Transportation Research Record

Quantifying Privacy Vulnerability Under Linkage Attack Across Multi-source Individual Mobility Data

--Manuscript Draft--

Full Title:	Quantifying Privacy Vulnerability Under Linkage Attack Across Multi-source Individual Mobility Data
Abstract:	With the advances in detector and sensor technologies, identity detection-based intelligent transportation systems---such as license plate recognition (LPR) system and parking electronic toll collection (ETC) system---have been widely deployed in urban transportation, generating large quantities of multi-source individual-based mobility data set (e.g., LPR data and parking data). Given the high frequency, precision and wide coverage, these individual-based mobility data can be used in many transportation research areas, such as transportation planning, traffic prediction and individual mobility pattern profiling. With the increasing demand for publishing and sharing these individual-based data sets to researchers and practitioners, the privacy issue of data publishing has been a major concern since true identities of individuals can be revealed by linkage attack. In this paper, we quantitatively measure the privacy disclosure risk caused by linkage attack across multi-source individual-based mobility data sets. Taking an example of LPR data and parking data, a traffic-knowledge-driven adversary model is proposed for linkage attack conducting among LPR data and parking data. Two common modes of LPR data publishing are examined and two quantitative criteria are introduced to present the risk of privacy leakage under linkage attack. The experimental results demonstrate that anonymized individual still under high risk of being linked successfully (71.63% under mode 1 and 36.55% under mode 2). This study serves as a wake-up call for relevant agencies and data owners about the privacy vulnerability caused by linkage attack across multi-source individual-based mobility data.
Manuscript Classifications:	Data and Information Technology; Urban Tranportation Big Data ABJ30SC; Big Data; Data Analysis; Data Fusion; Statistical Methods ABJ80; Modeling
Manuscript Number:	20-01384
Article Type:	Presentation
Order of Authors:	Jing Gao
	Qinglong Lu
	Ming Cai

QUANTIFYING PRIVACY VULNERABLITY UNDER LINKAGE ATTACK ACROSS MULTI-SOURCE INDIVIDUAL MOBILITY DATA

Jing Gao
Guangdong Provincial Key Laboratory of Intelligent Transportation Systems
Research Center of Intelligent Transportation System
&
School of Intelligent Systems Engineering
Sun Yat-Sen University
Guangzhou, Guangdong 510006, China
Email: gaoj26@gmail.com

Qinglong Lu
Department of Civil, Geo and Environmental Engineering
Technical University of Munich
Arcisstra. 21, 80333 Munich
Email: qinglong.lu@tum.de

Ming Cai, Corresponding Author
Guangdong Provincial Key Laboratory of Intelligent Transportation Systems
Research Center of Intelligent Transportation System
&
School of Intelligent Systems Engineering
Sun Yat-Sen University
Guangzhou, Guangdong 510006, China
Email: caiming@mail.sysu.edu.cn

Word count: 5929 words text + 4 tables x 250 words (each) = 6929 words
Submission Date: August 1, 2018

ABSTRACT

With the advances in detector and sensor technologies, identity detection-based intelligent transportation systems—such as license plate recognition (LPR) system and parking electronic toll collection (ETC) system—have been widely deployed in urban transportation, generating large quantities of multi-source individual-based mobility data set (e.g., LPR data and parking data). Given the high frequency, precision and wide coverage, these individual-based mobility data can be used in many transportation research areas, such as transportation planning, traffic prediction and individual mobility pattern profiling. With the increasing demand for publishing and sharing these individual-based data sets to researchers and practitioners, the privacy issue of data publishing has been a major concern since true identities of individuals can be revealed by linkage attack. In this paper, we quantitatively measure the privacy disclosure risk caused by linkage attack across multi-source individual-based mobility data sets. Taking an example of LPR data and parking data, a traffic-knowledge-driven adversary model is proposed for linkage attack conducting among LPR data and parking data. Two common modes of LPR data publishing are examined and two quantitative criteria are introduced to present the risk of privacy leakage under linkage attack. The experimental results demonstrate that anonymized individual still under high risk of being linked successfully (71.63% under mode 1 and 36.55% under mode 2). This study serves as a wake-up call for relevant agencies and data owners about the privacy vulnerability caused by linkage attack across multi-source individual-based mobility data.

