Revealing Privacy Vulnerabilities of Anonymous Trajectories

Shan Chang, *Member, IEEE*, Chao Li, Hongzi Zhu, *Member, IEEE*, Ting Lu, *Member, IEEE*, and Qiang Li

***Abstract*—The proliferation of various mobile devices equipped with GPS positioning modules makes the collection of trajectories more easier than ever before, and more and more trajectory datasets have been available for business applications or academic researches. Normally, published trajectories are often anonymized by replacing real identities of mobile objects with pseudonyms (e.g., random identifiers); however, privacy leaks can hardly be prevented. In this paper, we introduce a novel paradigm of *de-anonymization* attack re-identifying trajectories of victims from anonymous trajectory datasets. Different from existing attacks, no background knowledge or side channel information about the target dataset is required. Instead, we claim that, for each moving object, there exist some mobility patterns that reflect the preference or usual behavior of the object, and will not change dramatically over a period of time. As long as those relatively stable patterns can be extracted from trajectories and be utilized as quasi-identifiers, trajectories can be linked to anonymous historical ones. To implement such kind of de-anonymization attacks, an adversary only needs to collect a few trajectory segments of a victim, the durations of which do not necessarily overlap with that of trajectories in the target dataset (in simple terms, those trajectory segments are not necessary sub-trajectories included in the target dataset). Since the movements of victims in public areas could be observed openly, an adversary can obtain traces or locations about the victims either by direct monitoring them (e.g., tracking) or from third parties (e.g., social-networks). Then, the adversary extracts useful patterns from both the historical trajectories in the accessible dataset and newly obtained trajectory segments of victims, the historical trajectory with most similar patterns to that of a victim is considered as belonging to the victim. In order to demonstrate the feasibility of such attacks, we conduct extensive trace-driven simulations. We extract road segment preferences and stop of interests from trajectories of vehicles, and construct feature vectors (mobility patterns) of vehicles according to them, used for trajectory comparisons. Simulation results show that the adversary could re-identify anonymous trajectories effectively.**

***Index Terms*—De-anonymization attack, published anonymous trajectories, road segment preference, stop of interest, quasi-identifier.**

I. Introduction

NOWADAYS, GPS modules have become the standard equipments of a majority of mobile devices, which makes the collection of spatio-temporal data convenient. A series of consecutive position records from certain object during a period of time forms its spatio-temporal trajectory, which provides a good deal of valuable information to many applications. For instance, trajectories collected from vehicles can be used for improving the transportation managements, while moving path data of human beings could be utilized by the government for guiding the development of urban space. Plenty of trajectory data have been gathered and available for public in the last few years, e.g., Movebank [2], GeoLife [1], crowdflow [4] and RAWDAD [3]. Considering that spatio-temporal data are sensitive, the misuse of which may jeopardize the privacy of corresponding objects, e.g., health and marital statuses, home addresses and consumption capabilities, it is necessary to apply effective schemes on trajectories before being published, such that privacy of the objects can be guaranteed.

Currently, existing techniques for trajectories privacy-preserving fall into two groups: distortion and pseudonym. Different kinds of cloaking techniques are applied to trajectories for trajectory distortion, such as adding noises into trajectories designedly or decreasing the resolution of trajectories, in order to mask their precise temporal and spatial information. However, those distortion-based schemes could seriously impact the integrity and availability of trajectory datasets. In the other type of solutions, a unique and consistent random identifier, i.e., pseudonym, is used to replace the true identity of an object. Additionally, the pseudonym cannot be linked to the real identity of the object in any way. Since pseudonyms are easy to generate and apply, and there is no need to modify the original trajectories, they have been exploited extensively on published trajectories, e.g., the GeoLife GPS trajectory dataset published by Microsoft Research Asia [1]. However, even pseudonyms and true identities are disconnected, the effectiveness of pseudonyms on anonymizing trajectories is questionable. Researchers have claimed that due to the spatial and temporal features of

Manuscript received February 25, 2018; revised July 12, 2018 and September 14, 2018; accepted September 17, 2018. Date of publication September 24, 2018; date of current version December 14, 2018. This work was supported in part by the National Natural Science Foundation of China under Grants 61672151, 61402101, 61772340, 61472255, and 61420106010, in part by the Fundamental Research Funds for the Central Universities under Grant EG2018028, in part by the Shanghai Rising-Star Program under Grant 17QA1400100, in part by the DHU Distinguished Young Professor Program, in part by 2017 CCF-IFAA Research Fund, and in part by the Open Foundation of Symbol Computation and Knowledge Engineering of Ministry of Education, Jilin University. The review of this paper was coordinated by Prof. G. Mao. *(Corresponding author: Hongzi Zhu.)*

S. Chang, C. Li, and T. Lu are with the School of Computer Science and Technology, Donghua University, Shanghai 201620, China (e-mail: changshan@dhu.edu.cn; chaoli@mail.dhu.edu.cn; luting@dhu.edu.cn).

H. Zhu is with the Department of Computer Science and Technology, Shanghai Jiao Tong University, Shanghai 200000, China (e-mail: hongzi@cs.sjtu.edu.cn).

Q. Li is with the Key Laboratory of Symbolic Computation and Knowledge Engineering of Ministry of Education, Jilin University, Changchun 130021, China (e-mail: li_qiang@jlu.edu.cn).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TVT.2018.2871745

trajectories could be powerful quasi-identifiers, linkage attacks might re-identify anonymous trajectories [5]. A kind of de-anonymization attack has been formulated by C. Y. T. Ma *et al.* The authors indicated, as long as an adversary gathers some snapshots of trajectories of its moving victims, it is very likely that the adversary could recognize the trajectories of victims from a group of anonymized trajectories [6]. Those locations on certain trajectory are treated as quasi-identifiers. It should be noticed that, in order to launch the attack successfully, *the occurrence moment of those side information should within the duration of the anonymous trajectories to be identified*.

Actually, mobility pattern can also be quasi-identifiers, thus stronger and more general de-anonymization attacks might be launched, which implies that the spatio-temporal information obtained by adversaries (e.g., from side channel) need not overlap with the available anonymous trajectory dataset on time domain. We point out that, for two reasons, such attacks can be achieved. First, moving objects have mobility patterns recurring with relatively high frequencies, and those patterns will not change drastically over a period of time. Second, after anonymous trajectories being published, as long as the victims keep moving, the movements or whereabouts of them in public areas can be openly observed by adversaries through a variety of ways. For example, an adversary may obtain the movements of victims directly through following them by chance or engineered encounters. The adversary may also get spatio-temporal information either voluntarily or inadvertently. Such position information might be revealed in various way, including web blogs and social networks. Consequently, we introduce *a new paradigm of de-anonymization attacks utilizing mobility pattern-based quasi-identifiers*. In paradigm of the attack, adversaries learn mobility patterns of their victims from compromised moving trail or positions of the victims, and they have access to certain historical anonymous trajectory dataset and acquire mobility patterns of trajectories as well. By comparing mobility patterns of a victim with that of those anonymous trajectories in the dataset, the historical trajectory belonging to the victim can be identified. For example, an adversary who gathers several pieces of movement of a victim this week can identify the trace of the victim from an anonymous trace dataset published last month.

We conduct analysis on two large-scale real-world GPS-based trajectory datasets gathered from in Shanghai and Shenzhen, two metropolises in China, and extract road segment preferences and stops of interest (reflecting mobility patterns) from trajectories to construct their mobility feature vectors by using *Improved Term Frequency-Inverse Document Frequency* (ITF-IDF) values, and launch attacks successfully by comparing resulted feature vectors (trace-driven simulation results show the efficacy of the attacks), demonstrating the feasibility of such attacks. This also confirms that vehicles have relatively stable and distinguishable mobility patterns that can serve as quasi-identifiers.

The main contributions of this paper are highlighted as follows:

1) We formulate a new paradigm of de-anonymization attacks in which an adversary learns mobility patterns of victims (through real world knowledge-segments of movements or locations of them), and extracts mobility patterns of individual anonymous trajectories in a historical dataset accessible, and then trajectories of victims in the dataset can be recognized by comparing those mobility pattern-based quasi-identifiers.
2) We carry out intensive analysis on real-world trajectories of vehicles and discover that driving preference in terms of road segments of individual vehicle is stable and distinguishable, thus can be used for constructing a quasi-identifier.
3) We identify stops of interest of drivers from traces, which refer to the phenomenon a vehicle stops at a specific location for a while. Since stops of interest reflect the purpose of a driver, e.g., refueling at a gas station, having dinner in a shopping mall, sleeping at home, instead of blocked due to traffic jam, picking up or dropping off passengers, thus can also be used to build quasi-identifiers.
4) We present an effective strategy available to adversaries to make use of mobility patterns extracted. An ITF-IDF based method is introduced to form mobility pattern based feature vectors of a moving object. By comparing those feature vectors, an adversary can utilize compromised trajectories of its victims to identify their anonymous trajectories.
5) Extensive experimental results suggest that adversaries are able to re-identify anonymized trajectories with high probability, which means that such attacks are effective in breaching privacy.

