A New Trajectory Privacy Protection Scheme for Resisting Similarity Attack

Junjie Jia and He Huang
College of Computer Science and Engineering, Northwest Normal University, Lanzhou 730070, China
Email: 654205992@qq.com; 654179553@qq.com

Abstract. Aiming at trajectory privacy disclosure caused by high slope similarity in trajectory anonymous set, trajectory \((k, e)\) - anonymity model is proposed to resist the trajectory similarity attack. The algorithm uses the trajectory slope as sensitive value to select at least \(k\) different trajectories into an equivalence class and requires that the slope of different trajectories is at least \(e\). The experimental results show that the proposed algorithm can effectively prevent the trajectory similarity attack caused by the high similarity of the trajectory in the anonymity set and more effective trajectory privacy protection has been realized.

1. Introduction
With the rapid development of mobile devices and positioning technology, a large number of mobile object trajectory data are generated. With the issue of trajectory data, privacy protection technology not only protects the privacy of trajectory data, but also ensures the high availability of data. Therefore, how to protect the location of mobile objects and the privacy of trajectories has become a crucial problem.

At present, the technology of trajectory privacy protection is roughly divided into three categories. (1) False data: By adding false trajectories to interfere with the original data, the data can be protected. At the same time, the data of the disturbed data will not be seriously distorted, but this method has large storage capacity and low availability. (2) Generalization: By extending all the sampling points on the track to the corresponding anonymity regions to achieve the purpose of privacy protection, it is the mainstream trajectory privacy protection method. (3) Inhibition: There is a choice of release trajectory data, not the release of certain sensitive positions on the trajectory or the location of frequent visits, the disadvantage is that the loss of information is too large.

Trajectory similarity is an important factor affecting trajectory clustering and anonymity in trajectory privacy protection. If the similarity of \(k\) trajectories is very high, they will coincide at a certain track or a certain point or pass the same sensitive area, which will lead to privacy leakage. Literature [1] is proposed that there are some differences between the trajectories of \(k\) trajectories in \(k\)-anonymity groups, for reduce the risk of privacy leakage. It uses Minimum Bounding Rectangle (MBR) to ensure that the difference is greater than a threshold. Based on the literature [2], a personalized method based on graph partitioning is designed. Based on the distance and direction between trajectories, consider the weights between trajectories and transform the construction of \(k\)-anonymity sets into trajectory partitions. In literature [3], the method of using the angle and distance between the trajectories to measure the edge weight of the trajectory map is proposed. Literature [4] proposes a privacy preserving algorithm based on trajectory shape diversity. The algorithm avoids the risk of trajectory privacy leakage by constraining the angle between trajectories to avoid a large number of highly similar trajectories in the set. In this paper, we also aim at the problem of privacy
leakage caused by similar trajectories. We not only consider the time and spatial attributes of trajectories, but also consider the slope property of trajectories, and propose a trajectory (k, e) - anonymity model for resisting trajectory similarity attacks.

Based on trajectory k- anonymity, we use trajectory slope as sensitive attribute and track slope to evaluate trajectory similarity. We propose trajectory (k, e) - anonymity model. The trajectory (k, e) - anonymity model requires that the correlation coverage of trajectories is at least k, and the equivalence class internal slope difference is at least e. This model can effectively solve the problem of privacy leakage caused by high trajectory similarity in trajectory equivalence class.

2. Related Work
Trajectory privacy [5] is a special personal privacy. It refers to sensitive information contained in personal trajectories, such as certain locations, occasions, or other personal information (such as family address, working place, living habits, etc.). The trajectory privacy protection is to ensure that the sensitive information of the trajectory itself is not leaked, and to prevent the attacker from infer the user's personal information. At present, the leakage of trajectory privacy in the data release is mainly divided into two categories:

(1) The frequent access of the user trajectory to a certain location leads to the privacy disclosure of the mobile object. If someone visits a hospital in a certain period of time, an attacker can infer that this person is likely to have some kind of disease in the near future.

(2) Privacy leaks caused by the relationship between the trajectory of the moving object and the external knowledge. If someone fixed every morning from point A to point B of every night, fixed from point B to point A, through data mining, the attacker can easily infer that A is a home address of a person, and B is a unit of work, and then through the zip code, work units and other public information, personal privacy and trajectory privacy are all leaked.

