Journal of Information Hiding and Multimedia Signal Processing ©2018 ISSN 2073-4212
Ubiquitous International Volume **9**, Number **5**, September **2018**

Implicit Privacy Protection in Spatio-temporal Data Distribution

Hui Xia[a,*], Weiji Yang[b,*]

[a]School of Software, Shenyang Normal University,
No.253 Huanghe North Street, HuangGu District, Shenyang, China

[b]School of Life Science, Zhejiang Chinese Medical University,
548 Binwen Road, HangZhou, China

[*]Corresponding author: freund_xia@126.com, yangweiji@163.com

Received March, 2018; revised May, 2018

Abstract. *With the arrival of big data era, large numbers of users' location information is implicitly collected.Although these implicitly collected spatial and temporal data play an important role in the scientific and social fields,such as disease transmission and route recommendation, they cause new personal privacy disclosure problem when cross-referencing in spatio-temporal data distribution in big data era.The existing location privacy protection mechanism can not cross-reference the implicitly collected spatial data with the user's location data. In this paper, a nested-loop algorithm based on prefix filtering is proposed, which is based on the discovery-elimination of privacy protection. In particular, a nested-loop algorithm based on prefix filtering is proposed, Which is used to discover the records of implicitly collected spatio-temporal data that may reveal user privacy, and proposes a false data addition method based on frequent moving objects to eliminate these records.In addition, a more efficient anti-a priori algorithm and a graph-Data addition algorithm. Finally, the proposed algorithm is tested on several real data sets, and it is proved that these algorithms have high protection effect and performance.*

Keywords: Pspatio-temporal data, Big data, Privacy preserving, Implicit privacy protection

1. **Introduction.** In the era of big data, with the development of location technology, location-based services are becoming more and more popular, and users' space-time data is distributed through various services. While users actively publish their own spatio-temporal behavior through mobile social networking services such as sign-in, a large amount of spatio-temporal data recording people's behavior is implicitly collected[1] by mobile operators when people use mobile phones to make calls and receive short messages. Since the spatio-temporal data between mobile phones and mobile communication carriers is automatically collected by the signal base stations of mobile phones, these implicitly collected data have features of large data volume, implication of human behavior, and transmission of disease[2,3], poverty elimination [4], major social issues such as urban planning[5] and other important life applications play an key role in our lives,such as route recommendations[6] and travel by car[7]. However, these implicitly collected spatio-temporal data are cross-referenced with spatio-temporal data voluntarily released by users, will expose users' sensitive private information on personal identity, purpose of action, health status,hobbies,etc.[8,9].In recent years, with the increase of personal privacy and data dissemination in the use of laws and regulations. These implicitly collected

spatiotemporal data need to first eliminate records that may expose user privacy prior to scientific research and data mining.

In order to ensure that these sensitive personal information is not leaked, a large amount of work on privacy protection for spatio-temporal data is devoted to the anonymization of spatio-temporal data that may expose personal sensitive information. Some method,such as k-anonymity on location and trajectory data, will generalize users' data of the specific time range and the location record into the same spatial region, so that the attacker can not identify a specific user in a certain time range and space region. However, these methods do not consider that the attacker can actively publish with reference to the user. Spatio-temporal data finds records that reveal the privacy of users from spatio-temporal data that is implicitly collected. Therefore, spatio-temporal data sets protected by these methods will still leak data on user privacy.

The latest research[10] shows that the existing location or trajectory privacy protection method even if the location to be published is generalized to 15 square kilometers, timestamp generalization for one hour, more than 95% of users can also be four generalization of the time and space records are uniquely determined. If these users through the check-in service automatically released a number of their own behavior, they can easily be identified by the attacker.

Since attackers can collect spatio-temporal data voluntarily released by users through more and more social networking sites at the same time, the spatio-temporal data sets that are implicitly collected are more likely to expose user privacy. This protects these spatiotemporal records in effect. This poses serious challenges in protecting these time and space records in terms of effectiveness, efficiency and utility:(1) How the protection of spatio-temporal data sets is not diminished by the increased ability of attackers to collect spatio-temporal data that users actively publish; (2) Since the number of records in a data set that may reveal user privacy increases exponentially with the increase of spatiotemporal data sets, how to efficiently discover records to be protected; (3) How to ensure higher data utility when the space-time point set is protected for the release of spatio-temporal data.

In order to deal with these challenges, we first define implicit privacy on the implicitly collected datasets to ensure that no matter how many users actively collect the data, the implicitly collected spatial and temporal datasets. We will find two steps to protect the implicitly collected spatial and temporal data. In order to find out the records which may expose the user's privacy, this paper proposes a nested loop algorithm based on prefix filtering. An efficient antecedent algorithm is proposed to avoid the expose of the privacy of the user, so we design the false data adding algorithm based on the frequently moving objects, and propose a new algorithm based on the graph of the high degree of distortion of the data to ensure that the protected data set has high data utility.

In section 1, we introduce the technology of privacy protection in spatio-temporal data distribution. Chapter 2 introduces the implicit privacy problem and its discovery-elimination protection framework. The third section introduces the corresponding discovery and elimination algorithm. Chapter 4 is the experimental result display.

2. **Related work.** Anonymous methods[11,12] have been widely used to protect sensitive information in spatio-temporal data distribution and have achieved great success. Literature[13] gives a detailed overview. Anonymous methods can be classified into position k anonymity and trajectory k anonymity.

Location k anonymously generalizes each spatio-temporal data to be published from the two dimensions of location and time so that each generalized spatiotemporal region contains at least k moving objects[14]. Some of its variants use spatial indexing to organize individual moving objects to improve generalization and query performance[15]. Other variants make generalization more successful by considering the maximum speed limit for each spatiotemporal data to move between different generalization areas[16]. However, location k anonymous cannot solve the identity leakage caused by implicit location information collection and active location release in the era of big data.

The trajectory k anonymity[17,18] attempts to protect the entire trajectory anonymously, and make each trajectory satisfy k-anonymity. However, due to the diversity of moving object trajectories, it is impossible to completely anonymous to all trajectories[17,19,20]. Therefore, the existing method only slices the trajectory for different time periods, and performing k-anonymization on the trajectory after the slicing.

