pairs of users with higher similarity are selected as experimental data for comparison

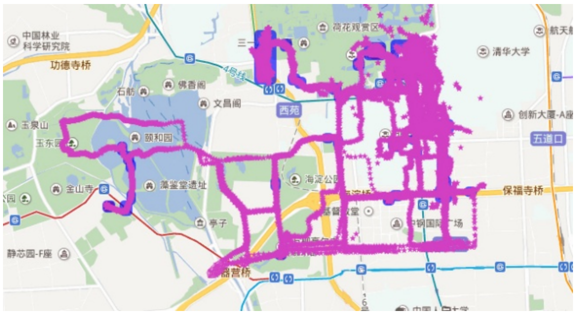


FIGURE 10. High relevance users diagram.

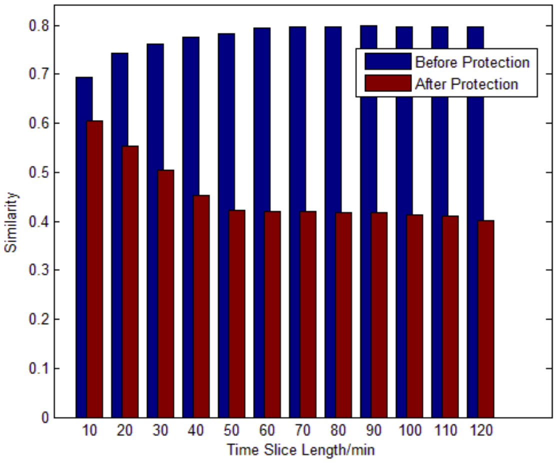


FIGURE 11. Relationship between the time period and user similarity in raw data.

of similarity in this paper. According to Jaccard's similarity coefficient, their similarity calculations are {0.8134, 0.8021, 0.7331, 0.8323, 0.7854}, where the calculated average is 0.7933. The article maps their data points individually to the map shown in Fig. 10, where it can be seen that these areas are located near the Peking University area through analysis, which means that these users are all students that visit the Summer Palace and Yuanmingyuan Park after school hours. As a result, these users have a higher positional similarity. The experiments on positional similarity performed below are based on the above-selected five groups of user position data for the verification analysis of the experiment.

The primary research analyses the user relevance of the time series selection for the original data, the effect of the construction of the influenced user correlation matrix, and how the user's correlation coefficient changes with the increase in the time period. As illustrated in Fig. 11, we can see that the abscissa is the time period from 10 minutes to 120 minutes, and the ordinate is the average of the similarity of the above five sets of user data. With the increase in the time period, the users' similarity from the figure will increase rapidly and then maintain a relatively stable trend, and then increase slightly to the highest value. By analyzing the experiment formed in the hotspot area, we can see that when the time period is short, the user will form many hotspots, while the same hotspots between the users are relatively fewer. Therefore, when the time period is short, the user's similarity obtained by the Jaccard similarity coefficient will be relatively small, but with an increase in the time period, the same hotspot area between users will increase, and the total hotspot area formed by the group users will decrease, which results in an increase in similarity among the total users. From Fig. 10 the user's similarity will be at a relatively stable level and increase with time up to an hour. In addition, because the users in most of the hotspot areas stay for approximately this time period, the merger of hotspots in the future is not obvious, which causes the similarity between users to be at a stable level. Therefore, this paper, combining the experimental analysis in Fig. 11, analyses the following data in terms of the time length of 50 minutes. Because the time period is too short to result in reasonable user similarity, many hotspots are excessively divided. However, overconsolidation of hotspots is caused if the chosen time is too long. The above two experiments show that time slice length of 50 is the optimal choice.

C. SENSITIVITY ANALYSIS EXPERIMENT

After selecting the appropriate length of time slice, the selection of the parameter ε is important in the differential privacy protection algorithm, where ε represents the privacy budget, which means that ε represents the protection level of the algorithm for the dataset. When ε decreases, the degree of protection of the algorithm increases, but ε depends on the actual situation so the value for the algorithm does not need to be as small as possible. In these experiments, NCDP FDP and SADP are three algorithms for comparison. Where NCDP is a method of cryptography that aims to provide a way to maximize the accuracy of data queries, and minimize the opportunity to identify its records when querying from a statistical database and FDP is a full-correlation differential privacy protection which is to model the relationship between the traffic volume and simple statistics about flows using a Hidden Markov Model and proposed by Chen et al. [22]. SADP is the algorithm proposed by this paper. According to our experiment, first, for the analysis of the availability of user datasets, this paper takes the user's trajectory distance for the experiment. From Fig. 12, the abscissa is size ε which varies from 0.1 to 1, and the ordinate expresses the track distance after protection, which causes ε to decrease, and the degree of protection of the three algorithms increases. However, for the fully correlated differential protection algorithm, the track distance reaches 1588.7 metres when the value of ε is 0.1, which produces considerable noise in the data. As ε becomes smaller, the degree of protection increases dramatically. For the privacy protection algorithm, regardless of the correlation protection and the adaptive correlation location protection algorithm, the performance is relatively good. When ε is 0.1, the protection degree of the two algorithms is 229.47 meters and 460.41 meters. Compared with the full-correlation protection algorithm, there is good data availabil-

