MedCrave
Step into the World of Research

International Robotics & Automation Journal

Research Article

Open Access

CrossMark

Spatio-temporal trajectory privacy-protection algorithm based on amending prefix tree

Abstract

The purpose of this paper is to prevent privacy leakage resulting from improper release of spatio-temporal trajectory data and achieve a balance between privacy protection and practicability. Therefore, this article builds a model of the self-evolution of attack patterns and proposes APT-PP algorithm based on the spatio-temporal correlation within trajectory data. The core thoughts of the APT-PP algorithm are as following: firstly, classify trajectory data into hold points and moving points, build trajectory chart and calculate the sensitivity of each point, which reduces the size of data set and improves operating efficiency; then, transform the above-mentioned chart into prefix tree to evaluate to what extent trajectories are kept secret, and prefix tree also raises the algorithm's retrieval efficiency; finally, amend the prefix tree with grafting operation, protecting privacy while keeping data utility. At the end of this article, a comparative experiment is introduced to assess the performance of APT-PP algorithm with real data set. The results of experiment proved that this algorithm can protect users' privacy and on the same time provide high-quality spatio-temporal trajectory data and positive user experience.

Keywords: trajectory, privacy protection, similarity, sensitivity

Volume 2 Issue 2 - 2017

HE Ming,[1,2,4] LIU Fangxin[1,4] ZHOU Huan,[3] CHEN Qiuli1,[2,4] MIAO Zhuang[1], ZHOU BO[1]

[1]College of Command Information System, PLA Science and Technology University, China
[2]The 61th Research Institute of PLA, China
[3]Institute of Vocational Education, Tongji University, China
[4]Nanjing University of Information Science and Technology, China

Correspondence: LIU Fangxin, College of Command Information System, PLA Science and Technology University, China, Email liufangxin_lgdx@163.com

Received: March 27, 2017 | **Published:** April 28, 2017

Abbreviations: TS, trajectory sequences; TISS, temporal information similarity set; TD, temporal data

Introduction

With the popularity of location services and the equipment with location function, a large number of location information and the path of moving objects are collected; analysis and digging of the internal relations and rules for these data are conducive to the needs of multiple applications,[1–2] for example: analysis and digging on the location information recorded by GPS of the city car can help the government to plan the urban road traffic, analysis and digging on the location information of the consumer group in business circle can provide a decision support for the site selection, advertisement putting and so on. Although more and more decision making related to location services benefit from the analysis and digging of spatio-temporal data, it is difficult to avoid the privacy threats caused by spatio-temporal data release: a lot of information about the privacy of the design user hides in the spatio-temporal data contained in the moving trajectories of the moving objects, and such hidden information, such as work habits, the work nature, work/residence address, financial status, etc., will be revealed in the digging of trajectory mode. In order to prevent the attacker from re-recognizing them in the released spatio-temporal data set based on the user's background information as well as part of data grasped, the privacy-protection algorithm is proposed.

Trajectory privacy protection has become the focus field of researchers in recent years. The attacker digs the moving mode of moving object in the in the released data set based on the user's background information as well as part of data stolen, so as to infer the corresponding user information of each trajectory. The protection of user privacy is not only to optimize the privacy level of the trajectory, but also to consider the availability of the optimized and released spatio-temporal data set, and the finally released data set will still be used for each location dependent application.[3] At present, some progress has been made in the research of trajectory privacy protection technology. Abul et al.[4] consider that the error exists in the location service provided by some devices, and propose the concept of (k,δ) anonymity as well as the NWA algorithm based on such concept by the trajectory clustering; Yang et al.[5] preserves the privacy of the trajectory by using the method of fuzzy region based on the idea of graph theory, the trajectory of spatio-temporal data set is transformed into the form of graph, the trajectory privacy protection is transformed into the problem of k anonymity sub-graph partition, so as to achieve the purpose of k optimizing the privacy; the GC-DM algorithm proposed by Wang et al.[6] comprehensively considers the factors such as time span, position information and trajectory shape, carries out the trajectory clustering to the trajectory based on similarity measure, then carries out the trajectory reconstruction in different clusters so that each cluster contains k trajectories to achieve the purpose of protection the privacy; in Xu and Cai[7] the trajectory of the moving object is recorded by setting up a moving footprint table which records the historical trajectory, and the historical footprint is used to replace the trajectory which is not satisfied with the requirements of privacy by querying the moving footprint in the subsequent release process of trajectory data, so as to achieve the purpose of privacy protection; from the point of view of attackers, Zhao et al.[8] proposes two kinds of trajectory privacy protection schemes based on the possible attack model that might be conducted by the attacker, the first of which generates a number of false trajectories by turning and deforming the threatened trajectory, and adds them into the released spatio-temporal data set to achieve the k anonymity of trajectory; the second scheme is the optimization of the first scheme, to restrain the local threatened trajectory fragments, not allowing it to be added to the finally-released data set to optimize privacy and data availability, Gidófalvi et al.[9] presents a trajectory privacy protection approach of data collection under the server-client framework, in which the motion estimation of different users is separated; these trajectories are used in the server end to carry out the exchange and anonymous operation of trajectory points/fragment between different users before the demand service. However, although the above algorithms have achieved the

k anonymity of the trajectory, which meets the needs of the user's privacy, there are still the following problems:

i. Consider too simple about the attacker's attack, while the actual effect of privacy protection is not sufficient. The model has not considered that attackers would enrich their knowledge in accordance with the result of the first attack. Therefore, when they attack again, the owner of the trajectory may have more risk of privacy leakage.

ii. The retention of data availability is low.

iii. The algorithm does not consider the offset error acceptable to part of application.

iv. Algorithm execution efficiency is not sufficient.