The literature[21] proposes an efficient solution to the problem of privacy leakage caused by cross-referencing between two related data sets (such as a consumer's location data and shopping data in a store). But the method is against the background of the attacker. Knowledge is limited and cannot be applied to the era of big data with multiple sources of public data. At the same time, differential privacy has attracted the attention of researchers since it can protect against any background knowledge in recent years[22,23], but the differential privacy method can only be protected based on the statistical value of the data and does not apply to us to publish the entire spatiotemporal dataset for analysis. Other studies based on cloud computing or crowdsourcing have protected privacy in accurate data distribution by distributing data to attackers who do not communicate with each other[24]. However, these methods often use integer programming to seek balance in publishing data and privacy protection,which actually exposed location privacy.

3. **Problem definition and privacy protection framework.**

3.1. **Problem definition.**

When a user u at the time t in the location l use mobile phone to send and receive text messages, then call, shaped as triples $< u, l, t >$ space-time information implicitly recorded by mobile operators, which is called spatio-temporal dataset D. The two-tuple $< l, t >$ which represents the specific location and time of the user, is called the space-time point. Through aggregating the users appearing at the same time and space, the given location loc and time $time$, space-time point $< loc, time >$ can be associated with a set of users $Sp_{loc,time} = \{u \mid < u, l, t > \in l = loc, t = time\}$ in this space-time point co-occurrence.

Definition 1 (Valid time-space point). Given the implicitly collected spatiotemporal dataset D, and the temporal and spatial thresholds ε_1 and ε_2 , $\forall$ spatiotemporal point $< l_i, t_i > \in D$, if $\exists D' \subseteq D$, make $\forall < l_j, t_j > \in D'$, for $|t_i - t_j| < \varepsilon_1$ and $|t_i - t_j| < \varepsilon_2$, then we call $< l_i, t_i > \cup \bigcup\limits_{<l_j,t_j>\in D'} < l_j, t_j >$ effective spatiotemporal point.

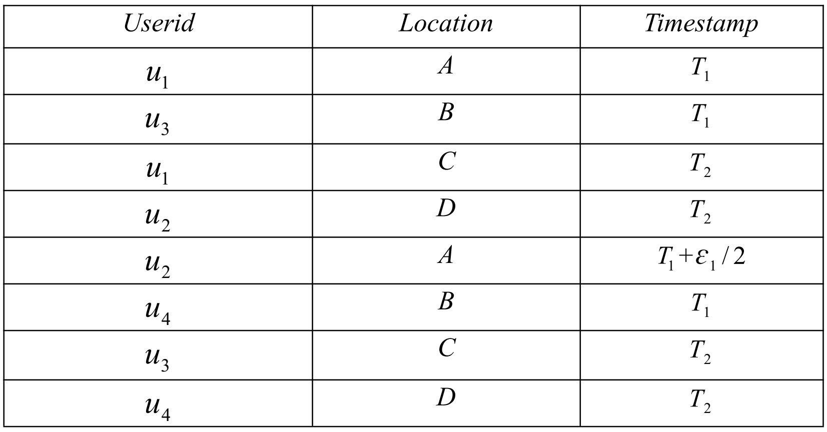
According to definition 1, table 1 has 6 effective spatio-temporal points are: $Sp_{A,T_1} = \{u_1\}, Sp_{A(T_1+\varepsilon_1/2)} = \{u_1\}$, $Sp_{B,T_1} = \{u_3, u_4\}, Sp_{C,T_2} = \{u_1, u_3\}, Sp_{D,T_2} = \{u_2, u_4\}$,

$Sp_{merge} = \{u_1, u_2\}$, wherein, the first five spatiotemporal points become valid spatiotemporal points because of their own existence, the sixth valid space-time point is formed by merging Sp_{A,T_1} and $Sp_{A,(T_1+\varepsilon_1/2)}$. For convenience, we write the time threshold and the distance threshold as $\varepsilon = \{\varepsilon_1, \varepsilon_2\}$.

If a user u through the sign-in service, etc. take the initiative to disclose their own in a number of spatiotemporal points set $\{< l_1, t_1 >, < l_2, t_2 >, ..., < l_n, t_n >\}(n \geq 1)$. While in the implicitly collected spatial-temporal data set D, the set $Sp_1, Sp_2, ...Sp_n$ of these spatiotemporal points satisfies $Sp_1 \cap Sp_2 \cap, ... \cap Sp_n = \{u\}$, then the attacker can associate the user who exposes his own location with the user u in the implicitly collected spatiotemporal data set. The attack process requires the attacker to collect enough public time and space points for the user, and we measure the attack with k. The ability to collect the number of spatiotemporal points exposed by the same user.

We present the (ε, k) privacy problem in the spatio-temporal data implicit collection based on the process of associating the user - published spatio-temporal points with the spatio-temporal points in the implicitly collected spatio-temporal data set.

TABLE 1. Example for valid spatiotemporal point ($T_2 > T_1 + 2\varepsilon_1$)



Userid	*Location*	*Timestamp*
u_1	A	T_1
u_3	B	T_1
u_1	C	T_2
u_2	D	T_2
u_2	A	$T_1+\varepsilon_1/2$
u_4	B	T_1
u_3	C	T_2
u_4	D	T_2

Definition 2. ((ε, k)Implicit privacy). Given the implicitly collected spatio-temporal data set D, the time threshold and the distance threshold $\varepsilon = \{\varepsilon_1, \varepsilon_2\}$ and the attacker's attack ability $k, \exists l \leq i \leq k$, effective spatio-temporal points to make $|Sp_1 \cap Sp_2 \cap, ... \cap Sp_n| = 1$, we say $\{Sp_1, Sp_2, ...Sp_n\}$ exposes (ε, k) implicit privacy.

3.2. **Privacy protection framework.** First of all, we allow the user to set the parameters of the privacy problem (ε, k) according to the different characteristics of the dataset. Specifically, when the public dataset is the microblogging data with poor time-space, the time threshold. The threshold value can be larger than the time-sensitive sign-in data; when the user (group) often publish their own time and space points, the attacker's attack ability k can be set larger.