Keywords: Individual-based mobility data, Linkage attack, License plate recognition data, Parking data, Disclosure risk

INTRODUCTION

With the advances in mobile sensing technologies, individual behaviors are widely captured and recorded. Large quantities of individual-based mobility data sets, such as license plate recognition (LPR) data and parking data, have been widely generated and collected. License plate recognition (LPR) data, generated by the LPR system which takes pictures of every passing vehicle and convert images to detailed spatiotemporal records automatically, capturing vehicle history trajectory with high precision and wide coverage. Given the high frequency, great precision, and extensive coverage, LPR data, have been applied to a wide range of transportation research areas, for example, transportation planning, traffic prediction and individual mobility pattern recognition. With the wide usage of LPR data, the focus has been increasingly centered on the privacy issue of publishing this kind of data. Previous studies adopted the concept of anonymity (*1*) and examined that individuals in anonymized LPR data set still have a high risk of being re-identified (*2, 3*). Specifically, Gao et al. (*2*) selected several spatiotemporal records in one's history trajectory to form his quasi-identifier and the anonymity of an individual was defined as the number of occurrence of the quasi-identifier in all individual's history trajectory. An anonymized individual is called re-identified if his anonymity is one, i.e., his quasi-identifier is only contained in his own trajectory. However, an individual with his anonymity equaling 1 doesn't mean that his true identity (vehicle plate number in our case) would be revealed. To get the true identity of the target for obtaining his privacy information, the further linkage should be made between anonymized data and external information (with true identity), which is called linkage attack.

This paper aims at quantifying the privacy disclosure risk under the linkage attack across multi-source individual mobility data. Taking an example of LPR data and parking data, we want to examine to what extent the true identity of anonymized individuals can be revealed by linkage attack between anonymized LPR database and parking database. In doing so, we propose a traffic-knowledge-driven adversary model to conduct linkage attack. Two common modes of LPR data publishing are investigated and two quantitative criteria are introduced to measure the privacy disclosure risk under linkage attack. To the best of our knowledge, this is the first study that empirically quantifies the privacy risk of LPR data by conducting linkage attack among LPR data set and parking data set from urban transportation system.

The remainder of this paper is structured as follows. Section 2 reviews the related research in privacy risk measurement and privacy protection in individual-based mobility data. Section 3 introduces the LPR data set and the parking data set. In addition, two common modes of LPR data publishing are introduced section 3. Then, in section 4, we propose the traffic-knowledge-driven adversary model concerning two different modes of LPR data publishing and empirical experiments and results analysis are presented in section 5. Finally, Section 6 presents some concluding remarks and proposes future work of this study.

LITERATURE REVIEW

With the emerging data in the traffic area and more and more data are available to the public, the latent insecurity of privacy has attracted the attention of many researchers. To prevent attackers from extracting private information from anonymized database, data suppression techniques are presented in (*4*) to reduce such reidentification risks. Concerning the privacy protection of Location Based Service (LBS) users, (*5*) investigate an adversary model to attack the LBS data with long-term pseudonyms and uncover insecurity of users privacy under this releasing pattern. Li et al. (*6*) establish a traffic monitoring scheme to preserve driver privacy by acquiring driving data and

integrating it with a weighted proximity graph to filter out false reports uploaded by malicious drivers to intervene the storage system.

In (*7*), 3 months of credit card records are investigated and the result shows that by only leveraging four spatiotemporal points 90% of individuals could be re-identified which illustrates a very serve problem in the metadata. (*8*) indicates that 10 pieces of side information of a victim are enough to identify the trace of (s)he by a probability of 30% to 50%. Factors affecting anonymity are summarized in (*9*), and concerning these factors, two typical techniques for improving data security are discussed, i.e. suppression and generalization. To evaluate the effectiveness of the techniques, an adversary model is proposed to conduct attacks. These studies, however, at an angle of revealing data security at a point level, have not noticed the potential consequence of linkage attack.