The remainder of this paper is organized as follows. Section II presents related work. In Section III, we explain the notations and attack model, as well as two real-world datasets used for verifying the effectiveness the proposed attacks, and the necessary pre-processing on the datasets. Section IV gives analysis on the frequently travelled road segments to demonstrate the existence of road segment preferences. In Section V, we introduce how stops of interest can be extracted from trajectories, which can also characterize mobile objects. Section VI elaborates a strategy of adversaries generating and utilizing mobility features. Trace-driven simulation results are shown in Section VII. Finally, we conclude our work in Section VIII.

II. Related Work

Privacy of published datasets has received more and more attention [7]–[10], and protection schemes have been proposed to safeguard published trajectories from being identified, one of which is k-anonymity, introduced by L. Sweeney [11]. k-anonymity refers to that for each individual, there exist k-1 other individuals which are indistinguishable from it. Nevertheless, optimal k-anonymity (i.e., with minimal distortion) is NP-Hard [12], thus it is hard to achieve. Cloaking refers to a class of widely used methods, where the granularity of spatial-temporal information is reduced for the purpose of privacy preservation [13]. C. Li *et al.* proposed a Dynamic Reversible Cloaking mechanism in which an anonymization secret key is used to uniquely generate a cloaking region allowing the original location data to be restored from the cloaked data using the secret key [14].

Y. Tian *et al.* inferred social ties from cloaked human trajectories. The main idea is to transform cloaking regions into semantic regions, and to extract features from those semantic regions for social ties discovering [15].

Pseudonyms (random identifiers) replace the true identities of objects. The utility of the pseudonyms comes from the convenience of generating and applying them, and no require for modifying original trajectories. However pseudonym is not enough to provide anonymity as there may exist distinguishing features (playing the role of quasi-identifiers) which are sufficient to distinguish individuals.

Recently, research results on the evaluation of privacy leakage of anonymized trajectories or locations have been reported. Through analysis on a large number of mobile phone call records, H. Zang and J. Bolot aim to recover sensitive information of users using only position information [16]. The experimental results reveal the truth that a large ratio of users could be re-identified from sector or cell level position information, which implies serious privacy vulnerabilities of anonymous position information. D. Kondor *et al.* verified the re-identifiability of records by matching them from two datasets. One is a mobile communication dataset in which each record represents the start or the end of a call, and another is a transportation dataset, based on the smart cards used by the electronic fare system on buses and trains of Singapore. They derived that for individuals who make 3-4 trips per day on average in the transportation system, it is expected to achieve a 16.8% success rate based on a one-week long observation of their mobility traces [17]. K. Emara focused on location privacy leakage of broadcasting messages of vehicles in VANET safety applications. He investigated several distortion-based privacy schemes, and considered both privacy gain and impact on safety applications. A traceability and distortion-based metric is proposed to determine how long and how accurate adversaries can track vehicles [18].

The use of location-based services (LBS) may also cause the damage of privacy. J. Freudiger *et al.* conducted investigation on mobility traces to evaluate the success of LBSs in deducing the true identities of users which are substituted by pseudonyms, and their points of interest. Their analysis explores the relations between the quantities and types of data acquired by LBSs and their powers to re-identify and to characterize customers [19]. Z. Xiao *et al.* distinguished the privacy issue of Vehicular LBS as an independent problem from the LBS privacy problem. They found that due to vehicle traces are mostly constrained by roads, the trajectories of vehicles are highly unique. If four points of a vehicle in an anonymous dataset were captured, with high probability, it can be identified uniquely [20]. A-M. Olteanu *et al.* proved that exploiting online social relationships can increase the accuracy of location inferring. They considered attackers who have knowledge about social relationships of their victims, and formalized the problem of deducing location of victims through co-location information with their friends on social networks [21].

Martin *et al.* quantified the impact of background knowledge obtained by an attacker on privacy damage [22]. They presented an algorithm to calculate the volume of unveiled private information with respect to the amount of the background knowledge in the worst case. An empirical risk model of *k*-anonymity based private data releasing was introduced by A. Basu *et al.* [23]. The authors consider the tradeoff between data privacy and utility, and attack cost. Y. Zhao *et al.* evaluated the strength of 41 privacy metrics for vehicular networks in terms of 4 criteria, i.e., monotonicity, extent, evenness and shared value range, using real traffic data. Their results showed that single metric cannot dominate all criteria and traffic conditions, thus metric suites were recommended [24]. R. Pellungrni *et al.* introduced linear regression-based privacy risk (level) estimation on mobility patterns of individuals, under re-identification attacks they mentioned [25]. A very recent work from F. Xu *et al.* demonstrated that only having access to statistic data of mobile users' mobility, it's possible for an attacker to recover the trajectories of those users, since individuals show unique and regular mobility pattern. Once the trajectories are recovered, then the privacy is immediately under the threat of re-identification attacks [26].

Chris Y. T. Ma *et al.* proposed a side information (i.e., location snapshots) based de-anonymization attack, where an attacker, obtaining a number of location snapshots of its victims, can recognize the trace of the victims from a set of anonymous traces [6]. The authors use Bayesian inference to break the unlinkability between location snapshots and anonymized trace. H. Wang *et al.* used two mobile social network datasets as side information to evaluate the performance of de-anonymization attacks using external information. A Gaussian and Markov based algorithm is adapted to deal with spatiotemporal mismatches in different datasets [27]. H. Li *et al.* measured the similarity between the disclosed locations in the MSN applications and the real mobility pattern, in terms of coverage rate and relative entropy, and presented an attack to infer MSN users' demographics from the disclosed locations by checking their similar point of interests [28]. Beside feature-location-based approaches, Markov Model based schemes are proposed for de-anonymization, for example in references [29] and [30]. These schemes require that an adversary is able to observe the movements of his or her victims for several days or weeks in the past in order to train a model of Mobility Markov Chain used in future testing.

The paradigm of de-anonymization attacks proposed is different from the previous work by particular considering the feasibility to make use of long-term mobility patterns of trajectories, which are stable and distinguishable, as quasi-identifiers, such that the historical anonymous trajectories can be re-identified. An adversary, after collecting a few number of trajectory pieces of a victim, may employ well grounded strategies to re-identify the victim from a set of historical anonymous trajectories.

III. Preliminaries and Attack Model

In this section, we declare some notations, introduce the attack model, and describe the datasets we used.

A. Notations

- *Trajectory Dataset:* a trajectory dataset D_{mt} is denoted as $\{\mathcal{V}, \mathcal{T}\}$. $\mathcal{V} = \{v_1, v_2, ..., v_n\}$ refers to the set of mobile objects. $\mathcal{T} = \{\mathcal{TR}_1, \mathcal{TR}_2, ..., \mathcal{TR}_n\}$ represents the set of

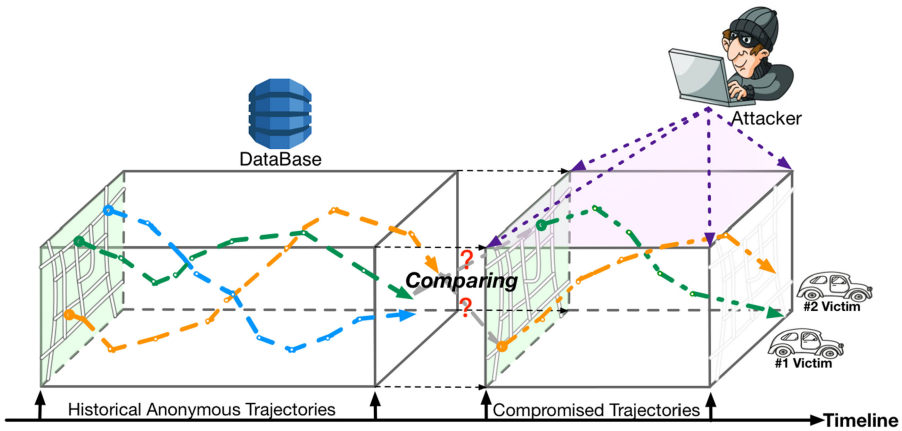


Fig. 1. An example of de-anonymization attack. The historical anonymous trajectory dataset contains three trajectories (orange, blue, and green polylines), and the adversary captures two trajectories from two victims, and wishes to link anonymous trajectories with the compromised trajectories.

trajectories of all mobile objects, where $\mathcal{TR}_i$ associates with the mobile object $v_i \in \mathcal{V}$.