2.1. Trajectory k-Anonymity

Definition 1(Trajectory k- anonymity)Given a number of trajectories, for any trajectory \( T_i \), if and only when \( t_i \) at any sampling time, at least \( k-1 \) trajectories at the corresponding sampling locations and \( T_i \) are in the same generalization area, these trajectories satisfy the trajectory k- anonymity. In general, the greater the value of \( k \) is, the better the privacy protection effect, but the greater the loss of information.

Definition 2(Trajectory equivalence class)If there is a set of trajectories \( \{ T_1, T_2, \ldots, T_i \} \), when \( T_i \in \{ T_1, T_2, \ldots, T_i \} \), the track \( T_i \) has the same starting time and ending time. Then we call this trajectory set an equivalent class of trajectories, that is, the trajectories of equivalent class have the same starting time and ending time.

2.2. Trajectories (k, e) – Anonymity

The trajectory k-anonymity model does not take account of the trajectory shape factor when achieving the anonymity protection. If the anonymity set and the internal trajectory pass through too high, it will lead to user privacy leakage. In this paper, a trajectory (k, e) - anonymity model is proposed for this problem. The trajectory (k, e) - anonymity model regards trajectory slope as sensitive attribute. When meeting the trajectory k-anonymity, it is required that the slope difference value of different trajectories in the trajectory equivalence class is at least e.

Definition 3(Trajectory slope)The slope of the trajectory fragment starting point and the end point is the slope of the line, and the slope of the trajectory segments between the sampling points in the same space-time condition is selected. If the location of the sample trajectory fragment \( T \) is \( (x_b, y_b) \) at the beginning time, the end position is \( (x_e, y_e) \), then the slope of the trajectory \( K_n \) is;

\[
K_n = \frac{y_e - y_b}{x_e - x_b}
\]

When \( x_e = x_b \) or \( y_b = y_e \), set \( K_n = 0 \).

Definition 4 (Trajectory (k, e) - anonymity) each trajectory equivalence class contains at least k
different slope trajectories, and the difference value of the trajectory slope is at least \( e \).

2.3. Trajectory Synchronization Processing

The two nonsynchronous trajectories are extended or reduced to make the synchronization trajectory, including two ways of synchronous extension and synchronous deletion. As shown in Figure 1, figure 1 (a) reflects the addition of some space-time sequences for \( T_2 \) with shorter motion time, so that it is the same as the time interval of the trajectories within the ensemble to be merged, and the trajectories should be regarded as uniform rectilinear motion. Figure 1 (b) is a synchronous deletion process, deleting the longer moving trajectory \( T_2 \).

![Figure 1. Two cases of synchronous processing methods](image)

**Definition 5** (Synchronous trajectory) The two trajectories have the same start and end time, and have the same time coordinate sequence, then the two trajectories are synchronous.

**Definition 6** (Maximum trajectory distance) The maximum value of the Euclidean distance between two synchronous trajectories at the same sampling point at the same sampling time in the time of the survival of the trajectory. For example, two synchronous trajectories \( T_p \) and \( T_q \), and their time coordinate sequences of their trajectories are \( \{(x_1^p, y_1^p), (x_2^p, y_2^p), \ldots, (x_n^p, y_n^p)\} \) and \( \{(x_1^q, y_1^q), (x_2^q, y_2^q), \ldots, (x_n^q, y_n^q)\} \), their maximum trajectory distance is;

\[
\max_{i \in \{p, q\}} \text{dis} \left[ (x_i^p, y_i^p), (x_i^q, y_i^q) \right] = \max_{i \in \{p, q\}} \sqrt{(x_i^p - x_i^q)^2 + (y_i^p - y_i^q)^2} \quad (2)
\]

**Definition 7** (Trajectory difference value) The trajectory sampling time is two synchronous trajectories \( T_p \) and \( T_q \) of \( M \), and the slope of \( T_q \) relative to \( T_p \) is poor. The slope difference of trajectories can reflect the difference degree of two track segments under the same time and space condition;

\[
TD(T_p, T_q) = \sqrt{(K_p - K_q)^2} \quad (3)
\]

For a trajectory anonymity set \( T: \{T_1, T_2, \ldots, T_n\} \), calculate the trajectory \( T_i(T_i \in T) \) the average difference of all values within the trajectory set.

\[
AD = \frac{\sum_{i=1}^{n}(T_p, T_q)}{n} ; T_p, T_q \in T \quad (4)
\]
Definition 8 (Trajectory information loss [6]) After the processing of the original trajectory $T$, the ID calculation formula at the sampling time $t$ is as follows:

$$ID_t(T, T^*) = \begin{cases} Dis(T[t], T^*[t]), & T^* \text{ is meaningful at } t \\ \Omega, & \text{other} \end{cases}$$  \hspace{1cm} (5)