It is worth noting that, in order to protect (ε, k) privacy, minimizing the spatiotemporal data that needs to be changed is an NP-hard problem.

Theorem 1. Minimizing the amount of data change to protect (ε, k) privacy problem is an NP-hard problem.

Proof. Since the minimization of data changes to protect the k-anonymity problem is an NP-hard problem, we need to prove that the k-anonymity problem can be reduced to (ε, k) in polynomial time. In the privacy protection problem, we set ε to $\{0, 0\}$. Thus, the solution of any k-anonymous problem is the solution of the privacy protection problem for $\varepsilon(k, \varepsilon)$ after $\{0, 0\}$. The solution of the privacy protection problem is also the solution of the k-anonymous problem. This mapping can be done in polynomial time. From this we can conclude that the protection of (ε, k) privacy problem is NP-hard problem.

Based on theorem 1, in order to efficiently protect (ε, k) privacy, we use the discovery-elimination framework. As shown in Fig. 1, the privacy protection framework is divided into three modules: data preprocessing module, finding violation of the privacy module, eliminating the privacy leak module.

According to the definition of implicit privacy, we first construct all spatio-temporal points according to time threshold and distance threshold ε in the data preprocessing module. Because we can not determine which time and space points are released by public users in advance, in large data era attackers can usually collect the same user-time-space

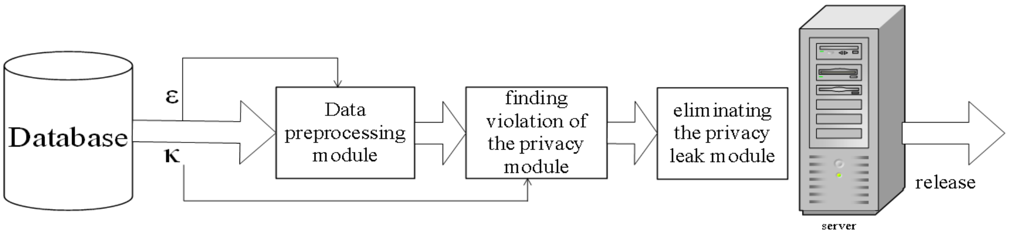

Database
ε
κ
Data preprocessing module
finding violation of the privacy module
eliminating the privacy leak module
server
release

Figure 1. Implicit privacy preservation framework

data from multiple social networking sources. In order to ensure that the user's privacy is not disclosed, we find any time-space combinations of $i(1 \leq i \leq k)$ spatiotemporal points in the discovery of violation of the privacy module, and check whether the attacker can match a user in the implicitly collected spatiotemporal data.Secondly, eliminate the privacy leak module will protect the set of spatiotemporal points discovered, eliminate the implicitly collected spatiotemporal data set pair (ε, k), in this way,we can find and eliminate spatiotemporal data which lead to (ε, k) privacy leak, regardless of the time-space point of the user's active publication.

4. **Algorithm and Analysis.** This section introduces algorithms for discovering and eliminating (ε, k) privacy disclosure.

4.1. **Discover (ε, k) privacy leaks.** Based on the privacy protection parameters ε and k, the basic idea of discovering spatiotemporal point data of privacy (ε, k) privacy is to enumerate all k spatiotemporal points and check whether each combination is associated with a unique user. The prefix filter based nest loop (*PF-NL*) is implemented by enumerating the k-bit non-repeating numbers to realize the nested loop of the k-points of the spatiotemporal points, and using the prefix filtering method in the enumeration. We first introduce the important properties of (ε, k) privacy.

Properties 1. Let U_k be the set of all space-time points that expose (ε, k) privacy, and let U_k be the set of all space-time points of size k that can be uniquely associated with the moving object, then $U_k = u_1 \cup \cdots \cup u_k$.

We skim the trivial proof of property 1, which states that we can start with a set of spatiotemporal points of size 1 until we enumerate the set of spatiotemporal points whose size does not exceed k. For this reason, we assign n space-time points from 1, and each set of time-space-points of size k can be regarded as a number with k-ary for n, and each of the same set of time-and-space points has k unique numbers representing[25] the time-space points in order.

Thus, a set of spatiotemporal points of size $k(k > 1)$ has a prefix of length $\{1, ..., k-1\}$. For example, we can denote the six valid spatiotemporal points are: $Sp_{A,T_1}, Sp_{A,(T_1+\varepsilon/2)}$ Sp_{B,T_1} , Sp_{C,T_2} , Sp_{D,T_2} and Sp_{merge} in Table 1 by the numbers 1 to 6, respectively. Considering $(\varepsilon = 0, k = 3)$ privacy, for a set of space-time points of size $\{1, 2, 3\}$ of 3, it has the prefix $\{1, 2\}$. We know $|Sp_1 \cap Sp_2| = 0$,therefore, the set of spatio-temporal points with this prefix must not expose (0,3) privacy. In the method based on prefix filtering, we avoid enumerating the set of spatio-temporal points that contain this prefix.

Combined with the above basic idea and the optimization method based on prefix filtering, we propose a basic algorithm $PF-NL$ which is found to violate implicit privacy.

Algorithm 1. $PF - NL$ Algorithm

Input: A set of spatiotemporal points of size n, privacy requirement (ε, k).

Output: All spatiotemporal points set R of violation of (ε, k) privacy.

```
num=[0,...,k-1],bound[n-k+1,...n],R=∅
**while** true **do**
**if** violate(num)==1 **then**
R.add(num)
**else**
prefix=$\min_p violate(num[1...p]) = 0$
**endif**
$(p = \max_{p<prefix} num[p] < bound[p])$
**if** $p < 0$
**break**
**endif**
num[p]=num[p]+1
num[p+1,...,k]=[num[p]+1,...num[p]+$k_p$+1]
**endwhile**
**return** R
```

In the following, we introduce an reverse aprior(RA) algorithm and use the idea of breadth-first search to avoid the existence of a small set of spatio-temporal points that do not violate (ε, k) privacy when searching for larger sets of points.

Algorithm 1 shows the reverse aprior (RA algorithm), where u_i, z_i and c_i denote the violations of size i, impossible to violate and spatio-temporal points set (ε, k) which may be violated privacy.