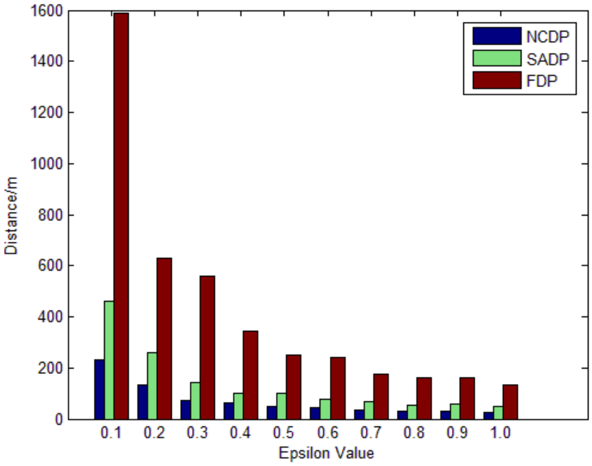


FIGURE 12. Trajectory distance.

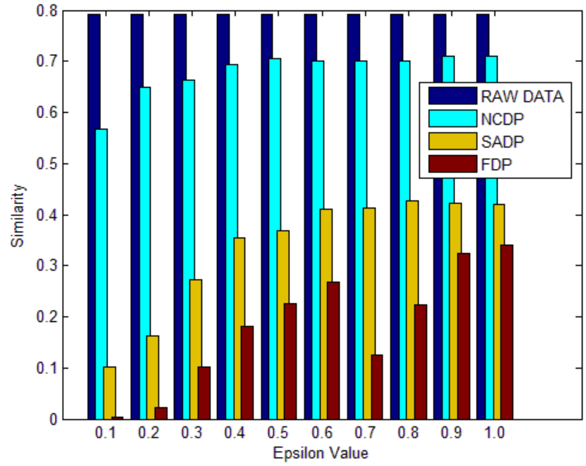


FIGURE 13. User similarity change chart.

ity, and the degree of change is relatively small, which gives rise to a good protection effect for the availability of the data.

The change degree of similarity of the user with the change of ε can be seen in Fig. 13. The degree of similarity of a user for unprotected data is 79.29%. For the differential privacy protection algorithm, without considering the correlation protection, although the similarity of the user decreases, the degree is relatively unaffected. When ε is 0.1, the user's similarity decreases to 56.84%. Because the user's location correlation is not considered, the degree of correlation with the user has hardly changed. For the full-correlation differential privacy protection algorithm and the adaptive correlation differential privacy protection algorithm, there is a relative effect on the user's similarity protection. Additionally, when ε is 1, the user similarity of the adaptive correlation differential privacy protection algorithm is 42.23%, and the fully correlated differential protection algorithm is 31.32%. The full-correlation differential meaning protection algorithm with a similarity of 31.32% decreases as ε decreases. In the case where ε is equal to 0.1, the user similarity obtained by the two algorithms is 10.23% and 0.31%, respectively. At the same time, as shown in Fig. 11, when ε is 0.1, the user's similarity is the lowest. Finally, According to the above similarity and data availability experiments, when the parameter ε is equal to 0.1, the performance of this algorithm is optimal.

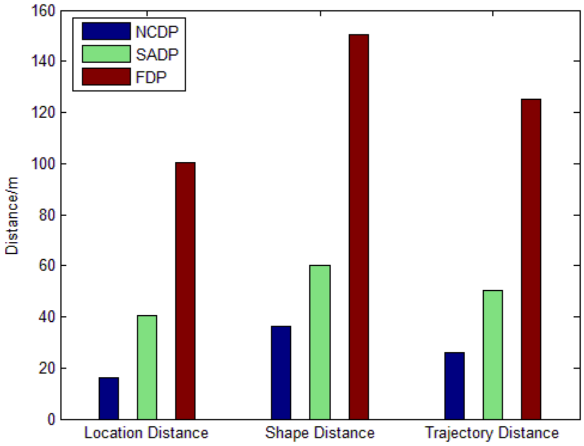


FIGURE 14. Distance measure analysis graph when $\varepsilon = 0.1$.