In view of the above problems, this paper enriches the attacker's attack model, based on which a privacy protection algorithm for amending prefix tree is proposed; this algorithm not only achieves the purpose of optimizing the trajectory data privacy, but also largely improves the data availability and the algorithm execution efficiency.

Problem description and model building

Problem description

In order to meet the needs of location-based services, a large number of spatio-temporal data are collected and distributed by all kinds of devices with positioning or sign-in function, which causes that the identity of the user is easily found by the attacker, further leading to the user's privacy reveal. Therefore, the research on trajectory privacy protection is carried out, which should not only ensure the privacy level of user's trajectory, but also guarantee the high quality experience of data provided for each kind of service, so the relationship between privacy level and data availability should be balanced.[10]

Attack model

The released spatio-temporal data set is shown in Table 1 (A); given a attacker set of $A=\left(a_1,a_2,\cdots,a_n\right)$, for any attacker a_i, the part of the data information grasped by him is shown in following Table 1 (B); according to the part of the information grasped by him, the attacker a_i can identify from the spatio-temporal data set that the trajectory of the Object_1 belongs to the user X_1, and the trajectory of the Object_3 belongs to the user X_2, while the trajectory of the Object_1 and Object_5 may belong to user X_3; if it is based on the condition that each trajectory has only one corresponding user, Object_1 can be excluded, so it is speculated that Object_5 may belong to user X_3. Then, the trajectory privacy protection is to modify the data in the spatio-temporal data set, so that probability of attacker re-identifying the corresponding trajectory of users from the released and new spatio-temporal data sets, namely the $P_{attack} \le \frac{1}{k}$; however, for the existing methods, the process of re-attacking still adopts the original information grasped by the attacker, which does not consider that the attacker will update their grasped information in the first time of stealing user privacy information, and the attacker will learn based on the stolen information, so the information grasped by attacker a_i in actual process of the second time of attack is not as shown in Table 2 (A), but the Table 2 (B); if deleting the trajectory therein or conduct transformation in a very abrupt manner and not considering the relevance and sensitivity of trajectory point in time and space, it is very likely that the attacker will regard such point as an abnormal point or a disturbing point, and not to consider such point, so the actual effect of privacy protection is not sufficient; for example, in order to preserve the trajectory of Object_5 not being found, the loc_8 is spatially transformed into loc_8, which does not exist in the information grasped by the attacker, regarding it as the disturbing point and choosing to ignore, so it can be speculated that Object_5 trajectory after being spatially transformed belongs to user X_3.

Table 1 the data included in the spatio-temporal data set

Moving object	**Moving trajectory**
Object_1	$loc_1 \to loc_2 \to loc_3 \to loc_4$
Object_2	$loc_1 \to loc_3 \to loc_5 \to loc_6$
Object_3	$loc_2 \to loc_7 \to loc_6 \to loc_2$
Object_4	$loc_2 \to loc_1 \to loc_5$
Object_5	$loc_2 \to loc_3 \to loc_4$

Table 2 The information obtained by the attacker in one and another attack, respectively

The information obtained by the attacker in one attack

Moving object	**Moving trajectory**
X_1	$\cdots \to loc_1 \to \cdots \to loc_4 \to \cdots$
X_2	$\cdots \to loc_7 \to \cdots \to loc_6 \to \cdots$
X_3	$\cdots \to loc_2 \to \cdots \to loc_4 \to \cdots$

The information obtained by the attacker in one attack

Moving object	Moving trajectory
X_1	$\left\{\begin{array}{l} \cdots \to loc_1 \to \cdots \to loc_4 \to \cdots \\ loc_1 \to loc_2 \to loc_3 \to loc_4 \end{array}\right.$
X_2	$\left\{\begin{array}{l} \cdots \to loc_7 \to \cdots \to loc_6 \to \cdots \\ loc_2 \to loc_7 \to loc_6 \to loc_2 \end{array}\right.$
X_3	$\left\{\begin{array}{l} \cdots \to loc_2 \to \cdots \to loc_4 \to \cdots \\ loc_2 \to loc_3 \to loc_4 \end{array}\right.$

First, the attacker a_i to steal spatio-temporal trajectory data without any privacy protection, then, the attacker a_i according to the part of the incomplete information and the background information he

