Walk Alone and Be Fast: Trajectory Privacy-preserving in Complicated Environment

Zheng Huo#, Ping He, Ruoyan Wei
Information Technology School
Hebei University of Economics and Business
Shijiazhuang, China
#huozheng@ruc.edu.cn

***Abstract*—Trajectories are location samples ordered by sampling time, which is useful to multiple mobility-related applications. However, publication of these trajectories may cause serious personal privacy leakage. In this paper, we propose an approach called Walk Alone and Be Fast (WABF) to protect trajectory privacy against semantic location attack and maximum moving speed attack. WABF reduces the whole trajectories' exposure probability. At last, we conduct a set of comparative experimental studies on a real-world data set, the results show that WABF is effective and the information loss is much lower than k-anonymity methods.**

Keywords—Privacy-preserving, Trajectory Data publication, Maximum speed attack

I. INTRODUCTION

With the development of positioning techniques and location-aware devices, numerous locations and traces of moving objects are collected and published. Mining and analyzing trajectories is beneficial to multiple novel applications. Although publishing trajectories is beneficial to mobility-related decision making processes, it may cause serious threats to individuals' personal privacy, such as, living habits, health conditions, social customs, work and home addresses, etc. Trajectory privacy-preserving techniques aim at protecting the sensitive information not to be revealed. Trajectory k-anonymity a classical technique in this area, it is proposed to anonymize k trajectories together in a broader similar time span [1–3].

But we argue that, it is not necessary to involve all location samples into the privacy strategy, since it may cause extra information loss. While, we have a key observation that, real trajectories are not randomly sampled spatio-temporal points, they have semantics, such as a stay in a semantic place, which denote where the moving object really visited.Suppose an adversary called Mr.J analyzes the published trajectories which are anonymized by replacing the identifiers with random and unique pseudonyms. Through his analysis with a reverse geo-coder and yellow pages, he found that a person in the data set lives in Jimen Li in BEIJING and works in Cuigong Hotel, also this guy traveled to Lijiang on Labor Day vacation of 2007, and his hometown is in HeNan province. The above features helped Mr.J to recognize this guy as Mr.Q by linkage with a published population data, consequently, several sensitive places where Mr.Q has visited were discovered by Mr.J, such as clubs, hospitals etc. Mr.Q's privacy is exposed through significant stays in his everyday traces.

To address the above problems, we propose a method called Walk Alone and Be Fast (WABF)to protect trajectory privacy against semantic location attack and maximum moving speed attack.

This research was partially supported by the grant from the Natural Science Foundation of Hebei Province (No. F2015207009) and Scientific research project of Hebei higher education institutions(No. BJ2016019).

II. RELATED WORKS

Trajectory privacy-preserving techniques are mainly in three categories: perturbation-based method, suppression-based method and generalization-based method. □

The main idea of perturbation-based methods is to add noises to the original data. In [7], authors propose to generate dummy trajectories to perturb the original trajectories. In order to confound fake trajectories with true ones, dummy trajectories are generated by rotating real users' trajectories. Differential privacy [8] is a new privacy strategy which adds Laplace noise to the input data to confuse the outputs. In [9], authors propose a differentially private trajectory data publication method, which publish trajectories in the form of prefix tree, each node of the tree represent a location sample of a trajectory. Laplace noise is added to the count value of each node. Authors also design a consistent processing method to improve the utility of the published data. □

Suppression-based methods try to suppress location samples if exposure of them may cause privacy leakage. Study in [4] is based on the assumption that different ad- versaries may have different and disjoint parts of users' trajectories. Trajectory pieces should be suppressed when publication of these pieces may increase the whole trajec- tory's*breach probability*. In [10], a suppression-based method is proposed to protect users' online trajectory privacy. Areas are classified as either *sensitive* or *insensitive* based on the proportion of visitors and the population of that area. Location updates are suppressed when users enter a sensitive area.