- *Trajectroy:* a trajectory $\mathcal{TR}_i$ is composed of a series of GPS reports, each of which contains a timestamp t, a pair of longitude and latitude coordinates (x, y).
- *Event of Stop:* an even of stop $\mathcal{E}_j^{(i)}$ refers that a vehicle v_i stays at certain place over a period of time δ, which is described by a pair of latitude and longitude coordinates (x_j, y_j), and a pair of beginning and ending timestamp (t_{b_j}, t_{e_j}). The duration $dura_i$ of a stop event is $t_{e_j} - t_{b_j}$.
- *Stop of Interest (SI):* a stop of interest $\mathcal{S}i_i$ a special kind of stop events, which describes the phenomenon v_i stops at a specific location for the driver's own purpose, e.g., refueling at a gas station, having dinner in a shopping mall, sleeping at home, instead of blocked due to traffic jam, picking up or dropping off passengers.
- *Road Segment (RS)*: a road segment r_i refers to a one-direction edge which is connected with two intersections $(\mathcal{I}t_i^{(bg)}, \mathcal{I}t_i^{(ed)})$.
- *Route:* a Route $\mathcal{R}_i$ is a set of consecutive road segment, i.e., $r_1 \rightarrow r_2 \ldots \rightarrow r_q$, where $\mathcal{I}t_j^{(ed)} = \mathcal{I}t_{j+1}^{(bg)}, 1 \leq j < q$.
- *Anonymization of Trajectories:* before releasing a trajectory dataset, the true identity of each trajectory in the dataset will be substituted by a unique pseudonym, e.g., a random number. There are two principles must be obeyed in the selection of pseudonyms. First, the true identities of trajectories should not be related to the pseudonyms, no matter what. Second, the same pseudonym will always be mapped to the same identity.

B. Attack Model

An adversary launching a *de-anonymization* attack attempts to unveil the true identity of an anonymized trajectory. A *victim* refers to a mobile object whose trajectory is exposed (namely compromised) to the adversary. The objective of the adversary is to re-identify as many anonymized trajectories as possible, utilizing compromised trajectories. To be precise, in the proposed paradigm of *de-anonymization* attack, the adversary breaks anonymity of trajectories through the following steps (an example is illustrated in Figure 1):

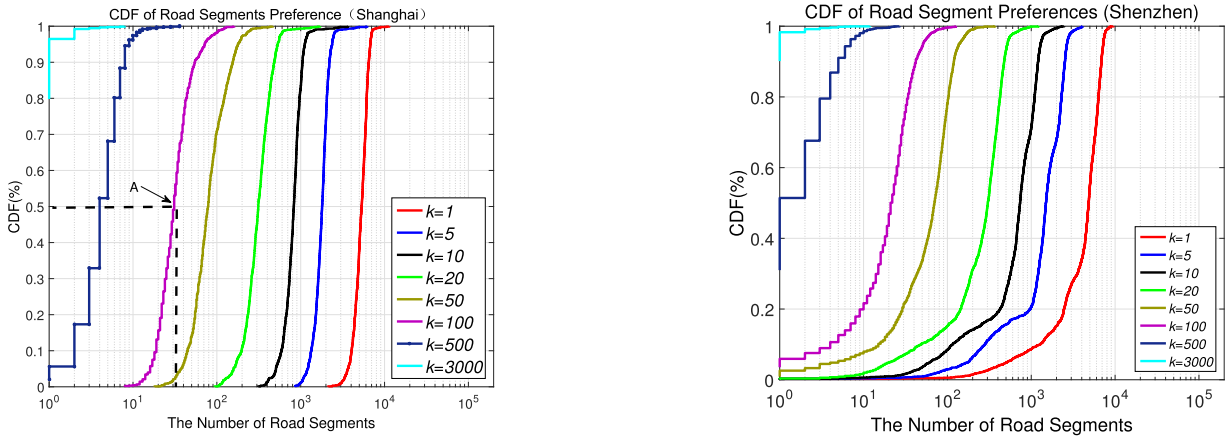
1) The adversary is able to access a collection of historical trajectories, which contains the trajectories of its victims and other mobile objects unconcerned.
2) The adversary is able to obtain several pieces of trajectories of the victims. There are various ways to know the movements of the victims, such as tracking or monitoring them in public areas directly, or learning from third parties like social networks. It should be noticed that the compromised trajectories are not necessary to overlap with historical trajectories in time domain.
3) The adversary carries out the strategy of re-identification (comparing compromised trajectories with historical ones) such that the victims can be associated to their historical trajectories.

TABLE I
STATISTICAL CHARACTERISTICS OF THE TRAJECTORY DATASETS

Dataset	Shanghai	Shenzhen
Number of vehicles	906	1945
From date	Feb. 1, 2007	Oct. 1, 2009
Duration (day)	28	31
Granularity (second)	15, 60	60
Number of Trajectories	25,368	60,295

C. Dataset and Pre-Processing

- *Datasets:* we choose two sets of trajectories, each of which contains more than 10 thousands of taxies, in two large cities of China, i.e., Shanghai and Shenzhen. Each GPS report in trajectories includes: *the ID of the corresponding taxi, longitude and latitude coordinates, moving speed, timestamp*, as well as *an operational status*, indicating whether there are passengers onboard or not. The GPS reports are collected at a granularity of 60 seconds (the granularity of reports in Shanghai dataset is 15 seconds if there are passengers onboard). Table I shows the statistical features of Shanghai and Shenzhen datasets.



(a) Shanghai (b) Shenzhen

Fig. 2. CDFs of the number of road segments on routes satisfying k strength of preferences under different k values.

- *Road Networks:* in Shanghai and Shenzhen datasets, there include 66,459 and 89,578 road segments in their road networks, respectively. Each road segment is assigned a unique ID in corresponding road network.

In each dataset, the whole trajectory $\mathcal{TR}_i$ of v_i is separated into num_{day} *1-day sub-trajectories* $\mathcal{TR}_i^{(q)}(0 < q \leq num_{day})$, which indicates the trace of v_i in the q-th day. Furthermore, in order to eliminate GPS errors, a ST-Matching algorithm [31] is utilized to map GPS reports to their most likely road segments, thus a trajectory $\mathcal{TR}_i$ (or $\mathcal{TR}_i^{(q)}$) can be simplified as a route $\mathcal{R}_i$ (or $\mathcal{R}_i^{(q)}$), which refers to sequence of road segments.

IV. Analyzing on Frequently Travelled Road Segments

In this section, we carry out analysis on the two datasets in order to answer two questions. First, given a trajectory, whether some road segments occur in the corresponding route with higher frequencies than others? In other words, the driver prefers traveling on those road segments rather than others. Second, whether different trajectories exhibit different preferences of road segments? If experimental results to above questions are "positive", then it's possible to select a collection of frequently travelled road segments for each vehicle to serve as its mobility features, such that it can be distinguished from others.

A. Road Segment Preferences

Intuitively, vehicles pass through different road segments with uneven probabilities. There are various kinds of factors affecting the selection of traveling routes, for example the area a driver's home located in, the gas station or restaurant the driver get used to visit to, and the importance of road segments to the road network. Consequently, road segments appears for different times on different routes.

For each dataset, in order to verify the existence of road segment preferences, we conduct the following experiments. We focus on the number of times that each road segment is passed through by certain vehicle (in other words, occurs on the corresponding route). We set k as a *preference strength parameter*. If a road segment occurs on a route no less than k times, then we call this road segment satisfies k strength of preference on this route. For each route $\mathcal{R}_i(1 \leqslant i \leqslant n)$, we calculate the number of road segments which satisfies k strength of preference, denoted as $\mathcal{Z}_i^{(k)}$. We choose different k values, and the CDFs of $\mathcal{Z}_i^{(k)}(1 \leqslant i \leqslant n)$ are shown in Figure 2(a) and (b). As we can see in Figure 2(a), on the curve of $k = 100$, the point A indicates that 50% routes have no more than 30 road segments satisfying 100 strength of preference, and the curve also demonstrates that no route contains more than 200 road segments satisfying 100 strength of preference. Furthermore, as k value increases, $\mathcal{Z}_i^{(k)}$ declines significantly.