Citation: Walton LM, Schbley BH, Milliner SW, et al. Investigating the reliability and validity of an intimate partner violence screening tool for use in physical therapy practice. *Int Phys Med Rehab J.* 2017;1(1):1–4. DOI: 10.15406/ipmrj.2017.02.00018

has mastered, the attacker a_i can infer the corresponding relationship between the trajectory and the moving object, and then analyze and find out more privacy information of the moving object, which threatens the privacy of the moving object. The attacker a_i will update the information already mastered based on the information obtained from the first attack, when the attacker a_i with more comprehensive information to attack again, the moving object will be subject to greater threat to privacy, more likely to lead to the disclosure of the privacy of mobile objects. The effect of privacy protection after optimizing the privacy protection algorithm is valued by the leakage of the targets' privacy, when they are attacked again by those attackers. At this time, the attackers' learning ability is not taken into consideration. (As shown in Figure 1)

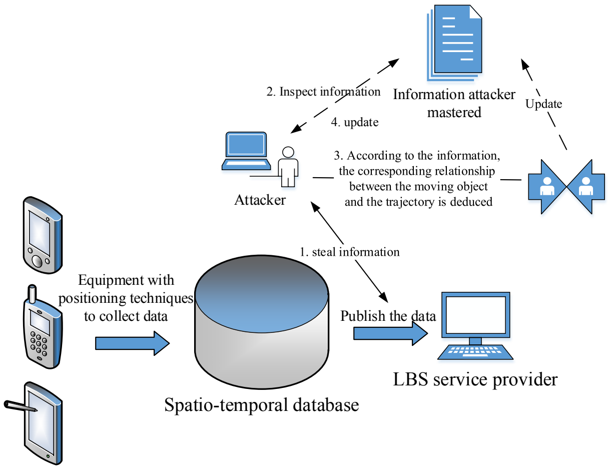
2. Inspect information
Information attacker mastered
Update
4. update
3. According to the information, the corresponding relationship between the moving object and the trajectory is deduced
Attacker
1. steal information
Equipment with positioning techniques to collect data
Publish the data
LBS service provider
Spatio-temporal database

Figure1 Sketch of attack model.

Algorithm design

Relevant definition

Definition 1: Spatio-temporal data se O represents the set of moving objects, recorded as $O = \{o_1, o_2, o_3, \cdots, o_n\}$ tra_{o_i}. Represents the record of n trajectories for moving object o_i.

$$tra_{o_i} = \begin{cases} \left\langle \left(lon_{o_{i\text{-}1}}^1, lat_{o_{i\text{-}1}}^1, t_{o_{i\text{-}1}}^1\right), \left(lon_{o_{i\text{-}1}}^2, lat_{o_{i\text{-}1}}^2, t_{o_{i\text{-}1}}^2\right), \left(lon_{o_{i\text{-}1}}^3, lat_{o_{i\text{-}1}}^3, t_{o_{i\text{-}1}}^3\right), \cdots, \left(lon_{o_{i\text{-}1}}^{m1}, lat_{o_{i\text{-}1}}^{m1}, t_{o_{i\text{-}1}}^{m1}\right) \right\rangle \\ \left\langle \left(lon_{o_{i\text{-}2}}^1, lat_{o_{i\text{-}2}}^1, t_{o_{i\text{-}2}}^1\right), \left(lon_{o_{i\text{-}2}}^2, lat_{o_{i\text{-}2}}^2, t_{o_{i\text{-}2}}^2\right), \left(lon_{o_{i\text{-}2}}^3, lat_{o_{i\text{-}2}}^3, t_{o_{i\text{-}2}}^3\right), \cdots, \left(lon_{o_{i\text{-}2}}^{m2}, lat_{o_{i\text{-}2}}^{m2}, t_{o_{i\text{-}2}}^{m2}\right) \right\rangle \\ \vdots \\ \left\langle \left(lon_{o_{i\text{-n}}}^1, lat_{o_{i\text{-n}}}^1, t_{o_{i\text{-n}}}^1\right), \left(lon_{o_{i\text{-n}}}^2, lat_{o_{i\text{-n}}}^2, t_{o_{i\text{-n}}}^2\right), \left(lon_{o_{i\text{-n}}}^3, lat_{o_{i\text{-n}}}^3, t_{o_{i\text{-n}}}^3\right), \cdots, \left(lon_{o_{i\text{-n}}}^{m3}, lat_{o_{i\text{-n}}}^{m3}, t_{o_{i\text{-n}}}^{m3}\right) \right\rangle \end{cases}$$

Where, a trajectory's length of moving object o_i is denoted as $\left|tra_{o_i}\right| = m$, triad $\left(lon_{o_i}^k, lat_{o_i}^k, t_{o_i}^k\right), k \le m$ is denoted as $point_k$, referring to the location of moving object o_i at time t^k, and each element in the triad respectively represents longitude, latitude and time.