Generalization-based methods try to generalize location samples on trajectories into areas, which can protect trajectories not to be re-identified. In [2], Abul et al. propose a

novel concept named (k,δ)-anonymity due to the imprecision of GPS data, where δ represents the possible location imprecision. Based on the concept of (k,δ)- anonymity, authors propose an approach called *Never Walk Alone* (NWA) to achieve (k,δ)-anonymity through trajectory clustering and space translation. In [11], authors refine (k,δ)-anonymity model and propose a model called (k, @d)-anonymity, which is based on location co-appearance. Instead of Euclidean distance, (k, @d)-anonymity uses EDR distance in trajectory clustering.

III. ATTACK MODEL

We first elaborate some assumptions for the attacker, explain the attack models we study in this paper, then we define the privacy model.

Definition 1 (**Knowledge of the Attacker**). Any party that owning the following information can be a potential attacker: (1) published trajectory data; (2) distribution of real-world places; (3) maximum moving speed of the moving object.

Published trajectories may be attacked in the following ways: firstly, significant stays on the trajectories are discovered; secondly, a whole trajectory may be re-identified by linkage of significant stays with background knowledge. e.g. by knowing a trajectory's several stays as background information (such as check-ins in Geo-social networks), adversaries may pick up a few trajectories which satisfy this demands, in the extreme case, there is only one trajectory, thus this trajectory is fully re-identified. This is a semantic location attack.

Definition 2 (**Maximum Moving Boundary**). Given an anonymity zone $Z_i$which is generalized by l real-world places, moving objects' maximum moving boundary of Z_i at time t_i is a round rectangle that extends Z_i by a radius of vmax($t_i - t_{i-1}$), denoted by MMB(t_i), where v_{max} is the maximum moving speed of the moving object.

Fig.1(a) shows moving objects' maximum moving boundary of Z_i, Fig.1(b) shows what is the maximum moving speed attack[4]. If adversaries know the anonymity zone at time t_i and the maximum moving speed v_{max}, they know the maximum moving boundary shown in Fig.1(b) as a round rectangle MMB(t_i). We know that, if MMB(t_i) intersects with Z_{i+1} at t_{i+1}, the moving object must be in the intersection area, the same happens when MMB(t_{i+1}) intersects with Z_i. Although each zone contains at least l real-world places, if there are several real-world places in the non-intersection area (The number of places in this area is definitely less than l), the privacy guarantee is less than 1/l. This is called maximum moving speed attack.

Definition 3 (**Privacy Model**). Given the maximum moving speed v_{max}, stay point $L_{spi}$is generalized to an anonymity zone Z_i, the intersection area between $Z_i$and MMB(t_{i+1}) should cover at least l real-world places, where l is a privacy parameter specified by users; pass-by points which are covered by $Z_i$are suppressed.

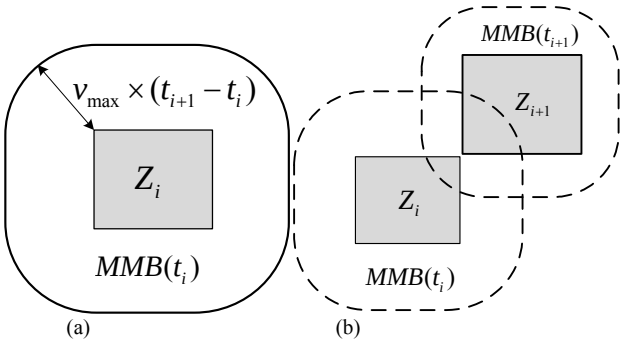

Fig.1 Maximum Moving Boundary and Maximum Moving Speed Attack

IV. ALGORITHM

Before our method, we assume the traces are already anonymized by replacing the true identifier with a random and unique pseudonym. Our goal in this paper is to anonymize original trajectory database D to a published version D* which satisfies conditions defined in the privacy model. The procedure of *WABF* is as follows.