To reach a trade-off both personal privacy and traffic data utility are at a comparatively high level, some studies have proposed valid systems for different applications. (*10*) built a decentralized architecture under which the location data of ridesharing users will be released in such a way that guarantees the security of user privacy without sacrificing the service quality. In (*11*), to protect the privacy of participants of probe vehicles who have to reveal their GPS positions to the traffic monitoring department, a VTLs-based system cooperating with an associated cloaking technique is provided. However, this system cannot protect traffic monitoring accuracy from any active attack. Sun et al. (*12*) construct a VTLs zone-based system to balance the data needs for general traffic research and privacy security, and to assess the effectiveness of this system, traffic-knowledge-based adversary models are proposed to conduct privacy attack.

Those systems and architectures are always focused on the privacy problem from one single data source, while in reality, different sources of traffic data are mixed and have a strong relationship with each other. To this end, a traffic-knowlwdge-driven adversary model is proposed for conducting linkage attack across multi-source individual-based mobility data. And the privacy disclosure risk under linkage attack among LPR data and parking data is quantitatively measured.

DATA DESCRIPTION

License plate recognition data

We use LPR data collected by LPR system in Guangzhou, China on 18th July, 2018. The LPR system is essentially a network of cameras that can take pictures of every passing vehicle and transform the image into a detailed spatiotemporal record automatically. The information in a single LPR record includes license plate number, color or the license plate, timestamp when vehicle capture by the detector, detector ID (representing different camera gantries) and driving direction of the vehicle. Table 1 lists the primary fields from an example LPR record. The trajectories of individuals can be tracked by summarizing a series of spatiotemporal records.

Considering the great advantage and increasing demand of LPR data, more and more agencies choose to publish and share the LPR data set for transportation research purpose. Taking into account the privacy concerns and issues in LPR data, just like the publication of other source of trajectory data set, data owners usually need to construct an anonymized LPR database for data publishing. Generally, there are two common alternative modes.

TABLE 1 : Primary fields in original LPR data set

Field	Example value
Vehicle plate number	@ABC123
Passing timestamp	2018-07-18 10:59:07 (one-second resolution)
Address	Inner Ring Road, Meizhou Building (East to West)
Detector ID	17068
Drive direction	0

Mode 1: full trajectory publishing

The first one is the most common LPR data publishing method. When LPR data is published in mode 1, the vehicle plate number of each vehicle will be replaced with a unique random identifier. Each record in the anonymized LPR database is a sequence of spatiotemporal tuples (in the form of $(T_i, LocR_i)$), which represents the full history trajectory of an individual/vehicle. In other words, under the data publishing of mode 1, complete history trajectories with its corresponding random identifiers will be published. Table 2 shows an example anonymized LPR database published in mode one.

TABLE 2 : Anonymized LPR database published in mode one

Rec_ID. #	Record value
5ac0bd6239f8b9a	(288,17062) $\rightarrow$ (302,17067) $\rightarrow$ (71898,14688) $\rightarrow$ (72058,16800)
5ac0bd6239f8b9b	(308,17063) $\rightarrow$ (512,17067) $\rightarrow$ (698,17069)
...	...

Mode 2: segmented trajectory publishing

Compared with mode 1 in which complete vehicle trajectories are published directly, mode 2 splits each individual's complete trajectory into segments according to parking activity and allocate a random identifier to each trajectory segment. Table 3 shows the example anonymized LPR database published in mode 2. Since individuals usually park several times for specific activities (e.g, shopping or working) during a day, it is quite common that the full history trajectory actually consists of several segmented trajectories. Consequently, spatio-temporal information within each O-to-D trip is revealed.

TABLE 3 : Anonymized LPR database published in mode two

Rec_ID. #	Record value
5ac0bd6239f8b9a1	(288,17062) $\rightarrow$ (302,17067)
5ac0bd6239f8b9a2	(71898,14688) $\rightarrow$ (72058,16800)
5ac0bd6239f8b9b	(308,17063) $\rightarrow$ (512,17067) $\rightarrow$ (698,17069)
...	...