In the RA algorithm, we first check all the spatio-temporal points, and the violation of (ε, k) privacy point of time and space added to u_1; because the space-time point must contain moving objects, they are all added c_1 (line 1).

Next, we consider a set of more spatiotemporal points consisting of a smaller set of spatiotemporal points that are likely to violate (ε, k) privacy (line 5). Since the set of size $i+1$ has a variety of combinations, different combinations may lead to different size sets of time-space points to be tested, in order to reduce the time overhead, we choose the two sets of the smallest space-time point of the Cartesian product to generate it. The set of spatiotemporal points used to check whether the violation of (ε, k) privacy should not contain a smaller set of spatiotemporal points that are unlikely to violate (ε, k) privacy (line 6). Finally, the violate sets the intersection of the set of moving objects contained in each space-time point in the set of space-time points and returns its size. The set of space-time points whose size is 1 violates (ε, k) privacy, is added u_{i+1} (row 6), and a set of spatiotemporal points with an intersection size of 0 would not be able to violate the (ε, k) privacy, thus adding z_{i+1} and removing it from c_{i+1} (lines 8, 9).

Algorithm 2. RA Algorithm
Input: Privacy requirements
Output: A set of spatiotemporal points that violate privacy requirements.

```
1. Scan database and form $u_1$ and $c_1$
2. $i = 1, R = \emptyset$
3. While ($c_i \neq \emptyset$ and $i \leq k$)
4. $w = \arg\min_w |c_w| \times |c_{i-w+1}|$
5. $c_{i+1} = c_w \times c_{i-w+1}$
6. $c_{i+1} = \{e \in c_{i+1} \,|\forall f \subseteq e, f \notin z_i, f \notin z_{i-w+1}\}$
7. $u_{i+1} = \{e \,|e \in c_{i+1}, violate(e) = 1\}$
8. $z_{i+1} = \{e \,|e \in c_{i+1}, violate(e) = 0\}$
9. $c_{i+1} = c_{i+1} - z_{i+1}$
```

```
Endwhile
$R = u_1 \cup ... \cup u_k$
```

4.2. **Eliminate spatiotemporal data that violates (ε, k) privacy.** In this section, we introduce the method of adding false data to eliminate spatiotemporal data that violates the (ε, k) privacy. In particular, we introduce a method of adding false data based on frequently moving objects that do not need to find a set of spatiotemporal points that violate (ε, k) privacy, and a privacy leak elimination method that can achieve higher data utility.

1) A false data add method based on frequent Moving objects

Algorithm 2 (FMO algorithm) shows the process of adding fake data based on frequent mobile users. We first find the two most frequently occurring moving objects, given the unique privacy parameter (ε, k), this process has a large number of fast algorithms [26]; secondly, we add these moving objects in each space-time point set.

Algorithm 3. FMO Algorithm

Input: Privacy requirements (ε, k), time-space points set D

Output: Publish data P satisfying (ε, k) privacy constraint

```
$P = \emptyset$
$F = \{\{u_1, u_2\} \,|NOT\exists\{u_3, ..., u_n\}, count(u) > \min\{count(u_1), count(u_2)\}\}$
For each $< u, l, t > \in D$
$P.add(< u.l.t >)$
$P.add(< u_1.l.t >)$
$P.add(< u_2.l.t >)$
Endfor
```

2) Graph - Based False Data Add Method

Algorithm 3 shows the process of graph-based dummy filling (G-DF) based on the set of spatiotemporal points of violation of (ε, k) privacy found in section A. In the algorithm G-DF, we consider position data as a graph, where each point represents a node in the graph, and if two points in time and space are exposed to a set of unique privacy, just add an edge to the diagram (line 2). We run Algorithm 2 only for each connected component in the graph, ie find the most frequent two moving objects in each connected component and add them to each node in the connected component (lines 3 to row 5).

Algorithm 4. G-DF Algorithm

Input: A Set of spatiotemporal points for exposure to uniqueness privacy

Output: Publish data P satisfying (ε, k) privacy constraint

```
1. $P = \emptyset$
2. Transform D and R into Graph g
3. for each connectivity sging
4. Perform Algorithm 3
5. Endfor
```

5. **Experiment.** We use two real data sets to compare the performance and effectiveness of the algorithm in the discovery-elimination framework to detect spatiotemporal data that violate (ε, k) implicit privacy and two algorithms to eliminate this violation. We will experiment to answer the following questions.

(1) How does the privacy protection parameter (ε, k) affect (ε, k) privacy in spatio-temporal data

(2)The influence of privacy protection parameters (ε, k) on the performance of privacy

(3)Effect of privacy protection parameters (ε, k) on the effect and performance of privacy protection algorithm

5.1. **Experimental Environment and Data Set Description.** We use Java 1.7 to implement Algorithm 1 $\sim$ Algorithm 3 in this paper, the experiment environment is a Linux server, Intel Xeon E5645 2.4Ghz processor, 128G RAM, 1T SATA hard disk. In addition to the RA method used in this paper to find the violation (ε, k) privacy space-time point set and the FMO and G-DF methods used in this paper are used to eliminate the set of spatio-temporal points that violate (ε, k) privacy. Other methods include:

(1) YCWA[18]: This method is the latest method to protect the privacy of spatiotemporal data by trajectory anonymity technology, which divides trajectories into dwell points and protects privacy information by anonymizing these dots. The method mainly focuses on the performance of trajectory privacy protection.

(2) SEQ-ANON[19]: This method focuses on the data availability of trajectory anonymity techniques, while minimizing the distance between the changed temporal and spatial points and the original temporal and spatial points while anonymizing the trajectory.

We experimented with two public data sets, GeoSocial[27] and GeoLife[28], as a user-implied collection of spatiotemporal data sets, with data size and number of users in Table 2.

Table 2. Dataset

DataSet	*Number of records*	*Number of moving objects*
GeoSocial	*4M*	18k
GeoLife	*20M*	17K

5.2. **Privacy protection effect comparison.** Figure 2 shows the effect of spatial data privacy protection on GeoSocial and GeoLife datasets under the more stringent (ε, k) privacy conditions (ε_1=10min, ε_2=1km, k=10). Our method RA-G-DF uses the best-performing RA method in the discovery phase and the best-performing G-DF method in the elimination phase. We can see that the trajectories still have a large number of (ε, k) violations privacy of spatio-temporal points set after YCWA and SQL-ANON processing, this is because these methods do not take into account cross-reference between the implicit collection of spatiotemporal data and user-published spatio-temporal data, which will lead to privacy disclosure.