When $T^*$ is meaningful at $t$, the distance between the original trajectory $T$ and the center trajectory $T^*$. The calculation formula is $Dis(T[t], T^*[t]) = \sqrt{(x_i - x_i)^2 + (y_i - y_i)^2}$, $\Omega$ is a constant, represents a penalty for the deletion of the position, in this paper set $\Omega = 10^5$. The information loss for the entire trajectory can be recorded as:

$$ID(T, T^*) = \begin{cases} n\Omega, & T \text{ was deleted} \\ \sum_{t \in \mathcal{T}_N} ID_t(T, T^*) & \text{other} \end{cases}$$  \hspace{1cm} (6)

Where $\mathcal{T}_N$ is the sampling time contained in the trajectory $T$, $n = |\mathcal{T}_N|$, the information loss of the $D^*$ formed by the trajectory database $D$ after anonymity

$$ID(T, D^*) = \sum_{t \in \mathcal{T}_N} ID_t(T, T^*)$$  \hspace{1cm} (7)

2.4. Trajectory Similarity Attack

If there are a number of highly similar trajectories of an equivalence class, especially the same slope track, then the attacker can easily be released through the trajectory to infer the approximate motion mode of these tracks and sensitive attributes, if the track is released through the same sensitive region, the attacker can infer all tracks of the set through the sensitive region, user privacy usually cannot be protected, such as the above attack mode called trajectory similarity attack. As shown in Figure 2 (a), when there is a large number of trajectories with similar slopes in equivalence class, once $T_3$ is released, because all trajectories in the collection are highly similar, attackers can easily deduce the motion patterns of other users, and user privacy is easy to be leaked. In addition, as shown in Figure 2 (b), when attackers match the movement patterns of the anonymity set with the geographical location, if the released trajectory $T_3$ passes through the sensitive area, the attacker can infer that $T_1$ and $T_2$ pass through the anonymous area, resulting in privacy leakage.

Figure 2. Two cases of privacy disclosure
Among them, the difference in slope trajectory anonymity set value is greater, indicating the similarity between anonymity set trajectory is lower, the attacker is difficult to use published trajectory to infer other trajectory motion patterns: if the difference in slope is smaller, indicating the similarity between anonymity set track high, there may be a large amount of similar trajectory, trajectory privacy are likely to leak.

A similar trajectory equivalence class, the slope of trajectory set may be very similar, therefore, trajectory (k, e) - anonymity model, not only to meet the requirement of trajectory equivalence class k-anonymity, but also meet the e slope difference, that is trajectory equivalence class is composed of k trajectories which the slope of trajectory satisfy e difference, prevent privacy leaks caused by a large number of similar trajectories in an equivalence class there are.

2.5 Measure method of leakage risk

The risk assessment of leak risk mainly evaluates the safety of anonymity tables, that is, the possibility of an anonymous table record recording the record of the original table. Literature [7] proposed a record link model based on distance to measure the risk of disclosure of anonymous tables.

**Definition 9 (Link success)** Any track record in the anonymity table is t. We calculate the distance from t to all the track records in the original table, get the nearest record set t` of distance t, and the next near record set t``. If t` or t`` is t anonymity, it is called t link success.

The idea of distance based evaluation is as follows: the proportion of successful records that is linked by anonymity tables is used as a measure of the risk of leakage. Link_records are the number of successful records linked by anonymity tables, and total_records is the total number of records. The risk value of leakage is LR.

\[
LR = \frac{\text{link_records}}{\text{total_records}}
\] (8)

3. Trajectory (k, e) - Anonymity Model

3.1. Algorithm Description

The trajectory (k, e) anonymity model is mainly aimed at the problem of high similarity in the trajectory anonymity set. The trajectory slope is regarded as the sensitive value on the basis of the original trajectory k-anonymity, and the difference value of the different trajectories in the equivalence class is at least e.

In order to minimize the information loss, we first pre-process the trajectory data set, so that the starting time and ending time of different trajectories in the trajectory equivalence class are the same. The start time and end time on the same trajectory in the same equivalence class, if an equivalence class in the trajectory of insufficient number of k, and find the equivalence class of the start and end of another equivalence class of the most similar moment, and the equivalence class as target synchronization processing, after the two equivalence classes combined, and so on, get a series of trajectory number is not less than k trace equivalence class. The algorithm 1 describes the process of pre-processing.

