Beyond K-Anonymity: Protect Your Trajectory from Semantic Attack

Zhen Tu, Kai Zhao, Fengli Xu, Yong Li, Li Su, Depeng Jin
Tsinghua National Laboratory for Information Science and Technology (TNLIST),
Department of Electronic Engineering, Tsinghua University, Beijing 100084, China
E-mail: liyong07@tsinghua.edu.cn

***Abstract*—Nowadays, human trajectories are widely collected and utilized for scientific research and business purpose. However, publishing trajectory data without proper handling might cause severe privacy leakage. A large body of works are dedicated to merging one's trajectory with others', so as to avoid any individual trajectory being re-identified. Yet their solutions do not provide enough protection since they cannot prevent semantic attack, which means the attackers are able to acquire individual's private information by using the semantics features of frequently-visited locations in the trajectory even without re-identification. In this work, we are the first to recognize the semantic attack, which is another severe privacy problem in publishing trajectory datasets. We propose an algorithm providing strong privacy protection against both the semantic and re-identification attack while reserving high data utility. Extensive evaluations based on two real-world datasets demonstrate that our solution improves the quality of privacy protection by 3 times, sacrificing only 36% and 30% of spatial and temporal resolution, respectively.**

I. Introduction

With the increasing prevalence of mobile devices, trajectories are widely collected in real-world by various location-based technologies, which benefits many areas of networking optimization [1, 2], urban computing [3], intelligent transportation [4], etc. Publishing trajectory dataset becomes a trend. However, trajectory exhibits individual's mobility, which contains plenty of private information such as home and working locations. Many studies have investigated $re\text{-}identification\ attack$ to prevent individuals being re-identified when their trajectories are published. For example, Montjoye *et al.* [5] investigate the privacy bounds of human mobility and find that four spatio-temporal points are enough to uniquely re-identify 95% of individuals in a fine-grained trajectory dataset. Zang *et al.* [6] reveal that anonymous location data severely leaks user's private information, *i.e.*, top 3 locations of a trajectory are enough to re-identify more than 80% individuals. Consequently, re-identification attack is a severe and well-recognized privacy threat in trajectory dataset.

Preventing individuals from being re-identified from the published trajectory data is a widespread topic in the last few years. Recently proposed algorithms NWA [7], GLOVE [8] and W4M [9] are all based on k-anonymity [10] which requires that each individual cannot be distinguishable from at least other $k-1$ individuals. Techniques of generalization, perturbation and suppression are utilized to achieve such indistinguishability. However, k-anonymity does not provide enough protection for trajectory data publishing. We introduce $semantic\ attack$ into the trajectory data, which indicates that the attacker is able to infer semantic information of locations in the trajectory when accessing the published dataset. Semantic information of a trajectory helps the attacker to know individual's behavior, which causes privacy disclosure. $Point\ of\ Interest$ (PoI), a feature on a map that occupies a particular point, reveals most of the semantic information of a location. A trajectory is formed by a series of locations and each location contains several PoIs. Some locations have limited categories of PoIs or their distributions differ from the whole city. In these cases, the attacker can enlarge his knowledge about the main functions of these locations and further infer the individual's motivation of visiting that location. For example, if an individual frequently visits a location that has most PoIs of hospital, the attacker infers that he may have health issues. Indeed, the PoI distributions of specific locations and overall city are quite different, which indicates that there is a huge vulnerability to be attacked.

The objective of our work is not only preventing individuals from being re-identified but more importantly protecting semantic information in the publishing of trajectory data. In order to achieve this goal, the published data needs to satisfy following three requirements. First, apparently it should be able to prevent re-identification attack, *i.e.*, the attacker cannot re-identify any individual in the dataset. Second, the published data should be able to prevent semantic attack, *i.e.*, the PoI distribution of each location in the trajectory should be as much close to that of the overall city as possible. Third, the truthfulness of data should be kept at record level, *i.e.*, only generalization and suppression rather than perturbation can be utilized to process the trajectories. The contribution of this paper can be summarized as the following three-folds:

- We are the first to introduce semantic attack to trajectory data to the best of our knowledge, which is a severe privacy problem in publishing trajectory data. We formally define the semantic attack protection of trajectory data as an optimization problem.
- We propose an algorithm to generalize trajectories to against both semantic attack and re-identification attack, which meets the requirements of k-anonymity, l-diversity and t-closeness at the same time. To the best of our knowledge, this is the first algorithm for protecting tra-

978-1-5090-6599-8/17/$31.00 ©2017 IEEE

jectory privacy by taking all these criterions into consideration.

- We evaluate our algorithm based on two real-world mobility datasets that are collected from cellular network as well as mobile device. The result demonstrates that our algorithm provides strong privacy protection against not only re-identification attack but also semantic attack, meanwhile the utility of data is well reserved, *i.e.*, it improves the quality of privacy protection by 3 times, sacrificing only 36% and 10% of spatial and temporal resolution decrease respectively.

The rest of the paper is organized as follows. Section II describes the dataset and introduces the attack and privacy model. Section III formalizes the problem and Section IV describes our anonymization algorithm for trajectory data. Section V presents the evaluation results. Section VI reviews the related work and finally Section VII concludes the paper.

II. Preliminaries

A. Dataset

In this work we utilize two types of data. The first one is $mobility\ data$, including two datasets collected both from the cellular network and mobile application. The key features of these two mobility datasets are summarized in Table I. The second type is $location\ semantic\ data$, which includes PoI distribution of the same city. Now we introduce the detailed information of these datasets.

1) Mobility Data:

- $Cellular\ Dataset$: This dataset is collected by a major mobile service provider in Shanghai, one of the major metropolitan in China. It contains complete trajectories of over 5,900,000 mobile users with a duration of one week, between April 1st and 7th, 2016. When the user accesses cellular network (i.e., making phone calls, sending texts, or consuming data plan), the accessed base stations and timestamps are recorded.
- $Application\ Dataset$: This dataset is collected from devices by a popular mobile application, which traces over 15,000 mobile users in Shanghai for two weeks, from November 1st to 14th, 2015. It records the timestamp and the mobile user's accessed base station when it is activated for service interactions.