According to the above analysis on real trajectories, we have two observations. First, for each vehicle, there do exist a collection of road segments frequently travelled. Second, although tens of thousands of road segments are included in both Shanghai and Shenzhen road networks, the size of the above collections of road segments are quite small for all vehicles. Both observations imply the existence of road segment preferences.

B. Difference of Preferences Between Routes

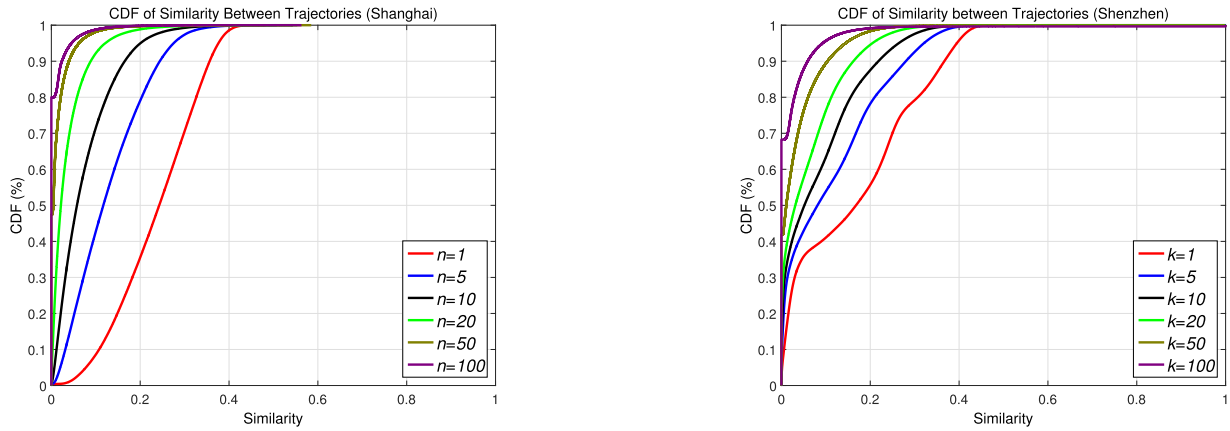
In order to compare the road segment preferences between routes, we introduce the following notations:

1) $\mathcal{C}_i = \{r_j | b_j^{(i)} > 0\}$, represents the collection of road segments occurring on route $\mathcal{R}_i$, where $b_j^{(i)}$ denotes the number of times that road segment r_j occurs on $\mathcal{R}_i$.
2) $\mathcal{C}_i^{(k)} = \{r_j | b_j^{(i)} \geqslant k\}$, i.e., $\mathcal{C}_i^{(k)}$ represents the collection of road segments satisfying k strength of preference on $\mathcal{R}_i$.
3) $\mathcal{C}_i^{(k)} \cap \mathcal{C}_j^{(k)} = \{r_l | r_l \in \mathcal{C}_i^{(k)}$ and $r_l \in \mathcal{C}_j^{(k)}\}$
4) $\mathcal{C}_i^{(k)} \cup \mathcal{C}_j^{(k)} = \{r_l | r_l \in \mathcal{C}_i^{(k)}$ or $r_l \in \mathcal{C}_j^{(k)}\}$

Then, given a value of k, we define the similarity between routes $\mathcal{R}_i$ and $\mathcal{R}_j$ as:

$$Sim_{i,j}^{(k)} = \frac{|\mathcal{C}_i^{(k)} \cap \mathcal{C}_j^{(k)}|}{|\mathcal{C}_i^{(k)} \cup \mathcal{C}_j^{(k)}|} \tag{1}$$

where $|\mathcal{C}|$ represents the cardinality of a set $\mathcal{C}$.



(a) Shanghai (b) Shenzhen

Fig. 3. CDF of similarities between pairs of routes under different k values.

For each dataset, we calculate the similarities between every pair of routes under different preference strength parameters, and the experimental results are shown in Figure 3(a) and (b) using CDFs curves. It can be seen that similarities between routes in both datasets are no larger than 0.45. It means that individual vehicle has unique road segment preference. In addition, increasing of k will further reduce of similarities between routes, which is because that interference from low preference strength road segments is eliminated. As illustrated in Figure 3, for $k = 100$, 90% pairs of route have less than 0.1 similarity. It confirms that routes have dissimilar road segment preferences, thus those preferences have potential to be quasi-identifiers.

V. Extracting Stops of Interest From Trajectories

In this section, we conduct analysis on the two trajectory datasets and try to identify stops of interest in the datasets, which can also reflect the unique behavior of a driver (e.g., the driver gets use to visit a gas station for refueling gasoline). However, there are mainly four different kinds of events of stop, i.e., stops for boarding and alighting passengers, for waiting traffic-lights, for traffic congestion, and stops of interest (only the last kind of stops are relative to the willingness of drivers), and it is very difficult to distinguish stops of interest from other three types of stops. We overcome the challenge through a three-step processing (the whole procedure of extracting stops of interest is illustrated in Figure 4).

A. Removing Stops for Boarding or Alighting Passengers

Noticing that a GPS report contains an operational status indicator, whose value is *0* or *1*, representing whether the corresponding taxi is vacant or occupied, respectively, thus stops for boarding or alighting passengers can be identified through the following rules: given a trajectory, if the operational status indicator in the corresponding GPS reports changes from *0* to *1* before and after an event of stop, then the event of stop is considered as a boarding passenger event. On the contrary, the indicator changes from *1* to *0* implies an event of stop for alighting passengers. According to this, we remove all stops for boarding and alighting passengers from the two datasets.

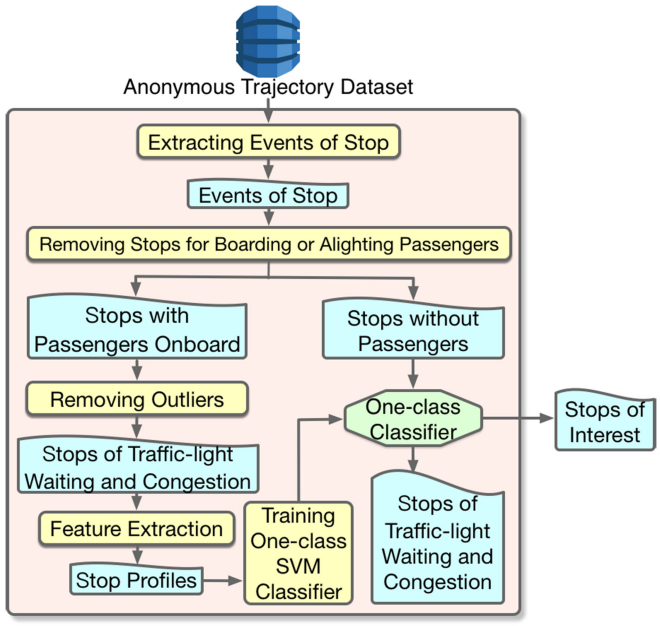
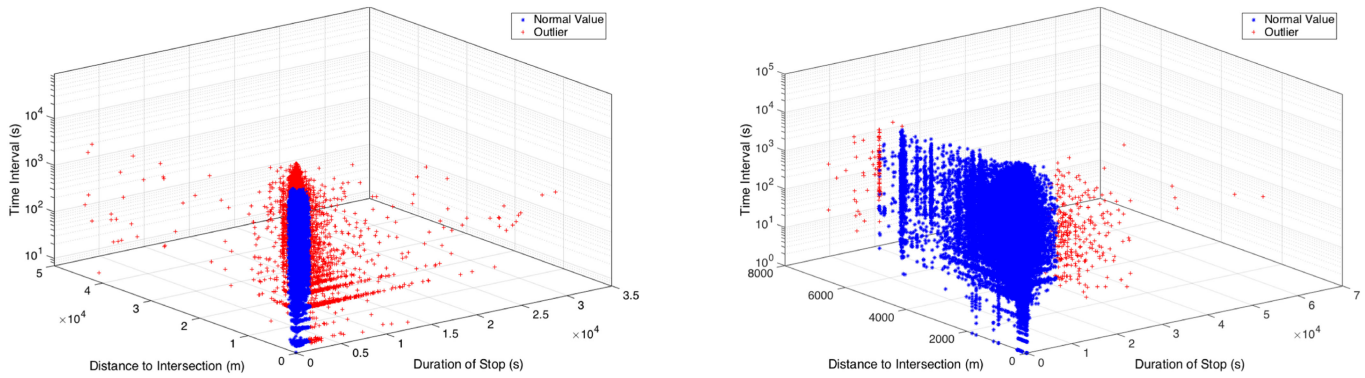


Fig. 4. The procedure of extracting stops of interest.