Definition 2: As for the two trajectories of sub trajectory data, namely $tra_{o_i} = \left\langle point_{o_i}^1, point_{o_i}^2, \cdots, point_{o_i}^l \right\rangle$ and $tra_{o_j} = \left\langle point_{o_j}^1, point_{o_j}^2, \cdots, point_{o_j}^r \right\rangle$, if there is a set of integers $1 \le k_1, k_2, \cdots, k_q \le l$ in the trajectory tra_{o_i} enable $point_{o_i}^{k_1} = point_{o_j}^1, point_{o_i}^{k_2} = point_{o_j}^2, \cdots point_{o_i}^{k_q} = point_{o_j}^r$, the article call the trajectory is the sub trajectory of tra_{o_j}, recorded as $tra_{o_j} \subseteq tra_{o_i}$.

Definition 3: Sensitivity, different moving objects have different definitions of privacy, so this paper uses sensitivity to represent the privacy of different trajectories. Sensitivity is the number of visits for each trajectory point $tra\_p = (lon, lat)$ in a trajectory data set TD; the larger the number of visits is the less sensitive the trajectory is and the higher the privacy level is.

Definition 4: Trajectory information gain; the data gain in this paper refers to the sensitivity carried by each point of each trajectory in the trajectory data set TD.

Definition 5: Temporal information similarity set, those in the trajectory data set TD meet the following two conditions belong to the same temporal information similarity set $TISS$:

i. Approximate time span of the whole trajectory.

ii. Have the same stagnation point.

Definition 6: Temporal information relevance set; those in the

Citation: Walton LM, Schbley BH, Milliner SW, et al. Investigating the reliability and validity of an intimate partner violence screening tool for use in physical therapy practice. *Int Phys Med Rehab J.* 2017;1(1):1‒4. DOI: 10.15406/jiratj.2017.02.00018

trajectory data set TD meet the following two conditions belong to the same temporal information relevance set $TIRS$:

i. Approximate sensitivity of each trajectory point in the whole trajectory.

ii. Approximate time span of the whole trajectory.

Definition 7: Sensitivity location, it is the set composed of the trajectory points $tra\_p = (lon, lat)$ with the tiny sensitivity in the given trajectory data set . It is worth noting that the sensitivity location set is variable, and it is about the sensitivity with the trajectory changed according to the privacy protection algorithm; the number of elements in original sensitive location set gradually turns to 0.

Definition 8: Safe trajectory sequence; given the trajectory data set TD and the sensitivity location set SLS , when and only when each trajectory point of certain trajectory $tra_{o_{i-k}}$ in the trajectory data set SLS does not exist in the sensitivity location set SLS , namely, $\forall tra\_p_i \in tra_{o_{i-k}}, tra\_p_i \notin SLS$ the trajectory is safe then.

Definition 9: k anonymity spatio-temporal data; given the trajectory data set TD , probability for each of these trajectories being successfully attacked by the attacker resulting in the privacy leakag $P_{divulge} \leq 1/k$, the spatio-temporal data is called as the k anonymity spatio-temporal data then.

Definition 10: Trajectory chart; given the spatio-temporal data set TD and the trajectory chart $TG = (V, E, S)$, where TG represents the node set of TG , which is the set composed of stagnation point, namely; $\forall e_i \in E, e_i \in MPS$ TG represents the side set of TG , which is the set composed of moving point, namely $\forall e_i \in E, e_i \in MPS$; S represents the sensitivity of each node and side; the average value of the corresponding sensitivity of the stagnation point and the moving point is obtained.

Definition 11: Tolerance error; in the positioning or sign-in service, some services are allowed to have errors; such as the sign-in service, in the releasing of personal position, it is allowed to have error within certain range; for example, the signing-in function for clock in by "ding talk" allows fine tuning within 500 meters. The $terror$ here represents the tolerance error.

The core steps of the algorithm in this paper are as follows (as shown in Figure 2):

a. Pre-process the spatio-temporal data sets TD , transform TD into trajectory chart for storage, and calculate the sensitivity of each trajectory point; tra_p

b. According to the definition 5, each trajectory is divided into different temporal information relevance sets $TIRS$ according to the sensitivity of each trajectory point in each trajectory and the time span of the trajectory.

c. Transform into prefix tree for storage according to the graph.

APT-PP algorithm

APT-PP algorithm proposed in this paper is as follows:

Step 1 Firstly; carry out the preprocessing of trajectory. k

Step 1.1 According to the given spatio-temporal data TD to generate the trajectory sequence, the trajectory generated is composed of the stagnation point and moving point.

Step 1.2 The sensitivity of each trajectory point is calculated according to the generated trajectory sequence.

Step 1.3 According to definition 5, the generated trajectory sequence is divided into each temporal information similarity set.

Step 1.4 Trajectory chart is generated based on the temporal information similarity set obtained by the above preprocessing.

Step 2 Trajectory chart obtained by preprocessing is transformed into prefix tree form for storage.

Step 3 Carry out "grafting" operation to prefix tree, with the specific operations are as follows.