For both datasets, the spatial granularity is related to the size of base stations. Therefore, by looking up the locations of these base stations, we are able to observe the trajectories of mobile users. As for privacy concern, users' identifications are anonymized.

2) Location Semantic Data:

- $PoI\ Dataset$: We crawl 0.82 million PoIs of Shanghai city from map service. Each PoI is a location with a specific label, *i.e.*, food and school. According to the functions, we further divide them into six categories, *i.e.*, entertainment, education, scenery spot, business, industry, residence. The detailed information is shown in Table II. PoIs often fall inside regions where people perform socioeconomic activities, *i.e.*, staying home, working and going to hospital, whose distribution reveals semantic information.

Datasets & Metrics	Cellular Dataset	Application Dataset
Source	Cellular network	Mobile application
Location	Shanghai, China	Shanghai, China
Time	Apr. 2016	Nov. 2015
Duration	one week	two weeks
User number	5.90 millions	15.50 thousands
Record number	1.54 billions	7.69 millions
Records/user	261	496

TABLE I: Major information and key features of mobility datasets.

Categories	Abbr.	# of PoI	Detailed information
Entertainment	Ent.	275366	food, hotel, gym, shopping, leisure.
Education	Edu.	8417	school, campus.
Scenery spot	Sce.	5735	scenery spot.
Business	Bus.	255352	finance, office building, company, *etc.*
Industry	Ind.	10880	factory, industrial estate, *etc.*
Residence	Res.	268305	residence, life services.

TABLE II: The detailed information of PoI dataset.

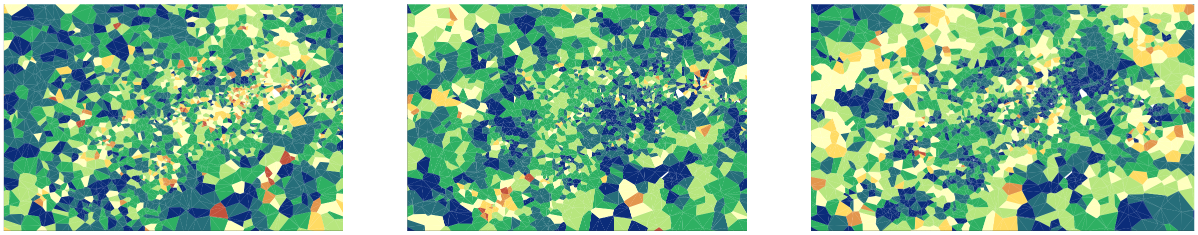
B. Attack Model

The attack model concerns what privacy information the attacker attempts to acquire about the victim from the dataset with some background knowledge. Thus, key elements are attacker's background knowledge and the interested privacy information. Trajectory is a sequence of spatio-temporal points and usually the background knowledge is some spatio-temporal points of the victim's trajectory. In addition, the attacker also has some open data such as the PoI distribution of the whole city. In this work, we tackle the following two attacks.

$Re\text{-}identification\ attack$: the attacker aims to identify individual in the dataset with some spatio-temporal points of the individual to be attacked. Matching these spatio-temporal points with the trajectories, the attacker might successfully re-identify the individual.

$Semantic\ attack$: the attacker aims to acquire individual's motivation at some spatio-temporal points in the trajectory by using location semantic information. For some locations in the trajectory, comparing its PoI distribution with the whole city, the attacker can obtain knowledge about the location's semantic information, *i.e.*, its main function of socioeconomic activities. Then the individual's motivation of visiting these locations can be inferred, which causes privacy disclosure.

Two scenarios of semantic attack are provided in Fig. 2. First, Fig. 2(a) shows two PoI distributions of the overall city and location A. Since location A only has one category of PoI $-$ business, the attacker can easily infer that this may be his working place if an individual frequently visits it during workdays. Such a privacy disclosure is caused by the unique of the PoI category. As a second example, Fig. 2(b) shows 6 categories of PoI all appear in location B. However, compared



(a) Business PoI distribution (b) Entertainment PoI distribution (c) Residence PoI distribution

Fig. 1: Geographical distributions of three categories of PoIs.

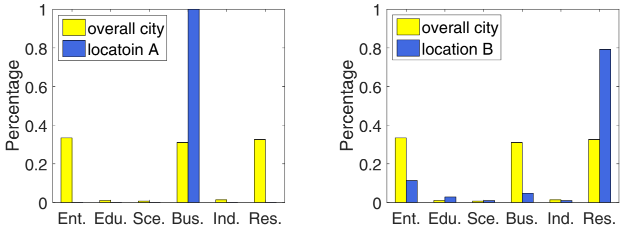


Fig. 2: Examples of semantic attack by comparing location PoI distribution with the whole city's distribution, where A and B are two locations in the user's trajectory.

with the PoI distribution of the whole city, the attacker can infer that this location is very likely to be a residential location. In conclusion, by analyzing location semantic information, the attacker is able to infer the individual's behavior at each timestamp of the trajectory.

Semantic attack can be easily applied to trajectories because the PoI distribution of location in each base station's coverage is quiet different from that of the overall city. Fig. 1 shows the geographical distribution of three kinds of PoIs in the downtown area of Shanghai. The color of each polygon refers to the proportion of that category of PoI under this base station. We find that their distributions are different from each other geographically, which indicates locations covered by different base stations have different urban functions.

C. Privacy Model

Our privacy model is consistent with the objective that publishing truthful data against both re-identification and semantic attack by satisfying criterions of k-anonymity [10], l-diversity [11] and t-closeness [12].

First, k-anonymity ensures the attacker cannot distinguish the victim from at least $k-1$ other individuals, which is utilized to against re-identification attack. Since we do not limit the attacker's knowledge about individual's trajectory, the victim's trajectory should be indistinguishable from at least $k-1$ other trajectories, which means these trajectories should be the same after generalization.