B. Modeling Stops of Traffic-Light-Waiting and Congestion

Considering that stops of interest may only occur when taxis are vacant (we explain the reason in the next paragraph), we separate the remaining stops into two groups, i.e., stops with and without passengers onboard (according to the operational status indicator), respectively. In the group of stops without passengers, stops for traffic-light waiting and traffic congestion, and stops of interest are contained. However, it is very challenging to distinguish stops of interest from the other two kinds of stops, since the locations and durations of stops of interest are diverse and hardly predicted.

To overcome the above difficulty, we also analyze on stops with passengers and get the following observations: 1) since the route is under the control of the passengers whose purposes are to arrive certain place as soon as possible, the corresponding trajectories are unlikely to contain stops of interest; 2) since passengers are charged not only by distance, but also by



(a) Shanghai (b) Shenzhen

Fig. 5. Distributions of stops with passengers in the three-dimensional coordinate system of the corresponding triples, where red points refer as outliers.

midway standby period as well, very few passengers require to stop the taxis, which implies that the overwhelming majority reasons for stops are traffic-light waiting and traffic congestion. Consequently, if we regard the stops for the requirements of passengers as outliers and filter them out from the group, then those stops remained (which are stops of traffic-light waiting or traffic congestion) can be used to build a model, characterizing the features of this two types of stops. Then the model can also be used to check the group of stops without passengers such that stops of interest can be excluded from it.

We also observe that, in urban area, traffic congestions may cause 1) short interval successive stops; 2) stops with short duration. While traffic-light waiting stops have the features of short duration and close to intersection. Thus, given an event of stop $\mathcal{E}_i$, we utilize the *duration of stop* ($dura_i$), *distance to the closest intersection* ($dist_i$) and *time interval between two adjacent stops* ($inter_i$) to construct a triple $\mathcal{S}_{tp_i}$, named *stop profile*, for charactering different kinds of stops.

Given a group of stops with passengers, after having prepared all stop profiles, we adopt *Mahalanobis distance,* denoted as $\mathbb{D}_m$, which is commonly utilized to measure the deviation of a point from a distribution in an unitless and scale-invariant way, to identify outliers. A stop profile $\mathcal{S}_{tp}$ with large $\mathbb{D}_m$ can be indicated as an outlier. A cutoff value of $\mathbb{D}_m$ for identifying outliers is recommended as $\sqrt{\chi^2_{(p,0.975)}}$ (p degrees of freedom) [32]. Given a stop profile $[dura_i, dist_i, inter_i]$, denoted as $\mathcal{S}_{tp_i}$, the $\mathbb{D}_m(\mathcal{S}_{tp_i})$ to a set of stop profiles $\mathbb{S}$ with mean value $\boldsymbol{\mu} = [\overline{dura}, \overline{dist}, \overline{inter}]$ and covariance matrix $\mathcal{V}$, is defined as:

$$\mathbb{D}_m(\mathcal{S}_{tp_i}) = \sqrt{(\mathcal{S}_{tp_i} - \boldsymbol{\mu})^T \mathcal{V}^{-1} (\mathcal{S}_{tp_i} - \boldsymbol{\mu})} \tag{2}$$

where $\mathcal{V}$ can be calculated as

$$\mathcal{V} = E\{(\mathbb{S} - \boldsymbol{\mu})^T (\mathbb{S} - \boldsymbol{\mu})\}. \tag{3}$$

There are 374,157 and 677,419 stops with passengers in Shanghai and Shenzhen datasets, respectively. We calculate $\mathbb{D}_m$ for all stops with passengers in the two datasets, and sort them according to their $\mathbb{D}_m$. Then we set the cutoff values of $\mathbb{D}_m$ for identifying outliers in Shanghai and Shenzhen datasets as 3.993 and 17.042, according to $\sqrt{\chi^2_{(2,0.975)}}$ [32], respectively.

For Shanghai and Shenzhen datasets, each stop without passengers is mapped to a point in a three-dimension coordinate system according to its $\mathcal{S}_{tp}$ in Figure 5(a) and (b), respectively, where x-axis and y-axis indicate the duration of stop and the distance to the closest intersection, and z-axis is the time interval between two adjacent stops. Blue points refer to stops for traffic-light waiting or traffic jam, while red points are outliers.

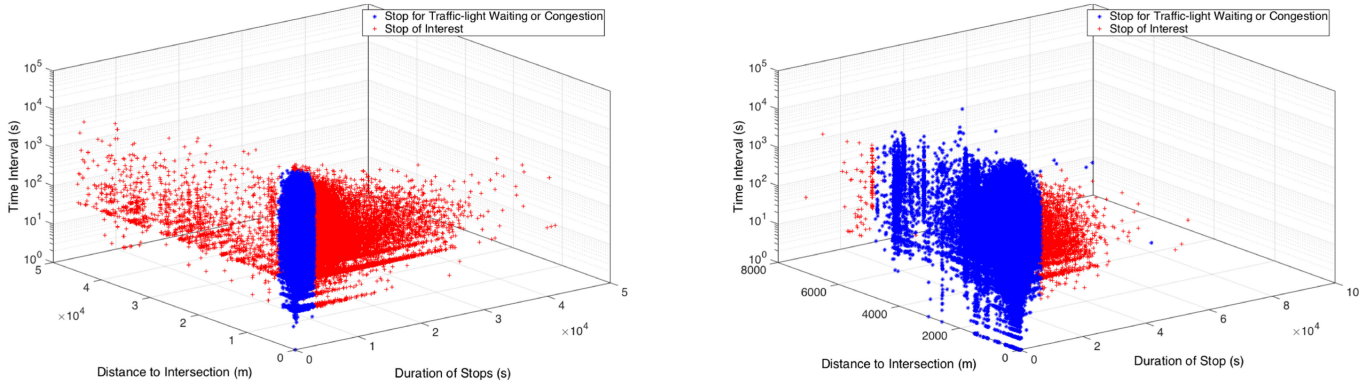
C. Identifying Stops of Interest

For stops with passengers in each dataset, after removing outliers, the purified stops for traffic-light waiting or traffic congestion can be used to train a one-class Support Vector Machine (SVM) classifier, with the Radial Basis Function (RBF) kernel function. Then stops without passengers are fed into the corresponding classifier for deciding whether the stop belongs to the class or not. Considering that in the group of stops without passengers, besides stops for traffic-light waiting or traffic congestion, only stops of interest exist, thus those stops excluded from the class are the stops of interest. Figure 6(a) and (b) illustrate the distributions of stops without passengers in Shanghai and Shenzhen, in the three-dimension coordinate system of stop profiles, where red points refer to stops of interests.

We conduct analysis on the resulted stops of interest, and figure out that individual vehicle presents relatively stable stops of interest each day. Figure 7(a) and (b) display stops of interest extracted from a trajectory in Shanghai and Shenzhen datasets, respectively, where points with different colors represent stops on different days. It can been seen in both figures that the corresponding vehicle stoped on similar positions almost every single day within one month. Thus we speculate that stops of interest of a vehicle could be utilized to build a quasi-identifier of it.

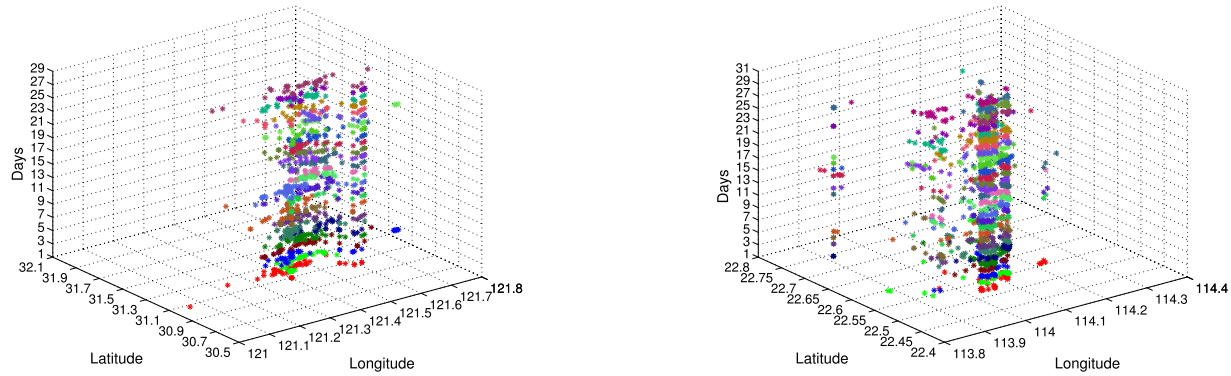
VI. A Strategy of Mobility Pattern-Based De-Anonymization

In this section, we give a simple strategy can be used by an adversary, to demonstrate the feasibility of the proposed mobility pattern-based de-anonymization paradigm. The key idea is to construct a *mobility feature vector* for each vehicle satisfying that vectors extracted from trajectories of the same vehicle have high similarities (i.e., matching scores), while vectors of different vehicles have low matching scores.