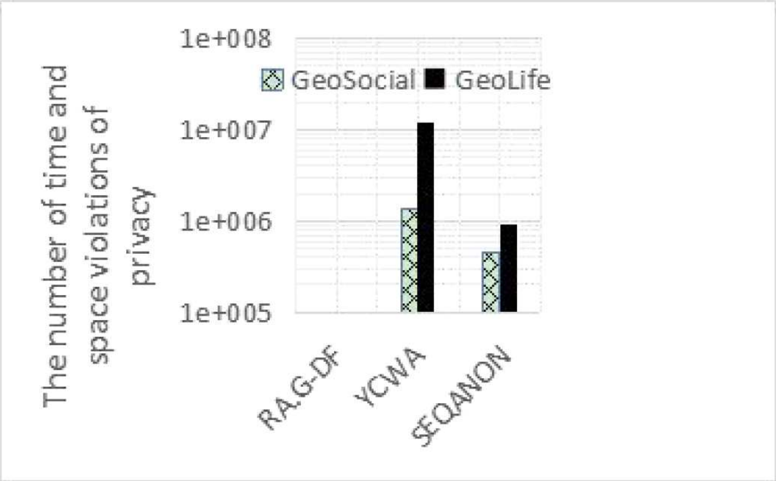


Figure 2. Implicit privacy preservation framework privacy preservation

5.3. (ε, k) **privace.** As shown in Fig. 3, when ε_2 is fixed and ε_1 is adjusted, the proportion of the spatio-temporal points set of (ε, k) privacy in the three data sets increases with increasing k for the three data sets. This phenomenon provides a basis for the high efficiency of RA algorithm.3(b) and 3(d) show that the proportion of the set of spatiotemporal points exposed to (ε, k) privacy on two datasets increases gradually as ε_2 increases, which is because more efficient spatio-temporal points are easier to match with the public positions of a user, thus violating (ε, k) privacy.

5.4. **Found to violate** (ε, k) **privacy algorithm performance.** We compare and verify the efficiency of our two (ε, k) implicit privacy discovery algorithms with real data sets GeoLife and GeoSocial.

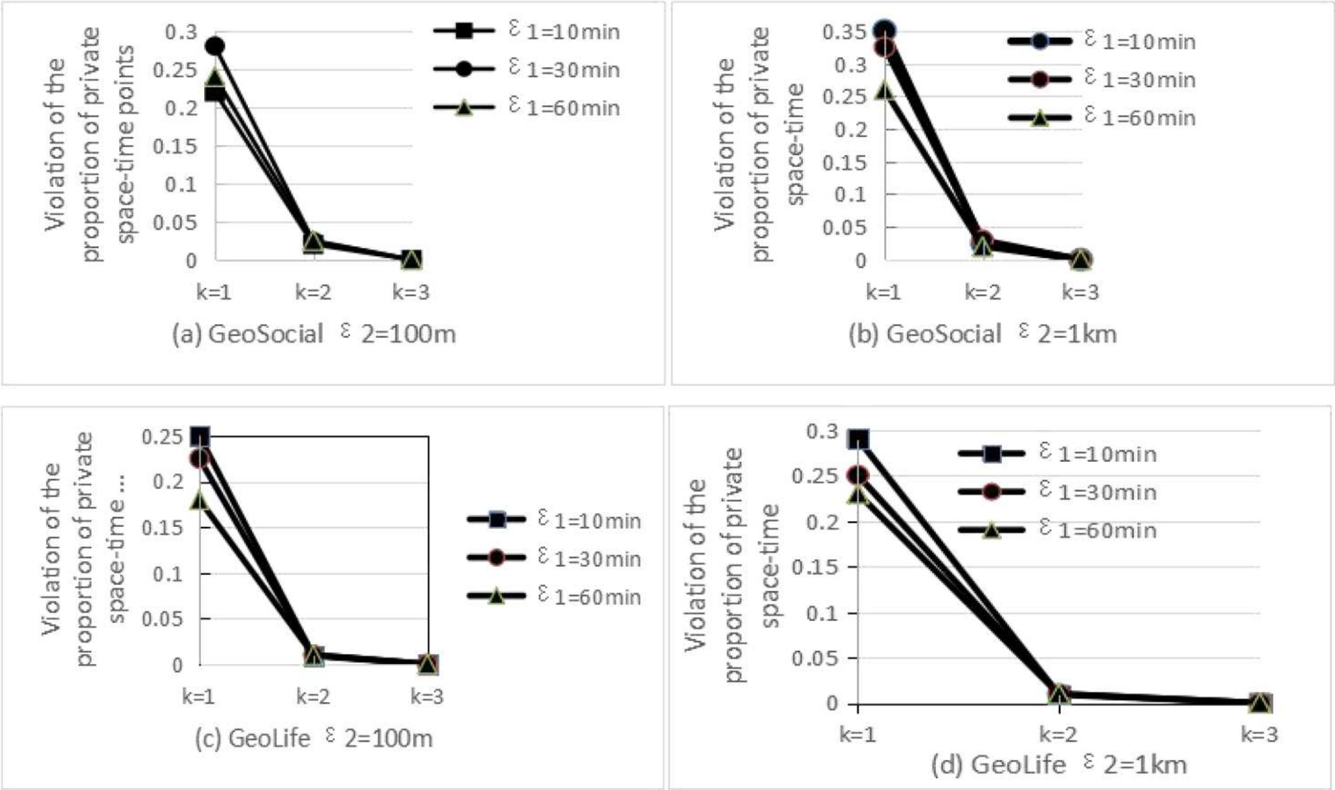
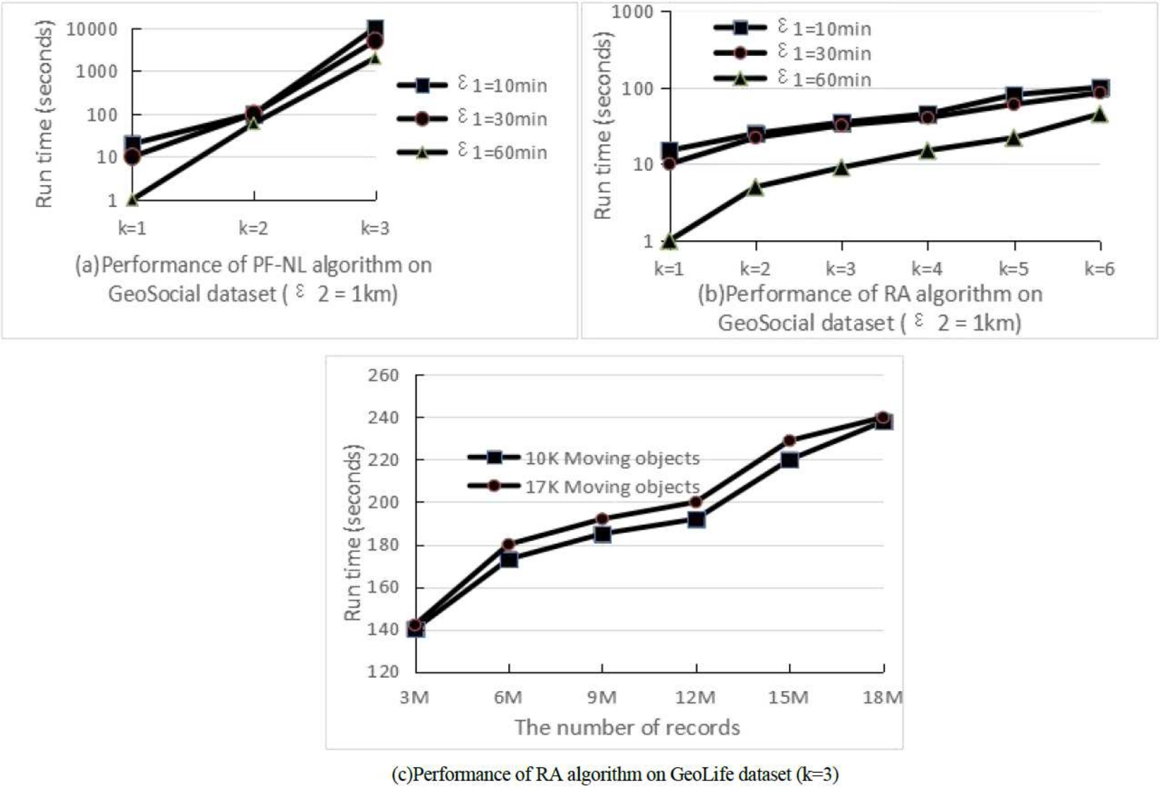