Second, in order to prevent from semantic attack, trajectories to be published should satisfy privacy criterions of l-diversity and t-closeness. A set containing indistinguishable trajectories from k individuals is called k-anonymous set. l-diversity requires that the sensitive attribute of a k-anonymous set contains at least l well-represented values for the sensitive attribute. A k-anonymous set is said to be t-closeness if the distance between the distribution of a sensitive attribute and that of the whole dataset is less than a threshold t. We formally introduce these two criterions into trajectory data. For l-diversity, we set the sensitive attribute as PoI, consequently the number of different PoI categories in a spatial position should be larger than l. On the other hand, satisfying t-closeness means the distance between the distribution of PoIs in a location and that of the whole city is within a threshold t.

Third, in order to maintain truthfulness of dataset, we only use spatio-temporal generalization and suppression to process the trajectory data. Spatial generalization is merging nearby base stations and temporal generalization is increasing temporal granularity to combine different trajectories into one. When merging some spatio-temporal points causes huge loss of spatio-temporal granularity, we just delete them, which is called suppression.

III. Problem Definition

In our privacy model, we adopt spatio-temporal generalization and suppression to make sure that every user's trajectory satisfies k-anonymity, l-diversity and t-closeness. Now we formally define how a trajectory meets above privacy requirements, and measure the loss of spatio-temporal resolution due to generalization.

Formally, we define the Mobility Dataset as $T = [T_1, T_2, ..., T_N]$ with T_i representing the original trajectory of user i and N representing the total number of users. Every user's trajectory contains continuous spatio-temporal points. For user i, the trajectory denotes $T_i = [p_1^i, p_2^i, ..., p_{s_i}^i]$, where p_j^i represents the j-th point of user i's trajectory and s_i represents the number of points in the trajectory. Each point records the user's temporal and spatial information, denoted by $p_j^i = (t_j^i, d_j^i, l_j^i)$, i.e., user i is at location l_j^i in time between t_j^i and $t_j^i + d_j^i$. As for the Location Semantic Data, we denote the PoI as $I = [I_1, I_2, ..., I_6]$, where I_i represents the subset of PoIs that belong to the i-th category. For the overall city denoted by R, we denote M_i as the number of PoIs in the i-th category. Then I_i can be denoted as $I_i = [I_1^i, I_2^i, ..., I_{M_i}^i]$. As every PoI has a specific geographical location, for a given region $r \in R$, we further obtain a PoI distribution vector,

$[m_1^r, m_2^r, ..., m_6^r]$, where m_i^r is the number of PoIs in the i-th category.

A. Privacy Formulation

In our privacy model, every user's trajectory needs to meet the requirements of k-anonymity, l-diversity and t-closeness after generalization. Now we discuss how to guarantee a trajectory satisfies these three privacy criterions.

k-anonymity: For trajectory dataset, the users that share the same full-length trajectory constitute an anonymous set. We define δ_k^i to represent the size of anonymous set A that user i belongs to, which means user i's trajectory T_i is indistinguishable from the other δ_k^i users. k-anonymity requires that $\delta_k^i \geqslant \delta_k$ with δ_k as the threshold.

l-diversity: Different from k-anonymity considered at the trajectory level, l-diversity is related to spatio-temporal points that belong to the trajectory. It requires that every spatio-temporal point has at least l categories of PoI. Given the j-th spatio-temporal point p_j^i of user i, $p_j^i = (t_j^i, d_j^i, l_j^i)$ and region $r = l_j^i$, the PoI distribution vector of region r is $[m_1^r, m_2^r, ..., m_6^r]$. We define δ_l^r as the number of different PoI categories that region r has, which is expressed as follows,

$$\delta_l^r = \sum_{u=1}^{6} O_u^r, \text{ where } O_u^r = \begin{cases} 1 & m_u^r \geqslant 1, \\ 0 & m_u^r = 0. \end{cases}$$

l-diversity requires that $\delta_l^r \geqslant \delta_l$ with δ_l as the threshold.

t-closeness: Similar to l-diversity, t-closeness is another criterion for spatio-temporal points in trajectory. It requires that each spatio-temporal point has a similar PoI distribution with the whole city, and the difference of these two distributions should be smaller than t. Again take p_j^i for example: the PoI distribution vector of region r is $[m_1^r, m_2^r, ..., m_6^r]$, and the PoI distribution vector of the overall city, denoted by R, is $[M_1, M_2, ..., M_6]$. We define X and Y as the PoI distribution of r and R respectively, expressed as follows,

$$X_u = \frac{m_u^r}{\sum_{h=1}^{6} m_h^r}, Y_u = \frac{M_u}{\sum_{h=1}^{6} M_h}.$$

We use **KL Divergence** [13] to measure the difference from Y to X, denoted by δ_t^r, which is defined as follows,

$$\delta_t^r = \sum_{u=1}^{6} X_u log \frac{X_u}{Y_u}.$$

Then t-closeness requires that $\delta_t^r \leqslant \delta_t$ with δ_t as the threshold.

B. Spatio-temporal Resolution Loss

We use spatial-temporal generalization to merge two trajectories, part of two users' trajectories (in blue and in pink) are shown in Fig. 3. We choose close spatio-temporal points from both trajectories and make them identical in temporal and spatial dimensions, which is called **merge**. For example, when merging two spatio-temporal points in the lower left corner, we generalize the time slot to 8am-10am, on the other hand we also add some extra base stations (in purple) to form a qualified region.

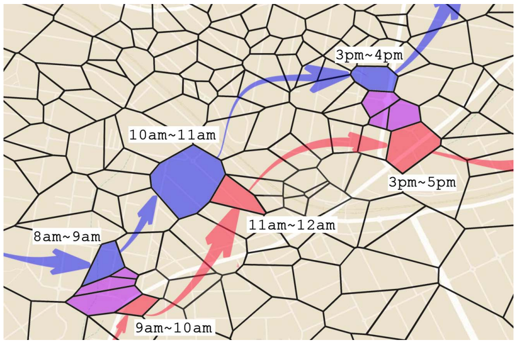


Fig. 3: Example of merging two trajectories.