Step 3.1 Count the number of leaf nodes of the prefix tree; if the number of leaf nodes of the prefix tree $|leaf| \geq k$, meaning that the stagnation point satisfies the $_{k-LL}$anonymity privacy protection operation. Skip to the Step3.2 only for the moving point, or both the stagnation point and the moving point skip to the Step3.2 for processing.

Step 3.2 Extract the points with the sensitivity for the stagnation point and the moving point less than , put into two sets respectively, namely $k-LHP$ and $k-LMP$; extract the points with the sensitivity larger than and put them into the two sets of $k-MMP$ and $k-MMP$, and then count the number of trajectory points contained in the twosets, namely $|k-LMP|$ and $|k-LMP|$; if $|k-LHP| \geq k$, selects nos. $\left\lfloor \frac{|k-LHP|}{k} \right\rfloor$ of trajectory points from the set $k-LHP$ to be the replacement point of set, if $|k-LHP| < k$, select the point that is the closest to itself from the set $k-MHP$ as the replacement point, namely $\{replace\_p \mid d_{replace\_tra} \leq terror\}$, where $replace\_p = (lon_0, lat_0)$ and $tra\_p = (lon, lat)$ respectively respects the replacement point in the $d_{replace\_tra}$ set and the point that is to be replaced in the $d_{replace\_tra}$ set currently; $d_{replace\_tra}$ respects the distance between the two points; if the $replace\_p$ does not exist, randomly select one point from the set $k-MHP$ as the replacement point; the operation for $|k-LMP| < k$ is as same as the above description.

Step 4 Traverse the entire prefix tree from the root node, to generate the safe trajectory sequence set SLS that is to be released Table 3.

The 1-29 lines of the above pseudo code are the part about the preprocessing for the spatio-temporal data set; the operation of this part is to compress the spatio-temporal data set into the trajectory sequence (forming the above nodes) to reduce the space. Among them, the 1 to 20 lines are the trajectory sequences set TS generated according to the input spatio-temporal data set TS , and the sensitivity of each trajectory point is calculated, while the 21 to 29 lines are the trajectory sequences set TS generated from the previous preprocessing; the trajectory sequence therein is divided into the temporal information similarity set $TISS$ that conforms the definition 5, namely the approximate time span of the trajectory sequence; trajectory sequences with the same stagnation point are divided into the same temporal information similarity set $TISS$. 32 to 44 lines are the process of generating the corresponding prefix tree for each temporal information similarity set; the trajectory sequence is stored in the form of prefix tree to improve the retrieval efficiency, so as to reduce the computation time. Finding the maximum common sequence, namely the maximum prefix, starting from the starting point of the trajectory sequence tra_i from the generated TP, add this trajectory sequence tra_i to the path $path_i$ of TP, and update the sensitivity of each node (namely the stagnation point and the moving point) in TP; if there is a maximum prefix, add the sensitivity of the intersection of the two paths, otherwise the sensitivity of the node on the path $path_i$ is equal to the sensitivity of the trajectory point in the trajectory sequence tra_i . 45 to 70 lines are the "grafting" process of privacy-protection optimization for prefix trees, which process is to replace the node (namely the trajectory point with $sensitivity < k$) that does not satisfy the privacy protection to other place by the algorithm, so as to achieve the purpose of protection the availability of data to a certain extent and strengthening the data privacy level at the same time. In the process of "grafting", the stagnation points and moving points on each path $path_i$ are divided into four sets $(k-LHP, k-MHP, k-LMP, k-MMP)$ with the as the threshold; determine the set after division, and if the set $k-LHP$ or the $k-LMP$ is larger than the threshold k, randomly select the point as the replacement point of set $k-LMP$ or $k-LMP$, otherwise calculate the distance to be replaced from set $k-MHP$ or $k-MMP$; if there is the replacement point set $\{replace\_point \mid d_{replace\_point} \leq terror\}$ that is not empty, select a point from it for replacement, otherwise randomly select the point from $k-MHP$ or $k-MMP$ for replacement. After the end of the "grafting" operation, start visiting from the root node, each path from the root node to the leaf node is the required safe trajectory SLS sequence at last The algorithm of this paper is divided into two parts, because the number of trajectory contained in the temporal information similarity set $TISS$ generated in the process of preprocessing $|TISS| \in [1, |TD|]$, considering the relationship between the complexity and the loss of data of the algorithm, if the number of trajectory $|TISS| < k/2$, there will be large consumption of complexity to use the algorithm 2; algorithm 2 is proposed based on this consideration for the trajectory sequence that has not been processed with the replacement of stagnation point in the algorithm 1, with the specific procedures as follows:

Step1. Preprocessing of spatio-temporal data is similar to algorithm 1, and the only difference lies in the Step1.3; the temporal information relevance set $TIRS$ in this algorithm is generated according to definition 6.

Step2. Generate the prefix tree TP based on temporal information relevance set $TIRS$, find the points of the stagnation point with the sensitivity less than TI from the prefix tree TP to put into set $S-LHP$, while the points with the sensitivity equal to or larger than k are put into the set $S-MHP$.