**Algorithm 1 Pre-processing of trajectory data set**

Input: Enter the original data table F
Output: The table \( F^e \) obtained after the synchronization process

1. while F ! =null;
2. {If \( T_i \) has the same start time and end time in F
3. Select these \( T_i \) as a class
4. i++;}
5. for all \( F_i \)
6. {if \( |F_i| < k \)
7. Select the nearest equivalent class \( F_j \);
8. Synchronizing \( F_j \) with \( F_i \);
9: merge $F_i$ and $F_j \rightarrow F_i$;
10: else
   
   // the number of $F_i \geq k$
   
   Take the $F_i$ out of the table $F$ and add to the $F^e$
11: put $F^e$

---

Algorithm 2 Trajectory (k, e) - Anonymity

1. For every equivalence class, calculate the slope of all trajectories, find the trajectory $A$ which slope is the smallest difference minimum of slope difference with the slope average $X$ trajectory, find out the farthest distance trajectory $B$ and find out the farthest trajectory $C$ from $B$.
2. Select the trajectory $B$ as the center trajectory, choose the nearest $k$ trajectories, if the slope difference value is more than $e$, make a class of them, or continue to search the nearest trajectory until meet the difference in slope value is more than $e$.
3. Select the trajectory $C$ as the center trajectory, choose the nearest trajectories, if the slope difference value is more than $e$, make a class of them, or continue to search the nearest neighbor trajectory until meet the difference in slope value is more than $e$.
4. If the remaining number of trajectories had difference in slope value is more than $2k$ and more than $e$, repeat step1-3;
5. If the remaining number of trajectories between $[k-2k]$ and the slope difference value is more than $e$, then form a class.
6. If the number of remaining trajectories is less than $k$, assigned to the nearest class.

3.2. Algorithm Analysis

Trajectory (k, e) - anonymity algorithm considers the difference of slope while considering the similarity, ensuring the k-anonymity and the slope e-difference. Compared with the trajectory k-anonymity, the privacy protection more effective, can effectively resist the attack trajectory similarity, the bigger value of $e$ is, the better privacy protection effect, but the loss of information will inevitably increase, the smaller the value of $e$, the information loss is smaller, but the privacy protection effect is poor, user can set the value of $e$ according to their own needs, the value of $e$ in the range of demand $(0.1\sim1)$ is better.

4. Experiment and Analysis

4.1. Experimental Environment

The experiment uses a road network data generator to generate 7035 synthetic trajectories, the area of the region is 23572m $\times$ 26915m, it contains 6105 positions, a total of 7035 trajectories. The sampling time of the longest trajectory is 40, the number of the shortest trajectory sampling time is 2, the maximum sampling time is 20s. Experimental environment: Inter(R) Core(TM)i5-7300HQ CPU @ 2.50GHz, 8G memory, the operating system is Windows 10, the algorithm is implemented using Visual C++ 6.0.

4.2. Analysis of Experimental Results

Trajectory (k, e) - anonymity algorithm is mainly improved on the traditional trajectory k-anonymity algorithm. Because the algorithm takes the trajectory slope as the sensitive attribute and restricts it, the information loss is slightly higher than the traditional trajectory k-anonymity algorithm, but the privacy protection effect is better.

4.3. Information Loss

Figure 3 reflects the change of the information loss of the trajectory k-anonymity model and the trajectory (k, e) - anonymity model with the value of k ($e=0.3$). The trajectory (k, e) - anonymity model has a slightly higher information loss than the trajectory k anonymity method, because when clustering, we should not only consider the k-anonymity, but also ensure that the slope difference value is at least $e$. In order to protect the privacy of the trajectory better, the information loss degree of the algorithm is acceptable.
4.4. Analysis of Operation Time

Figure 4 reflects trajectory (k, e) - anonymity model (e=0.3) compared with trajectory k anonymity in running time, trajectory (k, e) - anonymity model requirements in the construction of trajectory equivalence class, meet the k-condition of anonymity at the same time to meet the trajectory slope difference value is at least e, so trajectory (k, e) - anonymity model in k value under the condition of the running time increases gradually, but for the protection of privacy of the trajectory, the time increment is acceptable.