Consider the case that we merge spatio-temporal point $p_a = (t_a, d_a, t_a)$ and $p_b = (t_b, d_b, t_b)$ from T_i and T_j respectively, forming a new merged point $p_c = (t_c, d_c, l_c)$. Obviously, time range $[t_c, t_c + d_c]$ should contain $[t_a, t_a + d_a]$ and $[t_b, t_b + d_b]$, location l_c needs to cover l_a and l_b. In addition, l_c needs to satisfy the privacy criterion of l-diversity and t-closeness.

We calculate the loss of temporal and spatial resolution separately. Specifically, for a spatio-temporal point $p = [t, d, l]$, we use time interval d to represent the temporal resolution and use the coverage of location l to represent the spatial resolution. Therefore, the temporal resolution loss, denoted by θ_T and spatial resolution loss, denoted by θ_L, due to generalization can be obtained as follows,

$$\begin{cases} \theta_T^*(t_a, d_a, t_b, d_b) = \dfrac{(d_c - d_a)n_i + (d_c - d_b)n_j}{n_i + n_j}, \\ \theta_T(t_a, d_a, t_b, d_b) = min\left\{\dfrac{\theta_T^*(t_a, d_a, t_b, d_b)}{\theta_T^m}, 1\right\}, \\ \theta_L^*(l_a, l_b) = \dfrac{(S_{l_c} - S_{l_a})n_i + (S_{l_c} - S_{l_b})n_j}{n_i + n_j}, \\ \theta_L(l_a, l_b) = min\left\{\dfrac{\theta_L^*(l_a, l_b)}{\theta_L^m}, 1\right\}, \end{cases}$$

where S_l is the spatial area of location l. When T_i and T_j are both original trajectories, $n_i = n_j = 1$, θ_T^* and θ_L^* are unnormalized spatial and temporal resolution loss. By setting two thresholds $\theta_T^m = 8\ hours, \theta_L^m = 25\ km^2$, we normalize θ_T, θ_L to $[0, 1]$. When T_i and T_j are not from two single users but two groups of users whose trajectories have already been made identical, n_i and n_i equal to the number of users of each group respectively.

Combining the reduction of temporal and spatial resolution, we obtain the spatio-temporal resolution loss after merging two points as follows,

$$\theta(p_a, p_b) = \omega_T \theta_T(t_a, d_a, t_b, d_b) + \omega_L \theta_L(l_a, l_b). \quad (1)$$

If we attach equal importance to spatial and temporal resolution loss, normalization factors $\omega_T = \omega_L = 0.5$.

Considering merging two trajectories T_i and T_j, we define the reduction of spatio-temporal resolution, which is the merge cost, expressed as follows,

$$C_{ij} = \begin{cases} \frac{1}{m_i}\sum_{a=1}^{m_i} \min\limits_{b=1,..,m_j} \theta(p_a^i, p_b^j) & s_i > s_j, \\ \frac{1}{m_j}\sum_{b=1}^{m_j} \min\limits_{a=1,..,m_i} \theta(p_a^i, p_b^j) & s_i \leqslant s_j, \end{cases}$$

where s_i and s_j represent the total number of points in trajectory of T_i and T_j respectively.

C. Problem Formulation

In order to minimize the loss of spatio-temporal resolution under the condition that all the privacy criterions are satisfied, we formalize the optimal trajectory generalization problem as follows,

$$\begin{aligned} \text{Minimize} \quad & \sum_{i=1}^{N}(\sum_{j=1}^{N} C_{ij}X_{ij}/\sum_{j=1}^{N} X_{ij}), \\ \text{subject to} \quad & \forall\ G_i \in G,\ \delta_k^i \geqslant \delta_k, \\ & \forall\ p_j^i \in G_i, r = l_j^i,\ \delta_l^r \geqslant \delta_l, \\ & \forall\ p_j^i \in G_i, r = l_j^i,\ \delta_t^r \leqslant \delta_t, \end{aligned}$$

where G represents the generalized trajectory dataset containing N users, C is the merge cost matrix with $C_{ij} = C_{ji}$ and $C_{ii} = \infty$, and X is a connection matrix that each element has a value of 0 or 1. If the generalized trajectories G_i and G_j of user i and j are indistinguishable, $X_{ij} = 1$; otherwise $X_{ij} = 0$. It is apparently that $X_{ij} = X_{ji}$ and $X_{ii} = 0$. More importantly, $\sum_{j=1}^{N} X_{ij} = h$, which means that user i shares the same G_i with the other h-1 other users.

IV. Algorithm

The formulated problem is NP-hard, whose computational complexity grows exponentially with the number of users in the dataset. We propose a heuristic algorithm to obtain approximate solution. Our algorithm grants k-anonymity, l-diversity, and t-closeness of trajectories through specific generalization, while ensuring the smallest loss of spatio-temporal granularity. By continuously merging trajectories from the original mobility dataset, we obtain a new generalized dataset consisting of all merged trajectories, which is able to against re-identification and semantic attack.

A. Merging Spatio-temporal Points

Merging two spatio-temporal points is the basic process unit of merging two trajectories. Given spatio-temporal point p_a and p_b, we merge them into a new point p_c. We design an algorithm to achieve the merge process with the pseudocode listed in Algorithm 1.

In terms of temporal dimension, when time slot $[t_a, t_a + d_a]$ and $[t_b, t_b + d_b]$ are not exactly the same, we need to consider three different cases including no overlap, part overlap or total overlap, as shown in Fig. 4. In order to include both time slots, denoted by $[t_a, t_a + d_a]$ and $[t_b, t_b + d_b]$, the new time slot is obtained as follows,

$$[t_c, t_c + d_c] = [min\left\{t_a, t_b\right\}, max\left\{t_a + d_a, t_b + d_b\right\}].$$