D. USABILITY ANALYSIS EXPERIMENT

In the section, we perform the experiment of the location correlated data protection to obtain optimal data performance, and we need to think about degree of availability and similarity of a user's data after data protection. According to the above experimental analysis of the selection of the parameter ε, under the condition of $\varepsilon = 0.1$, we carry out the comparative experiment of three algorithms (NCDP, FDP and SADP). In this paper, the measurement of data availability uses the three distance measurement presentations described above, namely, location distance, shape distance, and tracking distance. At the same time, the paper adopts the Pearson correlation coefficient for the user similarity measure after data protection. We can determine that the analysis of the distance metric experiment under the condition of $\varepsilon = 0.1$ is that the degree of protection of the data is 25.98 meters for the differential privacy protection irrespective of location correlation. However, in the case where the degree of similarity of the users obtained is up to 58.12%, when the data are not protected, the similarity of the users is 79.29%, which demonstrates that the degree of correlation for the users is reduced. For fully correlated differential privacy protection, the perturbation of the data under the same parameters is relatively large at 125.29 meters. At the same time, a relatively good data similarity level of 10.04% is guaranteed. In addition, because the algorithm proposed in this paper protects the user's adaptive correlation degree, the degree of data protection obtained in this paper is 50.34 meters under the same parameters. Compared with fully correlated differential privacy, this reduces 150% of the noise, reducing the noise to 10.31% for the correlation degree of the user.

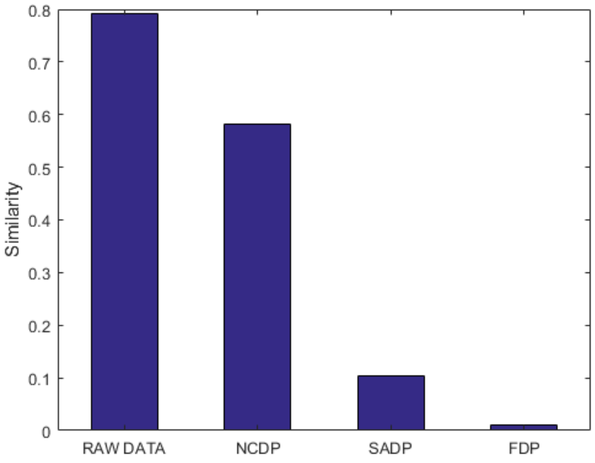


FIGURE 15. Similarity measure analysis graph when $\varepsilon = 0.1$.

From the above narrative, it is shown that the algorithm proposed in this paper has a relatively good overall performance.

VII. DISSCUSION

This paper focuses on the location correlated differential privacy protection. In the process of data protection, both the protection of user location data and the protection of user correlated information need to be considered. To solve this problem, we design a time - continuous hot spot extraction algorithm to obtain time - continuous hotspot area. Then the traditional differential privacy is extended, and we propose a personalized location data publishing mechanism by using the user correlation matrix which is based on the obtained group hotspot areas. That is, the user individual sensitivity matrix is constructed to guide the release of data by combining the user correlation matrix and user distribution.

Compared with previous approaches to location privacy protection, our method possesses some edges. Firstly, the time factor of location information is taken into consideration, and at the same time, the method can not only protect the user's location information, but also protect the multi-user's correlation information. Furthermore, in order to reduce the noise of differential privacy in privacy protection, we propose the concept of individual user correlation sensitivity, which makes full use of different correlations to personalize the user protection and reduce the introduction of noise. Besides, the proposed method has a relatively small time complexity. For example, the time complexity of hot spot extraction algorithm adopted by this paper is $O(m^2)$. The process of constructing the user correlation sensitivity is translated into the construction of the matrix, which also further improves the efficiency of the algorithm.

However, the algorithm itself is not perfect, and needs to be improved in the following aspects. This paper mainly aims at offline group location data to release, but for online data we can only perform online protection measures and can't update the relevant information in real time. In addition, when dealing with big data user feature analysis, this algorithm needs long time to run. Aiming to these problems, this algorithm may be optimized by using big data processing tools in our future work.