Parking data

The parking data we used was mainly collected by the parking toll collection system in Guangzhou, China on 18th July, 2018. Similar to the LPR system, the camera detectors at the parking gate take pictures of entering/leaving vehicles to get the vehicle plate number, the entry time and the departure time of vehicles. Table 4 lists the primary fields from an example parking record. It is quite natural that the parking actions arise from specific activity (e.g., commuting or shopping) and the parking data records the spatiotemporal information of origin or destination in a trip.

Since the parking lots are usually constructed and managed by the parking owners, the parking data set is consequently an accessible database to them. In addition, to get the parking data set, one only need to observe and record the access of vehicles at the parking gates. As a consequence, the parking database can be regarded as a public external database for linkage attack.

TABLE 4 : Primary fields in original Parking data set

Field	Example value
Parking ID	etcp_788
Parking name	GOGO Xintiandi
Address	No. 1 Middle Road, Xiaoguwei Street
Gate name	No. 1 entrance
Vehicle plate number	@ABC123
Intime	2018-07-18 11:08:07
Outtime	2018-07-18 11:32:07

So given the anonymized LPR database (without true identity of vehicle) published in two different modes and the external public parking database (with actual vehicle plate number), the adversary aim at inferring the true identity of records in anonymized LPR database by linking to parking database. Figure 1 shows a simple road network with detectors and parking lots, where the characters represents the locations of camera detectors and the serial numbers represent the locations of parking lots. Suppose an individual drove form a to c, then parked in the parking ③ for shopping nearby. After shopping, he left the parking and drove from e to f. So the goal of an adversary is to link the parking record in parking ③ to the LPR record (e.g., a $\rightarrow$ b $\rightarrow$ c $\rightarrow$ e $\rightarrow$ d $\rightarrow$ f in mode 1) in anonymized LPR database to obtain the vehicle plate number of the individual.

TRAFFIC-KNOWLEDGE-DRIVEN ADVERSARY MODEL

In this section, we propose a traffic-knowledge-driven adversary model to quantify the privacy disclosure risk under two different modes of publishing LPR data set. Specifically, the adversary model takes into consideration both the temporal (real-time travel time information) and spatial (spatial connectivity) relation to conduct linkage attack between LPR data (anonymized database) and parking data (public external database). The target of our model is to infer the true identities of records, i.e, the vehicle plate number, in the LPR database by linking to parking data. Two different modes of LPR data publishing proposed in Section **??** are examined.

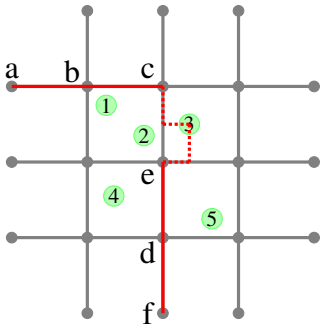

FIGURE 1 : Schematic illustration of road network with camera detectors and parking

Linking under full trajectory publishing (Mode 1)

As presented in Section 3, in LPR data publishing of mode 1, LPR data is anonymized by replacing vehicle plate number of each individual with a unique identifier and full history trajectory of each individual is published. Under this circumstance, each record in an anonymized LPR database represents the complete history trajectory of an individual and complete spatiotemporal information (a sequence of spatiotemporal tuples) of individuals is released. Given the parking database, we aim to link one LPR record to each parking data record to re-identify the true identity (vehicle plate number) of that anonymized LPR record.

Given a vehicle l with its parking record P_k^l in the parking database P_k of parking k ($k = 1,\ldots,m$), it can be naturally imagined that this vehicle passed through a detector, and then it entered the parking lot k. After that, it left the parking and captured by another detector. In other words, the parking record can be inserted into a pair of temporal consecutive spatiotemporal tuples in one's trajectory, i.e., a data record in the anonymized LPR database. Consider an individual j with an anonymized LPR record R^j, if he used to park in the parking k and since generated the parking data record P_k^l, without loss of generality, we suppose that parking occurred between i-th tuple and $(i+1)$-th tuple in R^j (denoted as $R_i^j = (T_i^j, LocR_i^j)$ and $R_{i+1}^j = (T_{i+1}^j, LocR_{i+1}^j)$, respectively), then we can estimate the timestamp of R_i^j (denoted as T_i^j) as the following Equation 1.