Algorithm 1: Merging Two Spatio-temporal Points

```
Input: Original Point $p_a$, $p_b$, Diversity Criterion $\delta_l$,
       Closeness Criterion $\delta_t$
Output: Merged New Point $p_c$
1  $t_c = min(t_a, t_b)$, $d_c = max(t_a + d_a, t_b + d_b) - t_c$
2  $l_{c1} \leftarrow getConnectedRegion(l_a, l_b)$
3  while $\delta_l^{c1} < \delta_l$ do
4      $X \leftarrow getNeighbors(l_{c1})$
5      for $i = 1 : size(X)$ do
6          $r = \{l_{c1}, X_i\}$, $B_i = \delta_l^r$
7      $i = argmax(B)$, $l_{c1} = \{l_{c1}, X_i\}$
8  $l_{c2} = l_{c1}$
9  while $\delta_t^{c2} > \delta_t$ do
10     $X \leftarrow getNeighbors(l_{c2})$
11     for $i = 1 : size(X)$ do
12         $r = \{l_{c2}, X_i\}$, $B_i = \delta_t^r$
13     $i = argmin(B)$, $l_{c2} = \{l_{c2}, X_i\}$
14 $l_c = l_{c2}$
```

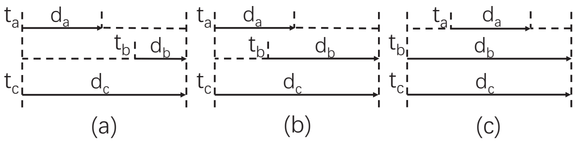
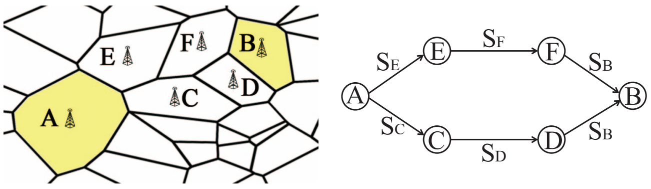


Fig. 4: Three different cases of merging time slot a and b into a new time slot c.

On the other hand, spatial generalization needs to satisfy the criterions of l-diversity and t-closeness, and it can be divided into three steps. The first step is to find a smallest connected region l_{c1} to cover l_a and l_b (Line 2). As l_a and l_b are both base stations, l_{c1} is a set of several base stations that include l_a and l_b. If l_a and l_b are the same, $l_{c1} = l_a = l_b$. If l_a and l_b are neighbors, $l_{c1} = \{l_a, l_b\}$. Otherwise, we need to add as fewer base stations as possible besides l_a and l_b to form a smallest connected region.

Fig. 5(a) shows the basic idea of how we find the connected region when merging two unconnected base stations. For example, when merging base station A and B, we need to add E and F, or C and D, or others to obtain a connected region. We transfer this problem into finding the shortest path in a directed graph, as shown in Fig. 5(b). Specifically, every base station can be regarded as a vertex and neighboring base stations have bidirected edges, while the weight of the edge is the area (like S_E) of the outgoing base station. Therefore, finding the smallest connected region covering A and B is equal to finding the shortest path from vertex A to B. The length of the path is the extra area increased from A to a smallest connected region. We utilize **Dijkstra Algorithm** [14] to find the shortest path and consequently obtain the connected region l_{c1}.



(a)

(b)

Fig. 5: Merging base station A and B to form a new connected region.

The second step is to find a region l_{c2} that includes l_{c1} and satisfies l-diversity criterion (Line 3-8). When l_{c1} already has δ_l or more categories of PoI, then $l_{c2} = l_{c1}$. Otherwise we need to add some extra neighboring base stations until l_{c2} meets the requirement. As shown in Line 3-7, at every iteration, we add a neighboring base station until l-diversity is satisfied.

The third step is to find the final region l_c that includes l_{c2} while satisfying t-closeness (Line 9-14). If δ_t^r has already been smaller than δ_t, $l_c = l_{c2}$. Otherwise, we need to add some extra neighboring base stations to make l_c satisfy the requirement. We adopt the same method as the second step to find the suitable neighboring base stations. The iteration process is shown in Line 9-13.

B. Merging Trajectories

As we have already considered l-diversity and t-closeness during the merging process of spatio-temporal points, we only need to satisfy k-anonymity criterion when merging trajectories. We design a greedy algorithm to achieve optimal merging with the pseudocode showing in Algorithm 2.

Algorithm 2: Merging Trajectories

```
Input: Original Dataset $T$, Merge Cost Matrix $C$,
       Anonymity Criterion $\delta_k$
Output: Generalized Dataset $G$
while $\exists\ T_i, T_j \in T, n_i < \delta_k, n_j < \delta_k$ do
    $i, j \leftarrow argmin(C)$
    $T_m \leftarrow merge(T_i, T_j)$, $n_m = n_i + n_j$
    $delete(T, T_i, T_j)$, $delete(C, i, j)$
    if $n_m < \delta_k$ then
        for $T_n \in T$ do
            $C_{mn} = calcCost(T_m, T_n)$
        $add(T, T_m)$
    else
        $add(G, T_m)$
```

The inputs of the algorithm are original dataset T, merge cost matrix C, i.e., containing the merge cost of all the trajectory pairs in T, and anonymity criterion δ_k. The algorithm will iterate until all the trajectories in T have been k-anonymized. At each iteration, we find the trajectory pair (T_i, T_j) with the smallest merge cost in C, merge them into a new n_m-anonymized trajectory T_m, and update both the datasets and merge cost matrix. We remove T_i, T_j from T and all merge cost related to T_i or T_j. If T_m has not been k-anonymized, we add T_m into T after calculating the merge cost between T_m and the rest trajectories in T one by one. Otherwise we add T_m into G. When all the original trajectories are merged into the k-anonymity trajectories in G, the merging process is finished and outputs the generalized dataset G.

In order to minimize the spatial-temporal resolution loss, we merge two trajectories step by step. Assume T_i is longer than T_j, we merge the first spatio-temporal point p_1^i of T_i with the smallest-merge-cost partner p_m^j in T_j and use the merged point to replace p_m^j. For the next spatio-temporal point p_2^i, we carry out the same operation until the last point $p_{s_i}^i$. Then we pick out the unmerged points of T_j and merge them into one of the rest merged points in T_j. At last, we carry out a reshaping operation to avoid time overlap between adjacent points.

After continuous iterations of merging two trajectories, finally we obtain the generalized trajectories that satisfy the criterions of k-anonymity, l-diversity and t-closeness.