Step3. If the $S-MHP$ is not 0, namely $S-MHP \neq \varnothing$, find the point in the set $S-MHP$ closest (namely within the tolerate error range, within $terror$ meters) to the point to be replaced in the set $S-LHP$ for replacement, otherwise find a point from the set $S-LHP$ that is the closest to itself for replacement, to get the amended prefix tree TP_{new}.

Step4. Carry out the "grafting" operation as algorithm 1 to the moving point in the amended TP_{new}. Divide the moving point in the amended TP_{new} into two sets of $k-LMP$ and $k-MMP$ according to k division, if $|k-LMP| \geq k$, selects $\left\lfloor \frac{|k-LMP|}{k} \right\rfloor$ nos. of trajectory points from the set $k-LMP$ as the replacement point, otherwise randomly select one point within the $terror$ -meter tolerate error of the point to be replaced from the set $k-MMP$ for replacement, namely $\{point(lon, lat) \mid d_{replace\_point} \leq terror\}$ .1-11 lines for the pseudo code of algorithm 2 Table 4 are about the preprocessing to the trajectory sequence that is not treated in algorithm 1; the trajectory sequence is divided into the corresponding temporal information relevance set $TIRS$ and the corresponding prefix tree is generated. The 12 to 34 lines of the pseudo code are the amendment of the prefix tree, and the processing of the amended prefix tree to the stagnation point changes from the original $O(n^2) \rightarrow O(n)$, which, to some extent, accelerates the efficiency of the algorithm.

Analysis for data availability and privacy level

Date availability and privacy level are discussed in this section, and the article will show that the spatio-temporal data republished by APT-PP algorithm of the paper has validly improved the privacy level in this section, and at the same time, the availability of data is also not greatly reduced, and the loss of quantity of information is in a relatively stable and acceptable range. This section will define the relevant measurement parameters in this section and discuss how to measure the availability of data and the protection level of privacy.

Protection level of privacy

[11–13] Measurement is carried out by probability recognized intensively from spatio-temporal data republished by attacker, and after optimizing the privacy according to the algorithm proposed in the paper, the k- anonymity demand of the trajectory is met, proving as follows:

The original given spatio-temporal data set SLS obtains the safety trajectory sequence set by APT-PP algorithm proposed in the paper. Set SLS is the prefix tree which is able to achieve the privacy protection after correction, composed by path (all other node paths except for the root node) traversed from root node to leaf node. Because that in the "grafting" process of prefix tree, the trajectory point of which the sensitivity does not meet the requirements is transferred or replaced to another place for the purpose of making the sensitivity reach threshold k above, finally the sensitivity of stagnation point and moving point included in amended prefix tree is greater than or equal to k , therefore, the final path sequence SLS acquired by depth traverse meets the k- anonymity demand of the trajectory.

Loss level of information and tolerance error of information

Loss level of information and tolerance error of information are used for measuring the information distortion level of spatio-temporal data caused by the modification of trajectory point at a certain extent in the process of algorithm processing and privacy optimization, respectively represented by MIL and $MITEL$. However, the former is the measure index aimed at accurate and strict service as demand, and the latter is the measure index aimed at service allowed error such as location and check in.

$$\begin{cases} MIL=1-\dfrac{\sum_1^{|PTS|}\left|PT_{new}\cap PT\right|}{|TD.point|} \\ MITEL=1-\dfrac{\sum_1^{|PTS|}\left|PT_{terror}\cap PT\right|}{|TD.po\,\mathrm{int}|} \end{cases}$$

PTS in the above formula means the prefix tree forest, $|PTS|$ refers to the quantity of prefix tree contained in the forest, PT_{new} and PT_{terror} respectively refer to the prefix tree after finishing the algorithm under the case of tolerance error for $terror=0$ and $terror\neq 0$, PT represents the prefix tree before carrying out the algorithm, and $TD.point$ refers to the number of all trajectory points included in given original spatio-temporal data set TD .

Algorithm complexity

The achievement of the algorithm in the paper is divided into three parts: preprocessing, generation of prefix tree and "grafting" operation, of which the complexity is $O\left(|tra|\times|point|\right)$ in the processing stage, $O\left(|tra|\times|pathpoint|\right)$ in the generation stage, and $O(|tra|\times|point|)$ in the "grafting" operation stage, therefore, the complexity of the algorithm is $O(|tra|\times|point|)$, where $|tra|$ refers to the number of trajectory included in the spatio-temporal set TD , $|point|$ means the number of trajectory point in one trajectory $|pathpoint|$ represents the number of trajectory point included in the path of the maximum prefix in prefix tree and represents the privacy level.