(a). Splitmapgeneration.Thispartisthekeycomponentofthealgorithm whichisthe guidance of the trajectory anonymization. First, we extract stay points from raw trajectories, then reconstruct semantic places using a *reverse geocoder*. After that, we construct anonymity zones containing l places through a grid-based and a clustering-based method respectively. The generated zones can also withstand the maximum moving speed attack.

(b). Trajectory anonymization. We divide trajectories into {move, stay} sequences, where stay points are replaced by corresponding zones. Then we check whether each zone can withstand the maximum moving speed attack, if not, we extend the zones. Pass-by points are either suppressed or un-processed, depending on whether it locates inside a zone or not. At last, D is transformed to D*.

(c). Information loss measure. We measure information loss of D*in this step. Since D*is always published for analyzing purpose, the utility of D*should be kept high. Here we adopt an information loss measure in [3], which is represented as the reduction of the probability with which people can accurately determine the position of a MOB.

Since the procedure above is illustrated in paper [5], we mainly focus on how to extend our method to protect trajectory privacy against maximum moving speed attack.

Trajectories are split and anonymized based on the split map, where an anonymity zone replaces stay points, and pass-by points are ignored. As we have explained previously, the intersection between the maximum moving boundary and the anonymity zones may cause privacy leakage. In order to avoid this attack, the maximum moving boundary of two consequent timestamps should cover the corresponding anonymity zones,

as explained in Fig.2. The maximum moving boundary MMB(t_{i+1}) should cover the anonymity zone Z_i, at the same time, the maximum moving boundary MMB(t_i) should cover the anonymity zone Z_{i+1}. if not, we should extend Z_{i+1} (Z_i) to make MMB(t_{i+1}) (MMB(t_i)) lager enough. In order to make the extended area size as small as possible, we adopt a concept called MaxMin distance, which is defined in the following.

Definition 4 (**MaxMin Distance**). Let Z_i and Z_j be two generated zones. The MaxMin distancefrom Z_i to $Z_j$is defined as:

$$MaxMinDist(Z_i, Z_j) = \max_{p \in Z_i} \min_{q \in Z_j} dis(p, q) \quad (1)$$

MaxMinDist(Z_i, Z_j) implies the maximum distance between a point $p \in Z_i$ and its closest point $q \in Z_j$ [4]. MaxMin distance is unsymmetrical, that is to say, MaxMinDist (Z_i, Z_j) $\neq$ MaxMinDist(Z_j, Z_i). If each anonymity zone of two consequent timestamps is fully inside the corresponding maximum moving boundary, the distance between zone Z_i and Z_j should satisfy: MaxMinDist(R_i, R_j) $\leqslant v_{max} \times (t_j - t_i)$ and MaxMinDist(R_j, R_i) $\leqslant v_{max} \times (t_j - t_i)$.

We have explained how to generate a split map in [5] and how to resist maximum moving speed attack, then we present how to anonymize trajectories with split map, as shown in Algorithm 1.

Input : D_{zones}, original trajectory dataset D
Output: D^*
1 Scan each location sample on trajectory T_i;
2 **for** *each stay point* L_{spi} **do**
3 $Z_i \leftarrow L'_{spi}$ corresponding zone;
4 $MaxMinDist(Z_i, Z_{i+1}) \leftarrow$ MaxMin distance between Z_i and Z_{i+1};
5 **if** $MaxMinDist(Z_i, Z_{i+1}) < v_{max} \times (t_{i+1} - t_i)$ **then**
6 $Z_i \leftarrow$enlarge the area size of Z_i;
7 **end**
8 **else**
9 Z_i is an anonymity zone;
10 **end**
11 **end**
12 **for** *each pass-by point* L_i **do**
13 **if** L_i *is covered by* Z_i **then**
14 suppress L_i;
15 **end**
16 **end**
17 **end**

Algorithm 1 TrajectoryAnoymization (D_{zones}, D)

The original trajectory database D is set as input, each location sample is scanned, stay points are replaced by the corresponding anonymity zones. The zones should be retreated, since the generated zones may not resist the maximum moving speed attack. We should examine each place in the zone to check whether it is fully covered by its consequent time-stamps' maximum moving boundary. If not, the zone should be extended and the extended size should be as small as possible to reduce the information loss(line 2-8).