C. Complexity analysis

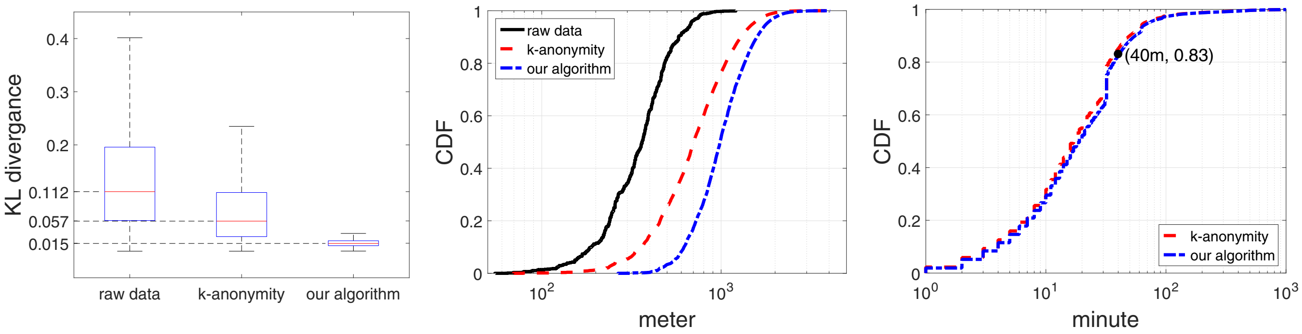
Assume the dataset has N users and the average number of spatio-temporal points of the trajectories is $\overline{m}$. The time complexity of calculating the merge cost matrix is $O(N^2\overline{m}^2M)$, where M is the total number of locations. Specifically, $O(M)$ is the time complexity of merging two spatio-temporal points. The time complexity of merging two trajectories and updating merge cost matrix is $O(\overline{m}^2M + N\overline{m}^2M)$, and the operation needs to be repeated $x = O(N)$ times, leading to $O(N^2\overline{m}^2M)$. Overall, our algorithm runs in polynomial time, which is effective in solving the optimal trajectory generalization problem.

V. Performance Evaluation

In this section, we evaluate the performance of our algorithm based on two real-world datasets. We filter active users that have relatively high probability to be attacked from the dataset. We also filter out trajectories that contain less than three different base stations because they are probably not collected from mobile devices carried by individuals.

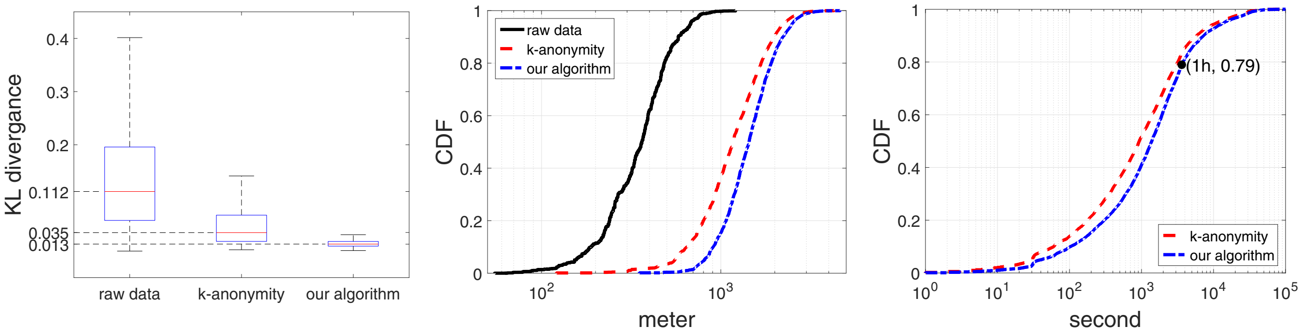
A. Comparison With Baseline

First, we evaluate the performance of our algorithm in the case that only satisfies k-anonymity, which is solved by GLOVE [8]. Our algorithm degenerates to this case if we do not add constrains on l and t (l=0, t=∞). Fig. 6(a) shows the KL divergence between PoI distributions of locations in trajectories and the whole city. It indicates that satisfying k-anonymity reduces the chances of being semantically attacked, as the median of KL divergence decreases from 0.112 of raw data to 0.057 of data satisfying k-anonymity. Fig. 6(b) and Fig. 6(c) respectively shows the spatial resolution and the temporal resolution of all the spatio-temporal samples in the dataset. By comparing the resolution of raw data and the data satisfying k-anonymity, we find that almost all samples reserve



(a) Privacy preserving (b) Spatial resolution (c) Temporal resolution

Fig. 6: Privacy preserving and spatio-temporal resolution of our algorithm compared with k-anonymity in the cellular dataset. For k-anonymity, k=2, l=0, and t=unconstrained. For our algorithm, k=2, l=6, and t=0.01.



(a) Privacy preserving (b) Spatial resolution (c) Temporal resolution

Fig. 7: Privacy preserving and spatio-temporal resolution of our algorithm compared with k-anonymity in the application dataset. For k-anonymity, k=2, l=0, and t=unconstrained. For our algorithm, k=2, l=6, and t=0.01.

a spatial resolution of 2 km and nearly 83% of samples reserve a temporal resolution of 40 minutes, which keeps high utility of data.

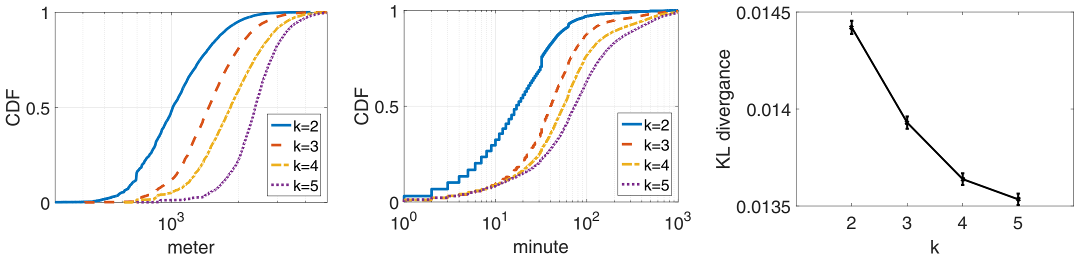
Second, we compare the performance of our algorithm with the case that only satisfies k-anonymity. We set $l = 6$, $t = 0.01$, providing strong privacy protection. In Fig. 6(a), the median of KL divergence decreases to 0.015, improving about 3 times than that of k-anonymity. Meanwhile, Fig. 6(b) shows that the average spatial resolution decreases from 776 meters to 1056 meters and the average deviation of these two CDFs is only 0.072. Fig. 6(c) shows that the average temporal decreases by 1.1, and the average deviation of two CDFs is only 4.18×10^{-4}. Thus, when l and t are strictly constrained, the spatial resolution decreases little and still reserves a spatial resolution lower than 2 km and the temporal resolution is almost unchanged.