Figure 3. Effect of ε and k on three datasets

From the comparison of Fig.4(a) and Fig.4(b), it can be seen that under the same privacy parameter (ε, k), RA algorithm is $1 \sim 2$ orders of magnitude faster than PF-NL algorithm under the same privacy parameter. In particular, the PF-NL algorithm can not even compute the case where $k > 3$.Figure 4 shows two algorithms in the GeoSocial dataset to find out all the run-time violations of (ε, k) privacy under different (ε, k)

For the RA algorithm, we adjust the GeoLife dataset size and the number of moving objects to test its performance, Figure 4(c) shows that the size of the GeoLife dataset has the greatest impact on the algorithm because it changes the number of spatio-temporal points in the data set, while the number of moving objects has little effect on the algorithm.

5.5. (ε, k) **privacy protection algorithm effec.** We are constantly adjusting the privacy parameters n and k, to see two privacy protection algorithm for data sets released after the privacy protection of data utility.As shown in Fig. 5(a) and Fig. 5(b), the FMO algorithm adds more than 80% of the fake data to the two datasets, while the G-DF

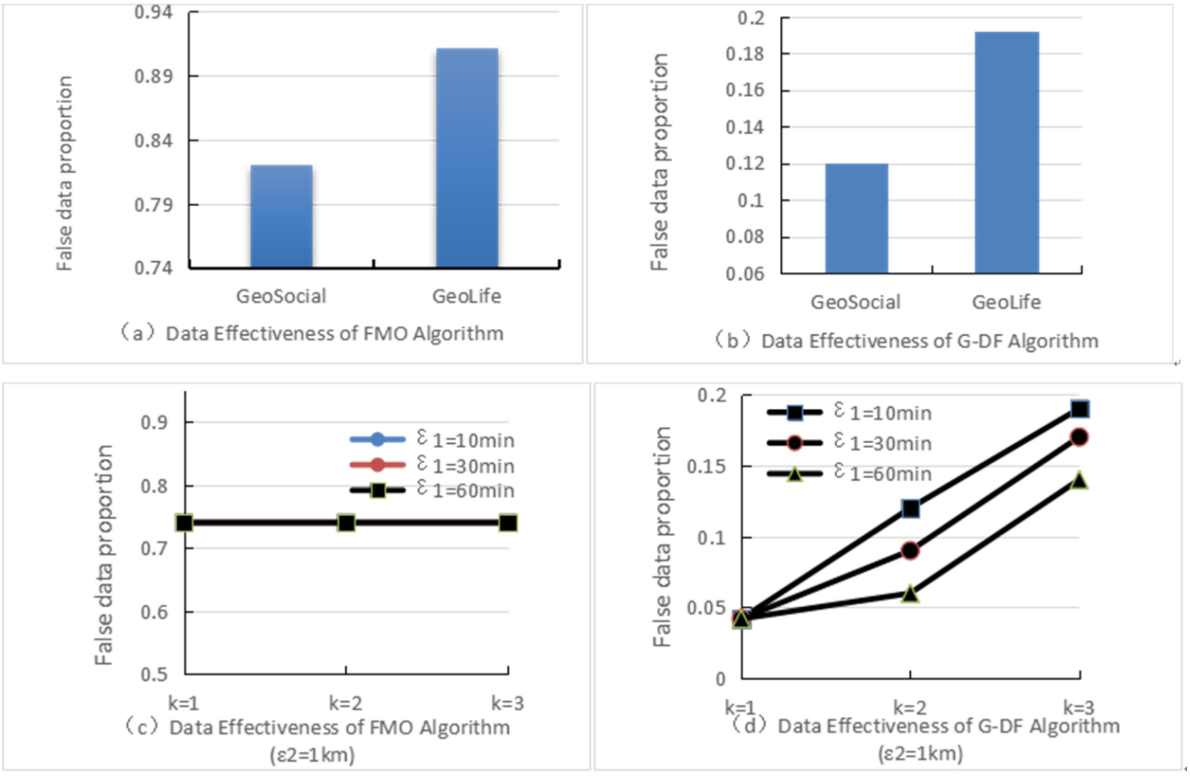


(a)Performance of PF-NL algorithm on GeoSocial dataset (ε 2 = 1km)