VIII. CONCLUSION

To solve the problem of location correlated privacy protection, this paper proposes a time - continuous hot spot extraction algorithm and a personalized correlation location data publishing mechanism. To solve the problem of discontinuity of time for group hot spots in the current study, we propose a cluster hotspot analysis method for Spatio-Temporal Series data, and also put forward to the necessity of analysis with time correlation to get the hotspot areas. In addition, the paper which refers to the idea of dynamic programming algorithms achieves an improvement for the traditional algorithm, which obtains time-correlated hotspot areas within linear time complexity, and short the algorithm time for the analysis of large data. At the same time, in the future work, the optimization issue of the parameters of the specific algorithm will be further studied.

ACKNOWLEDGMENT

The authors would like to thank the editor and the anonymous reviewers for their constructive comments and suggestions, which improve the quality of this paper.

REFERENCES

[1] Y. Xiao and L. Xiong, "Protecting locations with differential privacy under temporal correlations," in *Proc. 22nd ACM SIGSAC Conf. Comput. Commun. Secur.*, Oct. 2015, pp. 1298–1309.

[2] J. Hua, Z. Shen, and S. Zhong, "We can track you if you take the metro: Tracking metro riders using accelerometers on smartphones," *IEEE Trans. Inf. Forensics Security*, vol. 12, no. 2, pp. 286–297, Feb. 2017.

[3] M. Han *et al.*, "Cognitive approach for location privacy protection," *IEEE Access*, vol. 6, pp. 13466–13477, Mar. 2018.

[4] X. Shu, D. Yao, and E. Bertino, "Privacy-preserving detection of sensitive data exposure," *IEEE Trans. Inf. Forensics Security*, vol. 10, no. 5, pp. 1092–1103, May 2015.

[5] T. Wang *et al.*, "Trajectory privacy preservation based on a fog structure for cloud location services," *IEEE Access*, vol. 5, pp. 7692–7701, May 2017.

[6] W. Chao, Y. Jing, and Z. Jian-Pei, "Privacy preserving algorithm based on trajectory location and shape similarity," *J. Commun.*, vol. 36, no. 2, pp. 144–157, Feb. 2015.

[7] T. Peng, Q. Liu, G. Wang, Y. Xiang, and S. Chen, "Multidimensional privacy preservation in location-based services," *Future Gener. Comput. Syst.*, vol. 93, pp. 312–326, Apr. 2019.

[8] S. Hayashid, D. Amagata, T. Hara, and X. Xie, "Dummy generation based on user-movement estimation for location privacy protection," *IEEE Access*, vol. 6, pp. 22958–22969, Apr. 2018.

[9] L. Zheng, H. Yue, Z. Li, X. Pan, M. Wu, and F. Yang, "*k*-anonymity location privacy algorithm based on clustering," *IEEE Access*, vol. 6, pp. 28328–28338, Dec. 2017.

[10] M. Yang, T. Zhu, Y. Xiang, and W. Zhou, "Density-based location preservation for mobile crowdsensing with differential privacy," *IEEE Access*, vol. 6, pp. 14779–14789, Mar. 2018.

[11] K. Gu, L. Yang, Y. Liu, and B. Yin, "Efficient trajectory data privacy protection scheme based on Laplace's differential privacy," *Informatica*, vol. 42, no. 3, pp. 407–415, Sep. 2018.

[12] H. Wang and Z. Xu, "CTS-DP: Publishing correlated time-series data via differential privacy," *Knowl.-Based Syst.*, vol. 122, pp. 167–179, Apr. 2017.

[13] L. Cao, "Non-IIDness learning in behavioral and social data," *Comput. J.*, vol. 57, no. 9, pp. 1358–1370, Sep. 2014.

[14] L. Cao, Y. Ou, and P. S. Yu, "Coupled behavior analysis with applications," *IEEE Trans. Knowl. Data Eng.*, vol. 24, no. 8, pp. 1378–1392, Aug. 2012.

[15] H. Li, L. Xiong, and X. Jiang, "Differentially private synthesization of multi-dimensional data using copula functions," in *Proc. Adv. Database Technol., Int. Conf. Extending Database Technol.*, vol. 2014, Mar. 2014, pp. 475–486.

[16] T. Zhu, G. Li, W. Zhou, and P. S. Yu, "Correlated differential privacy for non-IID datasets," in *Differential Privacy and Applications. Advances in Information Security*, vol. 69. Cham, Switzerland: Springer, 2017. [Online]. Available: https://link.springer.com/chapter/10.1007%2F978-3-319-62004-6_14#citeas

[17] V. M. Sundaram and A. Thangavelu, "A delaunay diagram-based min–max CP-tree algorithm for spatial data analysis," *Wiley Interdiscipl. Rev., Data Mining Knowl. Discovery*, vol. 5, no. 3, pp. 142–154, Apr. 2015.