Fig. 7 shows similar results on the application dataset compared with the raw data and the case that only satisfies k-anonymity. The average spatial resolution decreases by 0.2 times while the average temporal resolution also decreases by 0.2 times. Thus the spatio-temporal resolution loss is limited. However, we greatly improve the KL divergence from 0.035 to 0.013. These results indicate that our algorithm provides strict privacy protection to trajectory data publishing while reserving high data utility.

B. *Influence of Different Parameters*

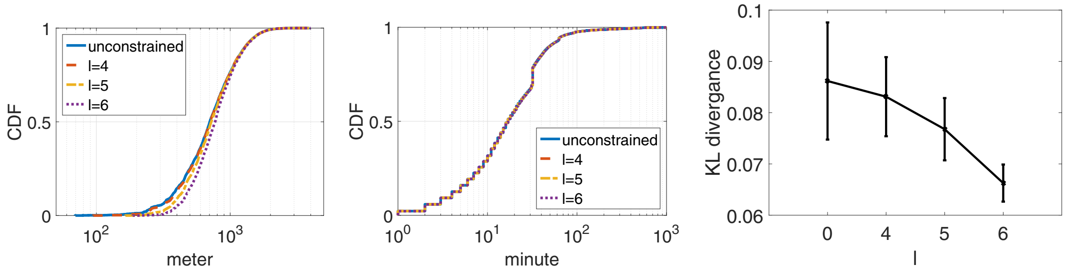
Influence of k. For k-anonymity, larger k guarantees higher privacy level, and we discuss how much resolution loss is required when improving the privacy level. We set $l = 6$, $t = 0.01$ and vary k from 2 to 5. Fig. 8(a) and (b) show the spatio-temporal resolution of trajectories. The average deviation between CDFs of $k = 2$ and $k = 5$ in spatial resolution is 0.34, and that of the temporal resolution is only 0.022. As k grows larger, the spatio-temporal resolution decreases but it is still within tolerable range. Thus, with little resolution loss, we protect trajectory data from re-identification attack. Fig. 8(c) shows the KL divergence with the error bar. It only decreases by less than 0.001 when k increases from 2 to 5. The reason is that increasing k dose not have a direct impact on the PoI distribution.

Influence of l. l is one of the major parameters that influence semantic attack, and now we investigate its impact on privacy preserving and spatio-temporal resolution. We set $k = 2$, do not constrain t and vary l from 4 to 6. As Fig. 9 shows, the spatio-temporal resolution decreases very slow and the KL divergence decreases by 0.01. The main reason is that many original base stations already have several different categories of PoI. When merging several base stations into a region, it can easily obtain all kinds of PoI. Therefore, larger l only guarantees the PoI diversity, but will not lead to smaller



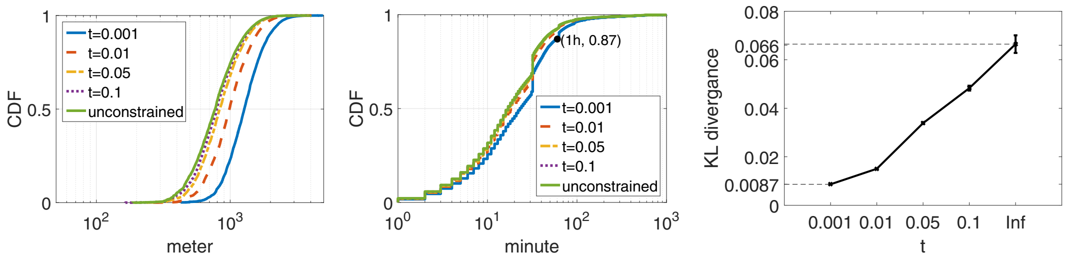
(a) Spatial resolution (b) Temporal resolution (c) Privacy preserving

Fig. 8: The influence of k on spatio-temporal resolution and privacy preserving (l=6, t=0.01).



(a) Spatial resolution (b) Temporal resolution (c) Privacy preserving

Fig. 9: The influence of l on spatio-temporal resolution and privacy preserving (k=2, t=unconstrained).



(a) Spatial resolution (b) Temporal resolution (c) Privacy preserving

Fig. 10: The influence of t on spatio-temporal resolution and privacy preserving (k=2, l=6).

KL divergence or larger loss of spatio-temporal resolution.

Influence of t. t is another parameter influencing semantic attack. We set $k = 2$, $l = 6$, and vary t from 0.1 to 0.001. From Fig. 10(c), the average of KL divergence decreases from 0.066 to 0.0087, which is a great improvement in privacy. However, Fig. 10(a) and (b) show that when gradually decreasing t, the spatio-temporal resolution loss increases slowly. All the samples have a spatial resolution lower than 3 km and nearly 87% samples have a temporal resolution lower than 1 hour. t is the key parameter that influences semantic attack, and smaller t can provide stronger privacy protection.

In conclusion, our algorithm reserves high data utility even when $k = 5$, which is able to resist re-identification attack forcefully. The parameter l is relatively not helpful against semantic attack while t can provide strong privacy protection when it's set with small value, *i.e.*, 0.001.

VI. Related Work

In traditional data publishing, there are numerous studies about k-anonymity [10], l-diversity [11] and t-closeness [12]. k-anonymity is a fundamental criterion of privacy against re-identification attack, but fails to provide protection against probabilistic attack, which motivates l-diversity and t-closeness. These privacy models achieve good performance in relational data [11, 12], where sensitive attribute is well-defined. In trajectory data, sensitive attribute is hard to define. In our work, we introduce PoI as the sensitive attribute, whose distribution reveals semantic information about the trajectory.