Experimental verification

Description for Experimental data and environment

To verify the superiority of APT-PP algorithm proposed in the paper, the algorithm is assessed in terms of detail by truthful data T-drive provided by Microsoft Asia Research Institute.[14–15] The data set T-drive collected the moving trajectory of 10357 taxies in Beijing area from February 2 to February 8, 2008, with the interval of 5 seconds for information collection, of which 15000000 trajectory points are included in total, and the total distance of moving trajectory reaches 9000000 km with the average sampling time for 177 seconds and average interval distance of sampling for 632m. The specific attribute of T-drive is shown in Table 5, and the format of moving trajectory recorded in data set is shown in Figure 3.

As shown in Figure 3, each line of data represents a spatio-temporal data collected by sampling point, separated by commas, respectively representing taxi ID, sampling data /time, longitude as well as latitude. As the time span involves the multi-day data, the data preprocessing is carried out before testing data. The data set contains partial repeating data, that is, multiple records in the same time, so the data needs to be cleaned before processing, and at the same time the position information data of moving object in the day composes a trajectory, that is, each moving object has seven trajectories (namely, a 7-day moving path is recorded). The APT-PP algorithm is written by Python, and is tested in hardware condition of Intel(R) Core(TM) i7-4610M CPU @ 3.00GHz, 8.00GB memory and Microsoft Windows 10, and the results are shown by MATLAB.

Experimental results and analysis

The algorithm in literature[4] (called NWA algorithm) and that in literature[6] (called GC-DM algorithm) are used as comparison algorithm in the test, and the contrast experiment is carried out respectively from the loss of information, tolerated loss of information as well as execution efficiency of algorithm to verify the superiority of algorithm. The reason for choosing the literature[4] is that such algorithm is a classic algorithm of privacy protection, and presently a lot of methods of privacy protection are derived from such algorithm and inspired by it, but the algorithm in literature[6] is an improved method of comparatively novel algorithm in terms of privacy protection of trajectory.

Experiment 1: Contrast test between size of privacy protection level and loss level of information. The test for contrast algorithm (NWA algorithm and GC-DM algorithm) and APT-PP algorithm of the paper in T-drive data set, and the comparison of information loss conditions MIL caused by different algorithms on data set with k value (anonymity level) are described in Figure 4. As can be seen from the graph that, with the increase of k value (anonymity level), the issued data set generated after processing of three algorithms on original data set and the loss conditions compared with original data set are all increased, however the algorithm of the paper represented by a blue straight line has a outstanding advantage compared with contrast algorithm, of which GC-DM algorithm has a few difference with the algorithm of the paper in the process of implementation when k value (anonymity level) is not large, but with the increase of k value (anonymity level), the advantage of algorithm in the paper is more and more obvious, and with the change of k value (anonymity level), there will be no sudden increase or decrease conditions on implementation of the algorithm of the paper with a smoothing curve.

Experiment 2: Contrast test between k size of privacy protection level and tolerated loss level of information. The test for contrast algorithm (NWA algorithm and GC-DM algorithm) and APT-PP algorithm of the paper in T-drive data set with $terror$ for 500m, and the comparison of information tolerance error loss conditions $MITEL$ caused by different algorithms on data set with k value (anonymity level) are described in Figure 5. Comparing with Figure 4,5 it can be found that, the loss conditions of the algorithm (NWA algorithm and GC-DM algorithm) have almost no difference in the same data set after two experiments, however the algorithm APT-PP has significant optimization in the experiment of loss level of tolerance error of information, and the APT-PP algorithm optimize the presence of tolerable offset $terror$ in the processing because the data issued in the experiment is served for the check in application of tolerance error. In the case of privacy protection, the geography location and time shall be fully considered in the offset change of trajectory point, so the quality of check in service is protected to the maximum extent, and the loss level of tolerance error of information has an outstanding advantage compared with other algorithms, there will be a good effectiveness when applying in check in demand.

Experiment 3: Test for k size of privacy protection level, tolerance error $terror$ and tolerated loss level of information. The test of APT-PP algorithm in T-drive data set, and the conditions for tolerated loss level of information changing with k (anonymity level) and tolerance error $terror$ are shown in Figure 6. As can be seen from figure, there is a positive correlation between APT-PP algorithm and k (anonymity level), but with a negative correlation for tolerance error $terror$, with the increase of tolerance error $terror$, the acceptable error offset is more greater, the correction of a lot of trajectory point in the process of prefix tree of corrected point is in the tolerance range, and the change of tolerated loss level of information MITEL is not influenced by the offset. The application of such positioning requirements is not influenced by the final formative trajectory safety sequence as an issued data set, and the location check also can be carried out for it, therefore as shown in the Figure 6, the APT-PP algorithm will be influenced by the change of tolerance error $terror$, but the NWA algorithm and GC-DM algorithm will not be influenced.