(a) Shanghai (b) Shenzhen

Fig. 6. Distributions of stops without passengers in the 3d coordinate system of the corresponding triples, where red points refer as stops of interests.



(a) Shanghai (b) Shenzhen

Fig. 7. Stops of interest extracted from a trajectory.

Consider an adversary who can access anonymized trajectory dataset D_{mt}, which contains n trajectories $\{\mathcal{TR}_1, \mathcal{TR}_2, \ldots, \mathcal{TR}_n\}$. Furthermore, the adversary acquires a piece of *compromised* trajectory of a victim, denoted as $\widetilde{\mathcal{TR}}_\sigma$. The goal of the adversary is to link $\widetilde{\mathcal{TR}}_\sigma$ to the *target trajectory* $\mathcal{TR}_\sigma$ belonging to the victim. Notice that there is no need for overlapping between $\widetilde{\mathcal{TR}}_\sigma$ and $\mathcal{TR}_\sigma$ on time domain.

The adversary performs a three-step strategy to launch the de-anonymization attack. We take stop of interest preference as example, while road segment preferences can be easily utilized with the proposed strategy.

- First, for each trajectory $\mathcal{TR}_i$ in $\mathcal{T}$, the adversary first extracts stops of interest from it, and then maps each obtained stop of interest to its nearest intersection. In this way, $\mathcal{TR}_i$ is converted into a sequence of intersections, denoted as $\mathcal{I}_i$. Thus, given a set of trajectories $\mathcal{T}$, the corresponding set of intersection sequences, i.e., $\mathcal{I} = \{\mathcal{I}_1, \mathcal{I}_2, \ldots, \mathcal{I}_n\}$, can be constructed. Similarly, a sequence of intersections $\widetilde{\mathcal{I}}_\sigma$ can be extracted from $\widetilde{\mathcal{TR}}_\sigma$;
- Second, given an $\mathcal{I}_i$, the adversary calculates an *improved term frequency-inverse document frequency* (ITF-IDF) value for each intersection in $\mathcal{I}_i$, the resulted ITF-IDF values are utilized to construct the Stop of Interest (SI) feature vector of $\mathcal{I}_i$. The ITF-IDF value of intersection $\mathcal{I}t_j$ indicates its significance in $\mathcal{I}_i$;
- Finally, the adversary proceeds a *feature vector matching algorithm*, and selects the *suspect trajectory* $\mathcal{TR}^*$, which has the most similar SI feature vector to that of $\widetilde{\mathcal{TR}}_\sigma$.

A successful attack refers that the suspect trajectory $\mathcal{TR}^*$ belongs to the victim.

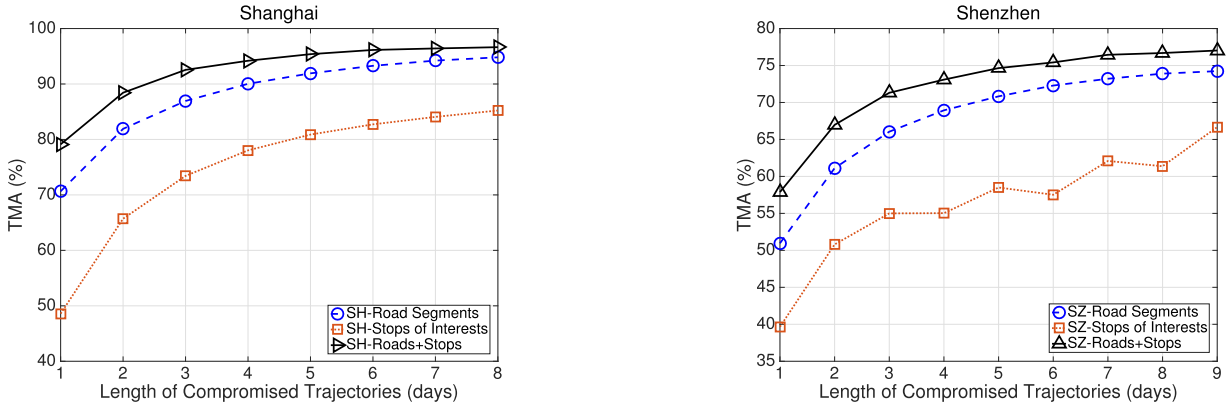
Remarks: A simple way to construct *Road Segment (RS) feature vectors* is to calculate an ITF-IDF value for each road segment on route $\mathcal{R}_i$ (the ITF-IDF value of r_j indicates the significance of r_j on $\mathcal{R}_i$), and then deal with those road segments just like what will be done to the intersections.

A. SI Feature Vector Extraction

We formulate the SI feature vector of a trajectory $\mathcal{TR}_i$ as $f_i = (w_{i,1}, w_{i,2}, ..., w_{i,\phi})^T$. where $w_{i,j}$ represents the ITF-IDF value of the intersection $\mathcal{I}t_j (1 \leq j \leq \phi)$, and ϕ indicates the total number of intersections in the corresponding road network. $w_{i,j}$ can be calculated by:

$$w_{i,j} = \frac{b_j^{(i)}}{\mathcal{N}_i} \times \log\left(c_j^{(i)} \times \frac{n}{c_j}\right)$$

where $b_j^{(i)}$ is the number of times $\mathcal{I}t_j$ appearing in $\mathcal{I}_i$ and $\mathcal{N}_i$ is the total length of $\mathcal{I}t_j$. Consequently, $b_j^{(i)}/\mathcal{N}_i$ is the frequency $\mathcal{I}t_j$ occurring in $\mathcal{I}_i$. c_j is the number of trajectories in $\mathcal{D}_{mt}$ whose corresponding intersection sequences contain



(a) Shanghai (b) Shenzhen

Fig. 8. TMA under different length of compromised trajectories.

$\mathcal{I}t_j$, and $c_j^{(i)}$ refers how many 1-day trajectories of $\mathcal{TR}_i$ (i.e., $\mathcal{TR}_i^{(q)}(1 \leq q \leq num_{day})$) whose corresponding intersection sequences contains $\mathcal{I}t_j$.

The proposed ITF-IDF value has the following characteristics:

- If $\mathcal{I}t_j$ appears in $\mathcal{I}_i$ for a large number of times, it implies that the driver usually stops the car, i.e., v_i, around intersection $\mathcal{I}t_j$, thus corresponding $w_{i,j}$ tends to be large;
- If $\mathcal{I}t_j$ appears in many intersection sequences, which implies that the occurrence of $\mathcal{I}t_j$ in $\mathcal{I}_i$ lacks of uniqueness, then corresponding $w_{i,j}$ tends to be small;
- If $\mathcal{I}t_j$ appears in several 1-day sub-sequence of $\mathcal{I}_i$, which implies that the occurrence of $\mathcal{I}t_j$ in $\mathcal{I}_i$ is stable in terms of a day, then corresponding $w_{i,j}$ tends to be large.

By the above design, if $\mathcal{I}t_j$ is significant on distinguishing $\mathcal{I}_i$ from others, the corresponding $w_{i,j}$ will be large.

B. SI Feature Vector Matching

We also construct an SI feature vector for the compromised trajectory $\widetilde{\mathcal{TR}}_\sigma$, which is $\widetilde{f}_\sigma = (\mu_{\sigma,1}, \mu_{\sigma,2}, ..., \mu_{\sigma,\phi})^T$. We calculate $\mu_{\sigma,j}$ as

$$\mu_{\sigma,j} = \frac{\widetilde{b}_j^{(\sigma)}}{\sum_{j=1}^{\phi} \widetilde{b}_j^{(\sigma)}}$$

where $\widetilde{b}_j^{(\sigma)}$ indicates the number of times $\mathcal{I}t_j$ occurring in $\widetilde{\mathcal{I}}_\sigma$.

Then, the matching score $\mathcal{M}_{Sc}$ between $f_i (1 \leqslant i \leqslant n)$ and $\widetilde{f}_\sigma$ can be calculated as

$$\mathcal{M}_{Sc}^{(i,\sigma)} = \frac{f_i \cdot \widetilde{f}_\sigma}{\|f_i\| \|\widetilde{f}_\sigma\|} = \frac{\sum_{j=1}^{\phi} w_{i,j} \times \mu_{\sigma,j}}{\sqrt{\sum_{j=1}^{\phi} (w_{i,j})^2} \times \sqrt{\sum_{j=1}^{\phi} (\mu_{\sigma,j})^2}}$$

Remarks: Consider that a large road network contains a huge number of intersections and road segments, implying a very large ϕ, thus the dimension of SI or RS feature vectors should be very high. Fortunately, SI or RS feature vectors are normally sparse vectors, the dimension can be reduced, e.g., by performing Principle Components Analysis (PCA).