For each pass-by point, the published version is kept as the original one, unless the pass-by point is covered by a zone. Cover is a spatial relationship between a zone and a pass-by point of the same trajectory. If a pass-by point L_j is covered by a zone, L_j is suppressed for privacy preservation purpose, since publication of location samples approaching to a zone may cause exposure of a stay point (line 9-11). At last, stay points in D is replaced by its anonymity zones, while pass-by points are either suppressed or ignored, depending on whether it is covered by anonymity zones. The published version D* contains no sensitive information taking by stay points.

V. Privacy Analysis

We formally show that by applying our methods, given the maximum moving speed, the published database D* will not expose any user's stay points during their travels. Privacy guarantee is always measured by re-identification probability that means the probability of adversaries to identify a stay point or a trajectory from the published database D*.

Theorem 1. Given a trajectory database D={T_1, T_2, ...T_n} and its published version D*={T_1*, T_2*, ...T_n*} generated by WABF, the average stay points re-identification probability is bounded by 1/*l*.

Proof. Adversaries have access to all the published trajectories, public knowledge and MOB's maximum moving speed. Adversaries know the distribution of places on the map. Given a published version D*, each stay point in D* is generalized to an area which contains at least *l*places. The re-identification probability depends on the number of places in an anonymity zone, which is bounded by 1/*l*.

To capture the information loss, we adopt the reduction in the probability with which people can accurately determine the position of an object in [3]. Given a published database D* of D, the average information loss is defined in the following equation:

$$IL_{avg} = \frac{\sum_{i=1}^{n} \sum_{j=1}^{j} (1 - 1/area(zone(O_i, t_j))) + \sum_{d=1}^{h} L_d}{n \times m} \quad (2)$$

IL_{avg}representstheaverageshrinksoftheidentifyprobabilityofalocationin D^*. Where area(zone(O_i, t_j)) represents the area size of the anonymity zone of $O_i$at time $t_j$when $O_i$stays. The probability of adversaries can accurately determine the location where the MOB stays shrinks from 1 to 1/area(zone(O_i, t_j)). If a location $L_d$is deleted, it is totally indistinguishable, so the information loss turns to be 1. n×m represents the total location samples in D. Obviously, IL_{avg}ranges from 0 to 1.

VI. EXPERIMENTS

We run our experiments on a real-world dataset. Thanks to the Geolife project [6], we get the published real trajectory data. The dataset contains more than 8000 trajectories of 155 users ranging from May 2007 to May 2010 mainly in Beijing. More than 23 million GPS records are collected. The dataset is represented as BEIJING henceforth. The experiments are run on an Intel Core 2 Quad 2.66HZ, windows 7 machine equipped with 4GB main memory.

We run a set of experiments on BEIJING to evaluate the performance of WABF, under different α and l value, as shown in Fig.2.

As shown in Fig.2(a), the information loss increased with the increasing of *l*, that is because with the increase of *l*, the area size of the anonymity zone is getting large, resulting in larger information loss. We can also see that, there is no obvious trend of the information loss on different α value, this is because α is a balancing parameter between Euclidean distance and semantic distance. For different real-world places, α value may act different.