(b)Performance of RA algorithm on GeoSocial dataset (ε 2 = 1km)

(c)Performance of RA algorithm on GeoLife dataset (k=3)

FIGURE 4. Performance comparison of algorithm PF-NL and RA



（a） Data Effectiveness of FMO Algorithm

（b） Data Effectiveness of G-DF Algorithm

（c） Data Effectiveness of FMO Algorithm (ε2=1km)

（d） Data Effectiveness of G-DF Algorithm (ε2=1km)

FIGURE 5. Performance comparison of algorithm of FMO and G-DF

algorithm can only add 10 to 20% false data. Fig. 5(c) shows that the change of privacy parameters does not affect the data utility of the FMO algorithm, because the FMO algorithm adds only two dummy data for each spatio-temporal point. Fig. 5(d) illustrates that the amount of dummy data to be added becomes large when decreases for the G-DF algorithm. This is because a set of spatio-temporal points of violation of privacy (ε, k) is increased when ε_1 decreases as shown in Figs. 3(a) and 3(c). We can also see that the data utility of the G-DF algorithm is much better than the FMO algorithm. This is because the G-DF algorithm does not add unnecessary false data.

5.6. (ε, k) **privacy protection algorithm performance.** We use the largest real data set GeoLife to compare the operating efficiency of two (ε, k) privacy protection algorithms.

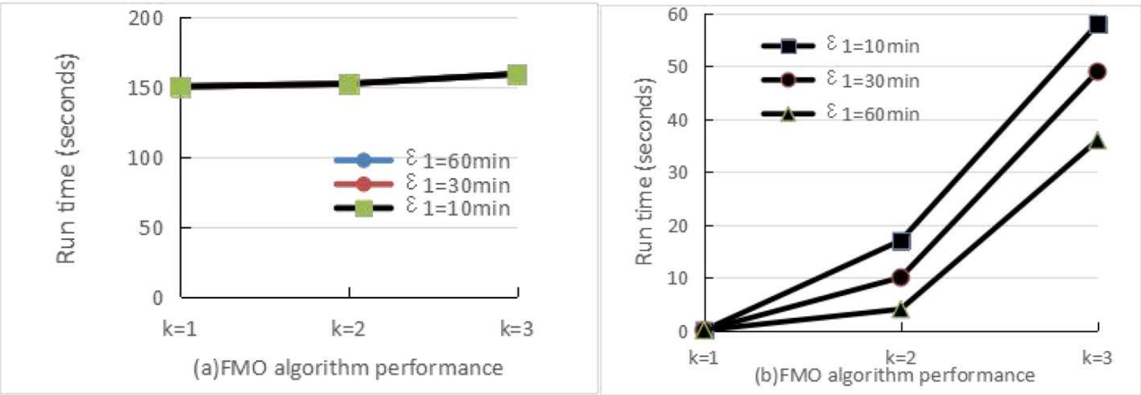


Figure 6. Performance comparison of algorithm of FMO and G-DF under GeoLife dataset (ε 2 =1km)

It can be seen that in Fig. 6, which violates the (ε, k) privacy spatio-temporal point set proportion, so its processing time is positively correlated with the number of sets of spatio-temporal points that violate privacy under the (ε, k) privacy parameter (ε, k).

6. **Conclusions.** In this paper, the definition of implicit privacy (ε, k) in the spatio-temporal data is proposed for the case that the user-published dataset is cross-referenced with the implicitly collected spatiotemporal data sets, and proposed the discovery-elimination privacy protection framework. In particular, two efficient algorithms are proposed for finding the set of spatio-temporal points that violate (ε, k) privacy. In addition, this paper proposes an anonymity protection method with false data addition. In order to improve the data utility, this paper proposes a new method of false data addition based on graph. Experiments on the real data sets show that the proposed algorithm is efficient. We will improve the performance of our method in the future work.

REFERENCES

[1] E. Nathan,P. Alex, Reality mining: Sensing complex social systems. *Journal of Personal and Ubiquitous Computing*, vol. 10, no. 4, pp. 255-268, 2006 [doi: 10.1007/s00779-005-0046-3]

[2] M. A. Le, A. J. Tatem, J. M. Cohen, Hay SI, Randell H, Patil AP, Smith DL. Travel risk, malaria importation and malaria transmission in Zanzibar. *Scientific Reports*, vol. 1, no. 7364, pp. 271-275. 2011 [doi:10.1038/srep00093]

[3] A. Wesolowski, Eagle N, Tatem AJ, Smith DL, Noor AM, Snow RW, Buckee CO. Quantifying the impact of human mobility on malaria. *J. Science*, vol. 338, no. 6104, pp. 267-270.2012 [doi: 10.1126/science.1223467]

[4] S. Hill, Banser A, Berhan G, Eagle N. Reality mining Africa. *In: Proc. of the AAAI Spring Symp. on Artificial Intelligence for Development.* 2010. http://www.seas.upenn.edu/ ngns/docs/References/Hill%202010%20realityminingafrica.pdf

[5] J. Yuan, Y. Zheng, X. Xie, Discovering regions of different functions in a city using human mobility and POIs. In: Yang Q, Agarwal D, Pei J, eds. *Proc. of the KDD. New York: ACM Press,* pp. 186-194. 2012. [doi: 10.1145/2339530.2339561]