VII. Evaluation

A. Methodology

To evaluate the effectiveness of the proposed de-anonymization attack, we carry out trace-driven simulations utilizing the datasets mentioned before. We choose the first 20 and 22 day trajectories in Shanghai and Shenzhen dataset acting as the anonymized trajectory dataset $\mathcal{T}$, respectively. The rest of trajectories (8 days long in Shanghai, and 9 days long in Shenzhen) in each dataset, denoted as $\mathcal{T}'$, are treated as compromised trajectories. In our experiments, we use $\mathcal{T}'$ to re-identify as many trajectories as possible in $\mathcal{T}$.

We introduce a metric, named *Trajectory Matching Accuracy* (TMA), to measure the effectiveness of the de-anonymization attacks, which is defined as

$$TMA = \frac{n_{crt}}{n}$$

where n_{crt} and n indicate the number of trajectories re-identified, and the total number of trajectories, in $\mathcal{T}$.

B. The Impact of the Length of Compromised Trajectories

We investigate how does the length of a compromised trajectory $\widetilde{\mathcal{TR}}_\sigma$ affect the success of a de-anonymization attack. To this end, we conduct the following experiments using Shanghai and Shenzhen datasets, respectively.

We vary the length of all compromised trajectories $\widetilde{\mathcal{TR}}_\sigma$ in $\mathcal{T}'$ from 1-day sub-trajectories to the whole trajectories (8 and 9 days long in each dataset) with an increment of one day. Then launch the RS feature based and SI feature based de-anonymization attacks under different length of compromised trajectories, respectively.

Simulation results are shown in Figure 8(a) and (b), where x and y axes represent the length of compromised trajectories and trajectory matching accuracies, respectively. From the figure we can see that, in Shanghai and Shenzhen dataset, by using 1-day long compromised trajectories, RS feature based TMAs are around 70% and 50%, and SI feature based TMAs are around

50% and 40%, respectively. We emphasize that TMAs of those mobility feature based attacks are much higher than that of random guessing attacks (Consider an adversary tries to re-identify $\mathcal{T}$ according to $\mathcal{T}'$, the random guessing attack launched by it can be seen as a sequence of n independent experiments, in each of which, for a compromised trajectory selected in $\mathcal{T}'$, the probability that the adversary guesses the corresponding trajectory from $\mathcal{T}$ successfully is $1/n$. Notice that since the adversary cannot know the result of the experiments, the trajectories which have been selected from $\mathcal{T}$ in previous experiments will not be excluded from subsequent experiments. Such a sequence of success/failure experiments is a Bernoulli process, i.e., $B(n, 1/n)$, then the average number of successful results will be $n \times 1/n = 1$. Comparatively, in the proposed attacks, n equals to 1945 in Shenzhen dataset, thus the average number of trajectories re-identified successfully by RS and SI features are 972 and 778, respectively). Moreover, simulation results illustrate that, as longer trajectories are available, higher TMAs can be achieved. Using 8 and 9 days long $\widetilde{\mathcal{TR}}_\sigma$ in Shanghai and Shenzhen datasets, RS feature based TMAs can reach up to 95% and 70%, and SI feature based TMAs can reach up to 85% and 67%, respectively.

C. The Impact of the Size of Suspect Trajectory Set

In the above de-anonymization attacks, given a compromised trajectory, the adversary treats the anonymized trajectory with the highest matching score to it as its suspect trajectory (i.e., deciding a single suspect trajectory). From another perspective, it is also significant for privacy-damaging if the adversary could almost certain that the target trajectory is in a small set of suspect trajectories (in other word, the target trajectory is in the set with a high probability). Thus the definition of TMA can be generalized as

$$TMA = \frac{n_{crt}^m}{n},$$

where m refers to the size of the set of suspect trajectories, and n_{crt}^m indicates the number of trajectories in $\mathcal{T}$ which fall into the corresponding m size suspect trajectory sets.

In order to verify the impact of m to TMAs, we set m from 1 to the number of the vehicles in each dataset (i.e., 906 and 1945, respectively) with an increment of one, and conduct both RS and SI feature based attacks under different m values. In the experiments, the adversary ranks all anonymized trajectories according to their matching scores to a particular compromised trajectory, and takes the top m trajectories to constitute its suspect trajectory set. Simulation results are shown is Figure 9, where x and y axes represent the size of suspect trajectory set and trajectory matching accuracies. Simulation results show that, even m takes a small value of five, for Shanghai dataset, the RS and SI feature based TMAs can reach up to 93% and 74%, and for Shenzhen traces, these two TMAs can reach up to 73% and 67%, respectively. Furthermore, TMAs keep increasing with the increase of m.

D. Improving TMA by Hybrid Feature Vector

In the last two experiments, the features of *road segments* and *stops of interest* are utilized separately in the attack strategy.

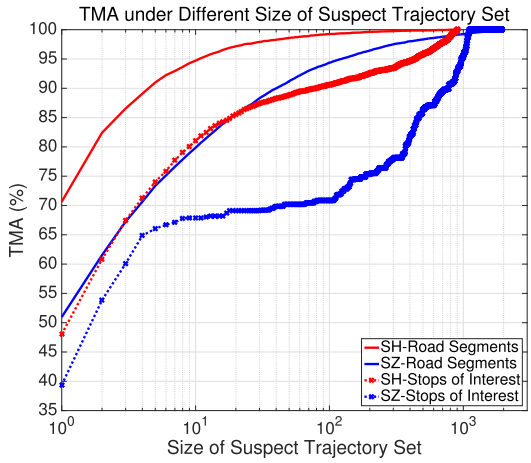


Fig. 9. TMA under different size of suspect trajectory set.

In this experiment, we verify whether utilizing both of them in the strategy can further improve TMA or not. To this end, a hybrid feature vector is obtained by simply concatenating RS and SI vectors of the same object. We also vary the length of compromised trajectories $\widetilde{\mathcal{TR}}_\sigma$ in $\mathcal{T}'$, the setting is the same as Subsection VII-B. Then launch de-anonymization attack using the hybrid feature under different length of compromised trajectories, respectively.

From Figure 8(a) and (b) we can see that, in both Shanghai and Shenzhen datasets, by using 1-day long compromised trajectories, hybrid feature can achieve around 80% and 58% TMAs, which means the TMAs are improved around 9% and 7% in Shanghai and Shenzhen datasets, respectively. As the length of comprised trajectories increasing, TMAs are further improved. Using 8 and 9 days long $\widetilde{\mathcal{TR}}_\sigma$, hybrid feature based TMAs can reach up to 96.64% and 77.03% in Shanghai and Shenzhen datasets, respectively.

VIII. Conclusion

In this work, we focus on the privacy vulnerabilities of anonymized trajectories, whose real identities are substituted by pseudonyms. We proposed a new paradigm of de-anonymization attacks based on mobility patterns of objects. In the attacks, an adversary who obtains a few pieces of trajectories of its victims, can learn mobility features of the victims effectively, and then can use the mobility features to recognize trajectories of the victims from an anonymized dataset available. We conducted experiments with two real-world datasets of mobile traces, and found that road segment preferences and stops of interest of vehicles can serve as quasi-identifiers to distinguish them, and proposed an ITF-IDF value based feature vector construction method by which mobility pattern based de-anonymization attacks can be implemented easily by adversaries. Simulation results demonstrate the effectivity of RS and SI features in re-identifying anonymized trajectories.