$$\hat{T}_i^j \approx T_{P_k^l}^{in} - t_{i,k}^j \tag{1}$$

Here $t_{i,k}^j$ represents the estimated travel time for the linkage between parking record P_k^l and the spatiotemporal tuple R_i^j, which can be calculated by the distance from $LocR_i^j$ to the location of parking k (indicated as $LocP_k$) divided by the estimated average speed in this route. To improve the precision of travel time estimation, we request traffic information using navigation API, by which we can get the distance of planned routes and the navigation duration of routes. Since traffic flow patterns are usually similar in neighboring links and over the same period, we approximate the average speed of vehicle j in the target route using the average speed in the previous link it traveled along, which can be calculated by the earlier pair of spatiotemporal tuples in R^j as $Dist_{i-1,i}^j/(T_i^j - T_{i-1}^j)$, where $Dist_{i-1,i}^j$ is the requested route distance from $LocR_{i-1}^j$ to $LocR_i^j$. This thus allows us to capture the real-time traffic dynamics (e.g., congestion status or free-flow status) and intrinsic driving characteristics of vehicles (e.g., heavy truck or private car), which serves as a key factor for the travel time estimation in the linkage process. Then t_{i,P_k}^j is calculated

by $Dist^j_{i,k}/V^j_{i,k}$, where $Dist^j_{i,k}$ is the requested route distance from $LocR^j_i$ to $LocP_k$. In addition, taking into account the existence of abnormal estimated travel time arising from traffic anomaly (e.g., traffic accidents, road constructions and aggressive driving, etc.), we calibrate the estimated travel time using the requested navigation duration from $LocR^j_i$ to $LocP_k$ (indicated as $Dura^j_{i,k}$) as equation 2.

$$t^j_{i,k} = \begin{cases} \frac{Dist^j_{i,k}}{V^j_{i,k}}, & \text{if } t^j_{i,k} \in [(1-\alpha)Dura^j_{i,k}, (1+\alpha)Dura^j_{i,k}]; \\ Dura^j_{i,k}, & \text{if } t^j_{i,k} \notin [(1-\alpha)Dura^j_{i,k}, (1+\alpha)Dura^j_{i,k}]. \end{cases} \tag{2}$$

Similarly, the travel speed of vehicle j on the route between $LocP_k$ and $LocR^j_{i+1}$ can be approximated by the calculated average speed on the later link between $LocR^j_{i+1}$ and $LocR^j_{i+2}$, and the estimated travel time on target route can be calibrated by the requested duration $Dura^j_{k,i+1}$ on it. Then the travel time of vehicle j from parking k to the following detector can be estimated by equation 3. If vehicle l and vehicle j are the same individual, then the timestamp that this vehicle went through the detector (i.e., T^j_{i+1}) can be approximated by $T^{out}_{P^l_k} + t^j_{k,i+1}$. See equation 4.

$$t^j_{k,i+1} = \begin{cases} \frac{Dist^j_{k,i+1}}{V^j_{k,i+1}}, & \text{if } t^j_{k,i+1} \in [(1-\alpha)Dura^j_{k,i+1}, (1+\alpha)Dura^j_{k,i+1}]; \\ Dura^j_{k,i+1}, & \text{if } t^j_{k,i+1} \notin [(1-\alpha)Dura^j_{k,i+1}, (1+\alpha)Dura^j_{k,i+1}]. \end{cases} \tag{3}$$