[6] K. Zheng, Shang S, Yuan J, Yang Y. Towards efficient search for activity trajectories. In: Jensen CS, Jermaine CM, Zhou XF, eds. Proc. of the ICDE. *Washington: IEEE Computer Society*, vol. 230-241. 2013.[doi: 10.1109/ICDE.2013.6544828]

[7] N. J. Yuan NJ, Zheng Y, Zhang L, Xie X. T-Finder: A recommender system for finding passengers and vacant taxis. *IEEE Trans. on Knowledge & Data Engineering,* vol. 25, no. 10, pp. 2390-2403.2013 [doi: 10.1109/TKDE.2012.153]

[8] S. B. Wicker The loss of location privacy in the cellular age. *Communications of the ACM,* vol. 55, no. 8, pp. 60-68. 2012[doi: 10.1145/22402 36.2240255]

[9] L. Wang, Meng XF, Information SO. Location privacy preservation in big data era: A survey. Ruan Jian Xue Bao/*Journal of Software,* 2014, vol. 25, no. 4, pp. 693-712 , no. in Chinese with English abstract). http://www.jos.org.cn/1000-9825/4551.htm [doi: 10.13328/j.cnki. jos.004551]

[10] Montjoye YAD, C. A. Hidalgo, M. Verleysen, V. D. Blondel . Unique in the Crowd: The privacy bounds of human mobility. Open Access Publications from Université Catholique De Louvain, 2013, vol. 3, no. 6, pp. 776-776.

[11] G. G. Bu, L. Liu, A customizable k-anonymity model for protecting location privacy. *In: Proc. of the Icdcs.* 2004. pp. 620-629.

[12] A. E. Cicek, M. E. Nergiz, Y. Saygin, Ensuring location diversity in privacy-preserving spatio-temporal data publishing. *VLDB Endowment,* 2014, vol. 23, no. 4, pp. 609-625. [doi: 10.1007/s00778-013-0342-x]

[13] B. C. M. Fung , Wang K, Chen R, Yu PS. Privacy-Preserving data publishing: A survey of recent developments. ACM Computing Surveys, 2010, vol. 42, no. 4, pp. 2623-2627. [doi: 10.1145/1749603.1749605]

[14] Sweeney L. K-Anonymity: A model for protecting privacy. *Int'l Journal of Uncertainty, Fuzziness and Knowledge-Based Systems,* 2008, vol. 10, no. 5, pp. 557-570. [doi: 10.1142/S0218488502001648]

[15] M. F. Mokbel, Chow CY, Aref WG. The new casper: Query processing for location services without compromising privacy. In: Dayal U, Whang KY, et al., eds. Proc. of the VLDB. New York: ACM Press, 2009. pp. 763-774.

[16] X. Pan, Xu J, Meng X. Protecting location privacy against location-dependent attack in mobile services. *IEEE Trans. on Knowledge & Data Engineering,* 2011, vol. 24, no. 8, pp. 1506-1519. [doi: 10.1109/TKDE.2011.105]

[17] O. Abul, Bonchi F, Nanni M. Never walk alone: Uncertainty for anonymity in moving objects databases. In: Alonso G, Blakeley J, Chen ALP, eds. *Proc. of the ICDE. Washington: IEEE Computer Society*, 2008. pp. 376-385. [doi: 10.1109/ICDE.2008.4497446]

[18] Z. Huo, Meng X, Hu H, Huang Y. You can walk alone: Trajectory privacy-preserving through significant stays protection. In: Lee SG, Peng ZY, et al., eds. *Proc. of the DASFAA. Berlin: Springer-Verlag,* 2012. 351-366. [doi: 10.1007/978-3-642-29038-1_26]

[19] G. Poulis, Skiadopoulos S, Loukides G, Gkoulalas-Divanis A. Apriori-Based algorithms for k m -anonymizing trajectory data. *Trans. on Data Privacy*, 2014, vol. 7, no. 2, pp. 165-194.

[20] J. Domingo-Ferrer, Trujillo-Rasua R. Microaggregation- and permutation-based anonymization of movement data. *Information Sciences,* 2012, vol. 208, no. 21, pp. 55-80. [doi: 10.1016/j.ins.2012.04.015]

[21] H. Hu, Xu J, On ST, Ng JKY. Privacy-Aware location data publishing. *ACM Trans. on Database Systems,* 2010, vol. 35, no. 3, pp. 53-56. [doi: 10. 1145/1806907.1806910]

[22] C. Dwork, Differential privacy. In: Bugliesi M, Preneel B, et al., eds. *Proc. of the ICALP. Berlin: Springer-Verlag,* 2006. pp. 1-12. [doi: 10.1007/11787006_1]

[23] M. Hay, V. Rastogi, Miklau G, Suciu D. Boosting the accuracy of differentially private histograms through consistency. VLDB Endowment, 2009,vol. 3, no. 1, pp. 66-69. [doi: 10.14778/1920841.1920970]

[24] T. Rekatsinas, Deshpande A, Machanavajjhala A. SPARSI: Partitioning sensitive data amongst multiple adversaries. *VLDB Endowment,* 2013, vol. 6, no. 13, pp. 1594-1605. [doi: 10.14778/2536258.2536270]

[25] D. Knuth, The art of computer programming. vol.4, Fascicle 2: Generating all Tuples & Permutations. Addison-Wesley Professional, 2008.

[26] A. Metwally, Agrawal D, El Abbadi A. Efficient computation of frequent and top-k elements in data streams. In: Eiter T, Libkin L, eds. *Proc. of the ICDT 2005. Berlin: Springer-Verlag,* 2005. pp. 398-412. [doi: 10.1007/978-3-540-30570-5_27]

[27] H. Geo, J. Tang, Liu H. Addressing the cold-start problem in location recommendation using geo-social correlations. D*ata Mining & Knowledge Discovery*, 2015, vol. 29, no. 2, pp. 299-323. [doi: 10.1007/s10618-014-0343-4]

[28] Y. Zheng, X. Xie, W. Y. Ma, GeoLife: A collaborative social networking service among user, location and trajectory. *Bulletin of the Technical Committee on Data Engineering,* 2010, vol.33, no. 2, pp. 32-39.