References

[1] GeoLife GPS Trajectories. 2011. [Online]. Available: http://research.microsoft.com/en-us/downloads/b16d359d-d164-469e-9fd4-daa38f2b2e13/
[2] Movebank. 2012. [Online]. Available: http://www.movebank.org/

[3] CRAWDAD: A community resource for archiving wireless data at Dartmouth. 2004. [Online]. Available: http://crawdad.cs.dartmouth.edu/
[4] Crowdflow.net. 2015. [Online]. Available: http://crowdflow.net/
[5] M. E. Nergiz, M. Atzori, and Y. Saygin, "Towards trajectory anonymization: A generalization-based approach," in *Proc. ACM Int. Workshop Secur. Privacy GIS LBS*, 2008, pp. 52–61.
[6] C. Y. T. Ma, D. K. Y. Yau, N. K. Yip, and N. S. V. Rao, "Privacy vulnerability of published anonymous mobility traces," *IEEE/ACM Trans. Netw.*, vol. 21, no. 3, pp. 720–733, Jun. 2013.
[7] A. R. Beresford and F. Stajano, "Location privacy in pervasive computing," *IEEE Pervasive Comput.*, vol. 2, no. 1, pp. 46–55, Jan.–Mar. 2003.
[8] M. Gruteser and X. Liu, "Protecting privacy in continuous location-tracking applications," *IEEE Secur. Privacy*, vol. 2, no. 2, pp. 28–34, Mar./Apr. 2004.
[9] J. Krumm, "A survey of computational location privacy," *Pers. Ubiquitous Comput.*, vol. 1, no. 6, pp. 391–399, 2009.
[10] L. Kulik, "Privacy for real-time location-based services," *SIGSPATIAL Special*, vol. 1, no. 2, pp. 9–14, 2009.
[11] L. Sweeney, "*K*-anonymity: A model for protecting privacy," *Int. J. Uncertainty, Fuzziness Knowl.-Based Syst.*, vol. 10, no. 5, pp. 571–588, 2002.
[12] A. Meyerson and R. Williams, "On the complexity of optimal *k*-anonymity," in *Proc. 23rd ACM Symp. Princ. Database Syst.*, Jun. 2004, pp. 223–228.
[13] M. Gruteser and D. Grunwald, "Anonymous usage of location-based services through spatial and temporal cloaking," in *Proc. ACM 1st Int. Conf. Mobile Syst. Appl. Serv.*, May 2003, pp. 31–42.
[14] C. Li and B. Palanisamy, "De-anonymizable location cloaking for privacy-controlled mobile systems," in *Proc. Int. Conf. Netw. Syst. Secur.*, Nov. 2015, pp. 449–458.
[15] Y. Tian, W. Wang, J. Wu, Q. Kou, Z. Song, and E. C.-H. Ngai, "Privacy-preserving social tie discovery based on cloaked human trajectories," *IEEE Trans. Veh. Technol.*, vol. 66, no. 2, pp. 1619–1630, Feb. 2017.
[16] H. Zang and J. Bolot, "Anonymization of location data does not work: A large-scale measurement study," in *Proc. ACM 17th Annu. Int. Conf. Mobile Comput. Netw.*, Sep. 2011, pp. 145–156.
[17] D. Kondor, B. Hashemian, Y-A. Montjoye, and C. Ratti, "Towards matching user mobility traces in large-scale datasets," *IEEE Trans. Big Data*, to be published, doi: 10.1109/TBDATA.2018.2871693.
[18] K. Emara, "Safety-aware location privacy in VANET: Evaluation and comparison," *IEEE Trans. Veh. Technol.*, vol. 66, no. 12, pp. 10718–10731, Dec. 2017.
[19] J. Freudiger, R. Shokri, and J-P. Hubaux, "Evaluating the privacy risk of location-based services," in *Proc. Int. Conf. Financial Cryptography Data Secur.*, Feb. 2011, pp. 31–46.
[20] Z. Xiao, C. Wang, W. Han, and C. Jiang, "Unique on the road: Re-identification of vehicular location-based metadata," in *Proc. Int. Conf. Secur. Privacy Commun. Syst.*, Oct. 2016, pp. 496–513.
[21] A.-M. Olteanu, K. Huguenin, R. Shokri, M. Humbert, and J-P. Hubaux, "Quantifying interdependent privacy risks with location data," *IEEE Trans. Mobile Comput.*, vol. 16, no. 3, pp. 829–842, Mar. 2017.
[22] D. J. Martin, D. Kifer, A. Machanavajjhala, J. Gehrke, and J. Y. Halpern, "Worst-case background knowledge for privacy-preserving data publishing," in *Proc. IEEE 23rd Int. Conf. Data Eng.*, Feb. 2007, pp. 126–135.
[23] A. Basu *et al.*, "A privacy risk model for trajectory data," in *Proc. IFIP Int. Conf. Trust Manage.*, Jul. 2014, pp. 125–140.
[24] Y. Zhao and I. Wagner, "On the strength of privacy metrics for vehicular communication," *IEEE Trans. Mobile Comput.*, to be published, doi: 10.1109/TMC.2018.2830359.
[25] R. Pellungrini, L. Pappalardo, F. Pratesi, and A. Monreale, "Fast estimation of privacy risk in human mobility data," in *Proc. Int. Conf. Comput. Saf., Rel., Secur.*, Sep. 2017, pp. 415–426.
[26] F. Xu, Z. Tu, Y. Li, P. Zhang, X. Fu, and D. Jin, "Trajectory recovery from ash: User privacy is not preserved in aggregated mobility data," in *Proc. Int. World Wide Web Conf. Committee*, Apr. 2017, pp. 1241–1250.
[27] H. Wang, C. Gao, Y. Li, G. Wang, D. Jin, and J. Sun, "De-anonymization of mobility trajectories: Dissecting the gaps between theory and practice," in *Proc. Netw. Distrib. Syst. Secur.*, Feb. 2018, pp. 449–458.
[28] H. Li, H. Zhu, S. Du, X. Liang, and X. Shen, "Privacy leakage of location sharing in mobile social networks: Attacks and defense," *IEEE Trans. Dependable Secure Comput.*, vol. 15, no. 4, pp. 646–660, Jul./Aug. 2018.
[29] Z. Chen, Y. Fu, M. Zhang, Z. Zhang, and H. Li, "The De-anonymization method based on user spatio-temporal mobility trace," in *Proc. Int. Conf. Inf. Commun. Secur.*, Dec. 2017, pp. 459–471.
[30] S. Gambs, M-O. Killijian, and M. N. P. Cortez, "De-anonymization attack on geolocated data," *J. Comput. Syst. Sci.*, vol. 80, no. 8, pp. 1597–1614, 2014.
[31] Y. Lou, C. Zhang, Y. Zheng, X. Xie, W. Wang, and Y. Huang, "Map-matching for low-sampling-rate GPS trajectories," in *Proc. 17th ACM SIGSPATIAL Int. Conf. Adv. Geographic Inf. Syst.*, Nov. 2009, pp. 352–361.
[32] P. J. Rousseeuw and B. C. van Zomeren, "Unmasking multivariate outliers and leverage points," *J. Amer. Statist. Assoc.*, vol. 85, pp. 633–639, 1990.

Shan Chang (M'08) received the Ph.D. degree in computer software and theory from Xián Jiaotong University, Xián, China, in 2013. From 2009 to 2010, she was a Visiting Scholar with the Department of Computer Science and Engineering, Hong Kong University of Science and Technology. From 2010 to 2011, she was also a Visiting Scholar with BBCR Research Laboratory, University of Waterloo. She is currently an Associate Professor with the Department of Computer Science and Technology, Donghua University, Shanghai, China. Her research interests include security and privacy in mobile networks and sensor networks.

Chao Li received the B.S. degree in computer science and technology from Anhui University, Hefei, China, in 2016. He is currently working toward the Postgraduate degree with the Department of Computer Science and Technology, Donghua University, Shanghai, China. His research interests include privacy and distributed system security. He is a member of ACM.

Hongzi Zhu (M'06) received the B.S. and M.S. degrees from Jilin University, Changchun, China, in 2001 and 2004, respectively, and the Ph.D. degree in computer science from Shanghai Jiao Tong University, Shanghai, China, in 2009. He was a Postdoctoral Fellow with the Department of Computer Science and Engineering, Hong Kong University of Science and Technology, and the Department of Electrical and Computer Engineering, University of Waterloo, in 2009 and 2010, respectively. He is currently an Associate Professor with the Department of Computer Science and Engineering, Shanghai Jiao Tong University. His research interests include vehicular networks, mobile computing, and smart computing.

Ting Lu (M'14) received the B.S. degree from the Harbin Institute of Technology, Harbin, China, in 2008, and the Ph.D. degree from Shanghai Jiao Tong University, Shanghai, China, in 2013. From 2013 to 2018, she was an Assistant Professor of computer science and technology with Donghua University, Shanghai, China. She is currently an Associate Professor with Donghua University. Her research interests include wireless networks, mobile computing, network management and control, cloud computing, mobile computing, and big data.

Qiang Li received the B.S., M.S., and Ph.D. degrees from Jilin University, Changchun, China, in 1998, 2001, and 2005, respectively. He is currently a Professor with the Key Laboratory of Symbolic Computation and Knowledge Engineering of Ministry of Education, Jilin University. His main research interests include network security and the detection of malicious code.