Experiment 4: Contrast test between k size of privacy protection level and execution efficiency of algorithm. The test for contrast algorithm (NWA algorithm and GC-DM algorithm) and APT-PP algorithm of the paper in T-drive data set, and the conditions of execution time of algorithm changing with k are described in Figure 7. There is a downtrend on execution time of three algorithms with k (anonymity level), because with the increase of k value (anonymity level), there is an increasing trend on privacy requirements. The trajectory point that the partial data cannot meet will be processed simply (namely directly remove and other operations), therefore the execution time of algorithm is also reduced, but comparing with Test 1 and Test 2, it is known that although the time is accelerating, the loss of information of spatio-temporal data set caused by sample processing of trajectory point is increasing. The APT-PP algorithm has a outstanding advantage compared with contrast algorithm in the execution efficiency of algorithm, however with the change of k (anonymity level), there is a few change on the execution time of algorithm, because in the process of implementation, the larger the k value (anonymity level) of algorithm of the paper is, the more the divided time information association set or the time information similarity set is, namely the size of single set is more and more small, but still each set will be processed. If the quantity of multi-set is 1, it will be uniformly reprocessed, so the descent range of execution time is not large, and is only slightly superior to GC-DM algorithm when k value (anonymity level) reaches 40 or 48.

Conclusion

To solve the privacy disclosure of spatio-temporal, the APT-PP algorithm is proposed in the paper, which not only considers the information of location but also considers the relation of time span in the process of optimizing tracing privacy, and defines the concept of time information similarity set and time information association set, so that the spatio-temporal data set is classified and processed in different levels; on the basis of it, the data storage in the form of prefix tree has validly improved the efficiency of data search, and the "grafting" operation is carried out on the trajectory stored in prefix tree to optimize the privacy level of issued spatio-temporal data and the loss level of data. Finally, the contrast verification is carried out for algorithm proposed in the paper by true data set, and the results show that APT-PP algorithm has a comparatively outstanding advantage in the execution efficiency and the quality of issued data set. APT-PP algorithm in the application of location based service has important significance. We will do further study on the privacy protection of different levels according to people's needs.

Acknowledgements

None.

Conflict of interest

The author declares no conflict of interest.

References

1. Zheng Y, Zhou X. *Computing with Spatial Trajectories*. USA: Springer; 2011.
2. Trujillo-Rasua R, Domingo-Ferrer J. On the privacy offered by (k, δ) -anonymity. *Information Systems*. 2012;38(4):491–494.
3. Barrios C, Motai Y, Huston D. Trajectory Estimations using Smartphones. *IEEE Transactions on Industrial Electronics*. 2015;62(12):1–1.
4. Abul O, Bonchi F, Nanni M. Never Walk Alone: Uncertainty for anonymity in moving objects databases. *icde*. 2008. p. 376–385.
5. Yang Jing, Zhang Bing, Zhang Jian-pei, et al. Personalized trajectory privacy preserving method based on graph partition. *Journal on communications*. 2015;3:1–11.
6. WANG Chao, YANG Jing, ZHANG Jian-pei. Privacy preserving algorithm based on trajectory location and shape similarity. *Journal on communications*. 2015;36(2):144–157.
7. Xu T, Cai Y. Exploring historical location data for anonymity preservation in location-based services. *IEEE Xplore Digital Library*. 2008. p. 547–555.
8. Zhao Jing, Zhang Yuan, Li Xing-Hua, et al. A trajectory protection approach via trajectory frequency suppression. *Chinese Journal of computers*. 2014;37(10):2096–2106.
9. Gidófalvi G, Huang X, Pedersen T B. Privacy: preserving trajectory collection. ACM Sigspatial International Symposium, Proceedings DBLP, USA: ACM; 2008.
10. Sampigethaya K, Li M, Huang L. AMOEBA: robust location privacy scheme for VANET[J]. *IEEE Journal on Selected Areas in Communications*. 2007;25(8):1569–1589.
11. Theodorakopoulos G, Shokri R, Troncoso C. Prolonging the hide-and-seek game: optimal trajectory privacy for location-based services. 2014. p. 73–82.
12. Komishani E G, Abadi M, Deldar F. PPTD: Preserving personalized privacy in trajectory data publishing by sensitive attribute generalization and trajectory local suppression. *Knowledge-Based Systems*. 2016;94:43–59.
13. Huo Z, Meng X, Hu H. You can walk alone : trajectory privacy-preserving through significant stays protection. *International Conference on Database Systems for Advanced Applications*. 2013;7238:351–366.
14. Yuan J, Zheng Y, Xie X. Driving with knowledge from the physical world. *ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. 2011. p. 316–324.
15. Yuan J, Zheng Y, Zhang C. *T-drive: driving directions based on taxi trajectories*. ACM Sigspatial International Symposium on Advances in Geographic Information Systems; USA: Acm-Gis 2010; 2010. p. 99–108.
16. Zhangjie Fu, Fengxiao Huang, Xingming Sun, et al. Enabling Semantic

Search based on Conceptual Graphs over Encrypted Outsourced Data. *IEEE Transactions on Services Computing*. 2016. 1 p.

17. Zhihua Xia, Xinhui Wang, Xingming Sun, et al. A Secure and Dynamic Multi-keyword Ranked Search Scheme over Encrypted Cloud Data. *IEEE Transactions on Parallel and Distributed Systems*. 2015;27(2):340-352.