$$\hat{T}^j_{i+1} \approx T^{out}_{P^l_k} + t^j_{k,i+1} \tag{4}$$

Adding a fluctuation of the estimated travel time, if the vehicle j leaved the previous detector within the period $[\hat{T}^j_i - T^{i,j}_f, \hat{T}^j_i + T^{i,j}_f]$ and arrived at the following detector within the period $[\hat{T}^j_{i+1} - T^{i+1,j}_f, \hat{T}^j_{i+1} + T^{i+1,j}_f]$, then the parking record P^l_k may be linked to the pair of spatiotemporal tuples R^j_i and R^j_{i+1} in the anonymized LPR record R^j and accordingly they may be the same vehicle. Differing from a constant time threshold was set for the fluctuation in (*12*), we allow the estimated travel time fluctuate within a proportion range, i.e, $T^{i,j}_f = \beta \hat{T}^j_i$, which can screen more reasonable estimated travel time. For conducting linkage attack, if two continuous tuples $R^j_i = (T^j_i, LocR^j_i)$ and $R^j_{i+1} = (T^j_{i+1}, LocR^j_{i+1})$ in one's LPR record R^j satisfy $(1-\beta)\hat{T}^j_i \le T^j_i \le (1+\beta)\hat{T}^j_i$ and $(1-\beta)\hat{T}^j_{i+1} \le T^j_{i+1} \le (1+\beta)\hat{T}^j_{i+1}$ simultaneously, (i, j) is added into a candidate set for parking record P^l_k (denoted as C^l_k). If C^l_k is not empty, then a optimal match between R^j_i, R^j_{i+1} and P^l_k can be selected by equation 5. Equation 5 means that the linkage result of our adversary model is the one with the minimal total time difference between the actual passing timestamps and estimated passing timestamps at the previous detector and the following detector. If the match is correct, i.e., $R^{\hat{j}}_{\hat{i}}$, $R^{\hat{j}}_{\hat{i}+1}$ and P^l_k indeed belong to the same vehicle, the parking record and anonymized LPR record thus are successfully linked, and the true identity of the anonymized record in LPR database, i.e., the vehicle plate number, is successfully inferred, which poses a threat to the individual's privacy. Applying the model to conduct linkage attack for each parking record in parking database, we can finally get the match set of the parking database.

$$\hat{i},\hat{j} = \underset{(i,j)\in C_k^l}{\arg\min} |T_i^j - \hat{T}_i^j| + |T_{i+1}^j - \hat{T}_{i+1}^j| \tag{5}$$

1 **Linking under segmented trajectory publishing (Mode 2)**

In mode 2, traffic data owner separates the complete trajectory to several segments at the nodes where the vehicle is in parking state, and each segment will have a unique identifier. To re-identify the vehicle by linkage attack, for a parking record (e.g. P_k^l), two trajectory segments, e.g. R^j and $R^{j'}$, will be taken into account if between which that parking record can be inserted into, mathematically, this condition can be formulated as Equation 6.

$$T_{end}^j \leq T_{P_k^l}^{in} \leq T_{P_k^l}^{out} \leq T_{st}^{j'} \tag{6}$$

where T_{end}^j and $T_{st}^{j'}$ indicates the passing timestamp of the last tuple of R^j and the first tuple of $R^{j'}$ separately, and $j \neq j'$. For better expound in the following content, variables with the subscription '*end*' are for the last tuple of a segment, while variables with the subscription '*st*' are for the first tuple of a segment. The adversary model proposed in section 4.1 can be applied to link the parking data to the LPR data released through mode 2 as well. But instead of applying the adversary model to each node, combined nodes, which are defined as the nodes formed by the last tuple of a segment and the first tuple of another segment, e.g. $(R_{end}^j, R_{st}^{j'})$, will be the only attacking objectives, as the parking behavior can only happen in the combined nodes. However, when we have a large LPR dataset, the number of combination of LPR records that can satisfy Equation 6 will be enormous, requiring a stricter restriction. To this end, a threshold for filtering the records swarm into the algorithm is introduced, see Equation 7. Equation 7 reduces the number of possible LPR records for constructing combined nodes by restricting the timestamps of the pre-parking tuple and the pos-parking tuple in a range of the timestamp estimated by leveraging the requested navigation duration. γ_t is a time-unit filtering parameter.

$$\begin{aligned} T_{end}^j &\in [T_{P_k^l}^{in} - (Dura_{end,k}^j + \gamma_t), T_{P_k^l}^{in} - (Dura_{end,k}^j - \gamma_t)] \\ T_{st}^{j'} &\in [T_{P_k^l}^{out} + (Dura_{k,st}^{j'} - \gamma_t), T_{P_k^l}^{out} + (Dura_{k,st}^{j'} + \gamma_t)] \end{aligned} \tag{7}$$