Other studies focus on the location privacy for location-based queries in mobile information delivery systems. The concept of k-anonymity was introduced in [15], and location cloaking algorithm for k-anonymity and l-diversity was proposed by [16]. Sematic information was considered in [17], while [18] proposed a mitigation technique to protect the location semantics from adversaries. Instead of protecting location privacy, our work concerns about trajectory privacy leakage. We protect mobility trajectories from both re-identification and sematic attack, by applying much stricter privacy criterions of k-anonymity, l-diversity and t-closeness.

In trajectory data, existing algorithms mainly focus on re-identification attack [7, 8, 9] and trajectory recovery attack from aggregated mobility data [19]. Our work is motivated by GLOVE [8], which applies k-anonymity to full-length trajectory anonymization. Our work differs from GLOVE mainly in two aspects. First, apart from re-identification attack, we take semantic attack into account. Second, instead of simply merging regular grids [8], we propose a novel spatio-temporal merging method based on irregular polygons that are more suitable for the scenario of cellular mobility.

VII. Conclusion

In this paper, we recognize and formally define the semantic attack of trajectory dataset. More importantly, we propose a novel algorithm to protect trajectories from both re-identification and semantic attack. The proposed algorithm is evaluated on two real-world trajectory datasets, and the results demonstrate that our algorithm provides desirable privacy guarantees while causing little data utility loss. We believe that this paper opens a new angle in trajectory privacy preserving.

Acknowledge

This work is supported by National Basic Research Program of China (973 Program) (No. 2013CB329105), National Nature Science Foundation of China (No. 61301080, No. 91338203, No. 91338102, and No. 61321061), and Research Fund of Tsinghua University-Tencent Joint Laboratory for Internet Innovation Technology.

References

[1] Y. Zheng, Y. Chen, X. Xie, and W.-Y. Ma, "Geolife2. 0: a location-based social networking service," in *2009 Tenth International Conference on Mobile Data Management: Systems, Services and Middleware*. IEEE, 2009, pp. 357–358.

[2] N. Eagle, A. S. Pentland, and D. Lazer, "Inferring friendship network structure by using mobile phone data," *Proceedings of the national academy of sciences*, vol. 106, no. 36, pp. 15 274–15 278, 2009.

[3] F. Xu, P. Zhang, and Y. Li, "Context-aware real-time population estimation for metropolis," in *Proceedings of the 2016 ACM International Joint Conference on Pervasive and Ubiquitous Computing*. ACM, 2016, pp. 1064–1075.

[4] L. Chen, A. Mislove, and C. Wilson, "Peeking beneath the hood of uber," in *Proceedings of the 2015 ACM Conference on Internet Measurement Conference*. ACM, 2015, pp. 495–508.

[5] Y.-A. De Montjoye, C. A. Hidalgo, M. Verleysen, and V. D. Blondel, "Unique in the crowd: The privacy bounds of human mobility," *Scientific reports*, vol. 3, 2013.

[6] H. Zang and J. Bolot, "Anonymization of location data does not work: A large-scale measurement study," in *Proceedings of the 17th annual international conference on Mobile computing and networking*. ACM, 2011, pp. 145–156.

[7] O. Abul, F. Bonchi, and M. Nanni, "Never walk alone: Uncertainty for anonymity in moving objects databases," in *2008 IEEE 24th International Conference on Data Engineering*. Ieee, 2008, pp. 376–385.

[8] M. Gramaglia and M. Fiore, "Hiding mobile traffic fingerprints with glove," *ACM CoNEXT*, pp. 1–13, 2015.

[9] O. Abul, F. Bonchi, and M. Nanni, "Anonymization of moving objects databases by clustering and perturbation," *Information Systems*, vol. 35, no. 8, pp. 884–910, 2010.

[10] L. Sweeney, "k-anonymity: A model for protecting privacy," *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems*, vol. 10, no. 05, pp. 557–570, 2002.

[11] A. Machanavajjhala, D. Kifer, J. Gehrke, and M. Venkitasubramaniam, "l-diversity: Privacy beyond k-anonymity," *ACM Transactions on Knowledge Discovery from Data (TKDD)*, vol. 1, no. 1, p. 3, 2007.

[12] N. Li, T. Li, and S. Venkatasubramanian, "t-closeness: Privacy beyond k-anonymity and l-diversity," in *2007 IEEE 23rd International Conference on Data Engineering*. IEEE, 2007, pp. 106–115.

[13] T. Li and N. Li, "On the tradeoff between privacy and utility in data publishing," in *Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM, 2009, pp. 517–526.

[14] S. Skiena, "Dijkstras algorithm," *Implementing Discrete Mathematics: Combinatorics and Graph Theory with Mathematica, Reading, MA: Addison-Wesley*, pp. 225–227, 1990.

[15] M. Gruteser and D. Grunwald, "Anonymous usage of location-based services through spatial and temporal cloaking," in *Proceedings of the 1st international conference on Mobile systems, applications and services*. ACM, 2003, pp. 31–42.

[16] B. Bamba, L. Liu, P. Pesti, and T. Wang, "Supporting anonymous location queries in mobile environments with privacygrid," in *Proceedings of the 17th international conference on World Wide Web*. ACM, 2008, pp. 237–246.

[17] M. Xue, P. Kalnis, and H. K. Pung, "Location diversity: Enhanced privacy protection in location based services," in *International Symposium on Location-and Context-Awareness*. Springer, 2009, pp. 70–87.

[18] B. Lee, J. Oh, H. Yu, and J. Kim, "Protecting location privacy using location semantics," in *Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM, 2011, pp. 1289–1297.

[19] F. Xu, Z. Tu, Y. Li, P. Zhang, X. Fu, and D. Jin, "Trajectory recovery from ash: User privacy is not p-reserved in aggregated mobility data," in *Proceedings of the 26th International Conference on World Wide Web*. International World Wide Web Conferences Steering Committee, 2017, pp. 1241–1250.