[18] S. Wang and H. Yuan, "Spatial data mining: A perspective of big data," *Int. J. Data Warehousing Mining (IJDWM)*, vol. 10, no. 4, pp. 50–70, Oct. 2014.

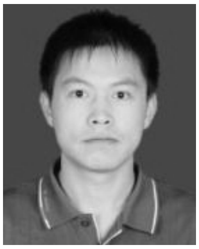
[19] X. Zhang, X. Gui, and F. Tian, "A framework for measuring query privacy in location-based service," *KSII Trans. Internet Inf. Syst.*, vol. 9, no. 5, pp. 1717–1732, May 2015.

[20] S. Gao, J. Ma, W. Shi, G. Zhan, and C. Sun, "TrPF: A trajectory privacy-preserving framework for participatory sensing," *IEEE Trans. Inf. Forensics Security*, vol. 8, no. 6, pp. 874–887, Jun. 2013.

[21] J. T. Meyerowitz and R. R. Choudhury, "CacheCloak: Enabling realtime location privacy for mobile users," *ACM SIGMOBILE Mobile Comput. Commun. Rev.*, vol. 13, no. 3, pp. 38–41, Jul. 2009.

[22] Z. Chen, J. Wen, and Y. Geng, "Predicting future traffic using hidden Markov models," in *Proc. IEEE 24th Int. Conf. Netw. Protocols (ICNP)*, Nov. 2016, pp. 1–6.

ZHENLONG PENG was born in 1977. He received the B.S. degree from Gannan Normal University, Ganzhou, China, in 1998, and the M.S. degree from Huaqiao University, Quanzhou, China, in 2004. He is currently pursuing the Ph.D. degree with the Department of Computer Science and Technology, Xi'an Jiaotong University, Xi'an, China. He is also a Researcher with the Key Laboratory of Computer Network, Xi'an Jiaotong University, and an Associate Professor with Quanzhou Normal University, Quanzhou. He has been researching digital image processing and information hiding. He is currently studying in crowdsourcing, crowd sensing, privacy protection, and the Internet of Things. His research interests include social networks and business intelligence.



JIAN AN (M'17) was born in 1983. He received the Ph.D. degree in computer science from Xi'an Jiaotong University, Xi'an, China, in 2013, where he is currently a Senior Engineer with the Department of Computer Science and Technology and a Researcher with the Key Laboratory of Computer Network. He has authored or coauthored more than 30 articles for scientific books, journals, and conferences. For the last four years, he has been involved in research on the social networks, service computing, and the Internet of Things. He and his team has been the recipients of the Science and Technology Progress Award of Shanxi and a number of National Natural Science Foundation Awards.

XIAOLIN GUI received the B.Sc. degree in computer science from Xi'an Jiaotong University (XJTU), China, and the M.Sc. and Ph.D. degrees in computer science from XJTU, in 1993 and 2001, respectively. Since 1988, he has been an Active Researcher in network computing, network security, and wireless networks with XJTU, where he is currently a Professor. He has been the Director of the Key Laboratory of Computer Network, XJTU, since 2008. His recent research interests include secure computation of open network systems, including grid, P2P, and cloud; dynamic trust management theory; and development on community networks.

ZHENXING WANG was born in 1995. He received the B.S. degree in computer science from the Anhui University of Technology, Ma'anshan, China, in 2017. He is currently pursuing the master's degree with Xi'an Jiaotong University, Xi'an, China. His research interests include privacy protection technology and mobile crowdsensing.

WENDONG ZHANG received the B.S. degree in mathematics and the M.S. degree in computer science and Technology from Xinjiang University, China, in 1998 and 2005, respectively. He is currently pursuing the Ph.D. degree in computer science with Xi'an Jiaotong University. His research interests include the Internet of Things and mobile crowdsensing.

RUOWEI GUI received the B.S. degree in automation and the M.S. degree in computer science from Xi'an Jiaotong University (XJTU), China, in 2013 and 2016, respectively, where he is currently pursuing the Ph.D. degree in computer science. He has been a Researcher in network security, data transfer, and data handling, since 2012, and continues to pursue research by working at the TCL Research Center. His recent research interest includes cloud and network security.

JINGXIAN XU was born in 1977. He received the B.S. degree from Minnan Normal University, Zhangzhou, China, in 2001, and the M.S. degree from Northeast Electric Power University, Jilin, China, in 2007. He is currently a Lecturer with Quanzhou Normal University, Quanzhou, China. He has been researching big data and cloud computing.