![Figure 3. Information loss of two different algorithms with different k values](image3)

![Figure 4. Running time of two different algorithms with different k values](image4)
4.5. Risk Assessment of Leakage

As seen in Figure 5, the trajectory \((k, e)\) - anonymity model is increasing with the slope value of \(e\), and the information loss degree of the different \(k\) values is increasing. Because with the increase of \(e\), the difference in the deviation of the slope of the trajectory equivalence class also increases, which leads to the increase in the distance of the trajectory equivalence class, thus increasing the amount of information loss. As shown in Figure 6, although the amount of information loss is increasing, the rate of increase is slow, and the risk of leakage is obviously declining. The trajectory \((k, e)\) - anonymity algorithm can obtain more secure anonymity data with the increase of the value of \(e\).

5. Conclusion
The trajectory \((k, e)\) - anonymity model is improved based on traditional trajectory \(k\)-anonymity
model, in this paper, the trajectory slope is treated as the sensitive attribute, we regard trajectory slope as sensitive attribute and increase the difference constraint of sensitive attribute in trajectory equivalence class, so that the k trajectories in anonymity sets satisfy e-difference, avoid a lot of highly similar trajectories, and reduce the risk of trajectory privacy leak.

From the security point of view, our trajectory (k, e) - anonymity model obtained by this algorithm is more capable of resisting the trajectory similarity attack. Due to constraint enhancement, information loss is slightly larger in this paper, but it effectively solves the problem of trajectory privacy leakage caused by the high similarity of track slopes in anonymity sets.

6. References

[1] MACHANAVAJJHALA A, KIFER D, GEHRKE J, et al. l-diversity: privacy beyond k-anonymity [J]. ACM Transactions on Knowledge Discovery from Data, 2007, 1(1): Article No.3.

[2] YANG J, ZHANG B, ZHANG J P, et al. Personalized trajectory privacy preserving method based on graph partition [J]. Journal on Communications, 2015, 36 (3): 1-11.

[3] GAO S, MA J, SUN C, et al. Balancing trajectory privacy and data utility using a personalized anonymization model [J]. Journal of Network & Computer Applications, 2014, 38 (1): 125 - 134.

[4] YUAN G, XIA S X, ZHANG L, et al. Trajectory clustering algorithm based on structural similarity [J]. Journal on Communications, 2011, 32 (9): 103-110.

[5] HUO Z, MENG X F. A survey of trajectory privacy-preserving techniques [J]. Chinese Journal of Computers, 2011, 34 (10): 1820-1830.

[6] SUN Dandan, LUO Yonglong, et al. Privacy protection algorithm based on trajectory shape diversity [J]. Journal of Computer Application, 2016, 36 (06): 1544-1551.

[7] Domingo-Ferrer J. Microaggregation for Database and Location Privacy [C]// International Workshop on Next Generation Information Technologies and Systems. Springer Berlin Heidelberg, 2006: 106-116.

[8] Sweeney L. K-anonymity: a model for protecting privacy. International Journal on Uncertainty, Fuzziness and Knowledge-based Systems, 2002, 10(5): 557-570

[9] ABUL O, BONCHI F, NANNI M. Anonymization of moving objects databases by clustering and perturbation [J]. Information Systems, 2010, 35(8): 884 - 910

[10] ABUL O, BONCHI F, NANNI M. Never walk alone: uncertainty for anonymity in moving objects databases [C]// ICDE 2008: Proceedings of the 2008 IEEE 24th International Conference on Data Engineering. Piscataway, NJ: IEEE, 2008: 376-385.

[11] Mohammed N, Fung B M, Debbabi M. Walking in the crowd: Anonymizing trajectory data for pattern Analysis//Proceeding of the 18th ACM Conference on Information and Knowledge (CIKM 2009). Hong Kong, 2009: 1441-1444

[12] Sun Y, Chen M, Hu L, et al. ASA: Against statistical attacks for privacy-aware users in Location Based Service [J]. Future Generation Computer Systems, 2016, 70.

[13] Gao S, Ma J, Sun C, et al. Balancing trajectory privacy and data utility using a personalized anonymization model [J]. Journal of Network & Computer Applications, 2014, 38 (1): 125-134.

[14] Ye A, Li Y, Xu L. A novel location privacy-preserving scheme based on l -queries for continuous LBS [J]. Computer Communications, 2016.

[15] Shin H, Vaidya J, Atluri V. Anonymization models for directional location based service environments. Computers & Security 2010a; 29 (1): 59 – 73.

[16] ABUL O, BONCHI F, NANNI M. Anonymization of moving objects databases by clustering and perturbation [J]. Information Systems, 2010, 35 (8): 884 - 910.