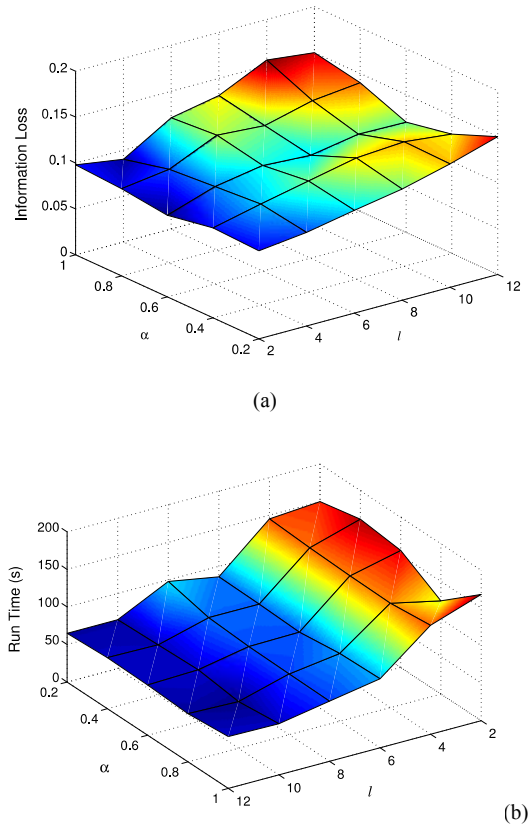

Fig.3. Information Loss of WABF

We then evaluate the run time of our approach. Fig.2(b) shows run time of WABF on different α and *l* value. With the increase of *l* value, run time of WABF decreases. In this experiment, we exclude the time consumption of stay points extraction and place reconstruction, because Google Maps API contains restrictions, it is only allowed to reverse one coordinate every 2 seconds.

VII. CONCLUSIONS

This paper proposes an approach called Walk Alone and Be Fast (WABF) to protect trajectory privacy against semantic location attack and maximum moving speed attack. WABF reduces the whole trajectories' exposure probability. At last, we conduct a set of comparative experimental studies on a real-world data set, the results show that WABF is effective and the information loss is much lower than k-anonymity methods.

REFERENCES

[1] Nergiz, M.E., Atzori, M., Saygin, Y., Baris, G.: Towards Trajectory Anonymization: A Generalization-based Approach. IEEE Transactions on Data Privacy.2, 47-75, 2009.

[2] Abul, O., Bonchi, F., Nanni, M.: Never Walk Alone: Uncertainty for Anonymity in Moving Objects Databases. In: 24th IEEE International Conference on Data Engineering, pp. 215-226. IEEE Press, Washington, 2008.

[3] Yarovoy, R., Bonchi, F., Lakshmanan, S., Wang, W.H.: Anonymizing Moving Objects: How to Hide a MOB in a Crowd? In: 12th International Conference on Extending Database Technology, pp. 72-83. ACM Press, New York, 2009.

[4] Pan, X., Xu, J. and Meng, X. Protecting Location Privacy against Location-Dependent Attacks in Mobile Services.IEEE Transaction on Knowledge and Data Engineering (TKDE).24(8): 1506-1519, 2012.

[5] Huo Z., Meng X., Hu H., Huang Y. You Can WalkAlone: Trajectory Privacy-preservingthroughSignificantStaysProtection. In: 17th International Conference on Database Systems forAdvancedApplications, pp. 351-366, April 15-18, 2012.

[6] Microsoft Research Geolife, http://research.microsoft.com/en-us/projects/geolife/

[7] Tou, T., Peng, W. and Lee, W. Protecting Moving Trajectories with Dummies. In: 2007 International Conference on Mobile Data Management, pp. 278-282. IEEE Press, Washington (2007)

[8] Dwork, C. Differential Privacy. In: 33rd International Colloquium on Automata, Languages and Programming, pp. 1-12. (2006)

[9] Chen, R., Fung, B.C., Desai B.C. and Sossou, N.M. Differentially Private Trasit Data Publication: A Case Study on the Montreal Transportation System. In: 18th ACM SIGKDD Conference on KnowledgeDiscovery and Data Mining, pp. 213-221. ACM Press, New York (2012)

[10] Gruteser, M., Liu, X.: Protecting Privacy in Continuous Location-tracking Applications. IEEE Security and Privacy. 2(2), 28-34 (2004)

[11]Abul, O., Bonchi, F., and Nanni, M. Anonymization of Moving Objects Databses by Clustering and Perturbation. Information Science. Vol. 35(8):884-910,2010.