For vehicles which do not have parking records, their complete trajectory will be released directly like mode 1. So in mode 2, the timestamp of the pre-parking tuple and the pos-parking tuple can be estimated by Equation 8.

$$\hat{T}_{end}^j \approx T_{P_k^l}^{in} - t_{end,k}^j \tag{8a}$$

$$\hat{T}_{st}^{j'} \approx T_{P_k^l}^{out} + t_{k,st}^{j'} \tag{8b}$$

And similar to the description in section 4.1, $t^{j}_{end,k}$ and $t^{j}_{k,st}$ can be estimated as Equation 9.

$$t^{j}_{end,k} = \begin{cases} \frac{Dist^{j}_{end,k}}{V^{j}_{end,k}}, & \text{if } t^{j}_{end,k} \in [(1-\alpha)Dura^{j}_{end,k}, (1+\alpha)Dura^{j}_{end,k}]; \\ Dura^{j}_{end,k}, & \text{if } t^{j}_{end,k} \notin [(1-\alpha)Dura^{j}_{end,k}, (1+\alpha)Dura^{j}_{end,k}]. \end{cases} \tag{9a}$$

$$t^{j'}_{k,st} = \begin{cases} \frac{Dist^{j'}_{k,st}}{V^{j'}_{k,st}}, & \text{if } t^{j'}_{k,st} \in [(1-\alpha)Dura^{j'}_{k,st}, (1+\alpha)Dura^{j'}_{k,st}]; \\ Dura^{j'}_{k,st}, & \text{if } t^{j'}_{k,st} \notin [(1-\alpha)Dura^{j'}_{k,st}, (1+\alpha)Dura^{j'}_{k,st}]. \end{cases} \tag{9b}$$

And $V^{j}_{end,k}$ and $V^{j'}_{k,st}$ can be calculated as Equation 10.

$$V^{j}_{end,k} = \frac{Dist^{j}_{end-1,end}}{T^{j}_{end} - T^{j}_{end-1}} \tag{10a}$$

$$V^{j'}_{k,st} = \frac{Dist^{j'}_{st,st+1}}{T^{j'}_{st+1} - T^{j'}_{st}} \tag{10b}$$

Similarly, we set a fluctuation range for the estimated travel time, for the LPR record R^{j} which can meet $(1-\beta)\hat{T}^{j}_{end} \leq T^{j}_{end} \leq (1+\beta)\hat{T}^{j}_{end}$ and $R^{j'}$ which can meet $(1-\beta)\hat{T}^{j'}_{st} \leq T^{j'}_{st} \leq (1+\beta)\hat{T}^{j'}_{st}$, (j, j') will become a candidate of the parking record R^{l}_{k} and be added into C^{l}_{k}. The optimum will be selected by Equation 11. It means the candidate whose passing timestamps result into the smallest difference from the estimated timestamps will be the output of the adversary model. If j, j' and l are equal, it will be counted as a success.

$$\hat{j}, \hat{j}' = \underset{(j,j') \in C^{l}_{k}}{\arg\min} |T^{j}_{end} - \hat{T}^{j}_{end}| + |T^{j'}_{i+1} - \hat{T}^{j'}_{i+1}| \tag{11}$$

EXPERIMENT AND RESULT ANALYSIS

In this section, we quantitatively measure the privacy vulnerability due to linkage attack among LPR data and parking data by applying the proposed adversary model to an actual scenario.Two different modes of LPR data publishing are examined and the corresponding results are analyzed.

The linkage attack was conducted in a scenario with five parking lots. Due to the wide coverage of camera detector in the LPR system, the accessible detectors of parking lots are not far away from them. Consequently, to improve the efficiency of the adversary model, we only consider detectors within a 5-km radius of one parking as its latent previous detectors and following detectors. In this case, 24 detectors are selected and the network geometry of the detectors and parking lots is shown in Figure 3. The parking data set and LPR data set were collected on the same day (July 18th, 2018). We construct the parking database by selecting the parking records of the five parking lots in our experiment area. After data cleaning (e.g., drop duplicated records), there are 145 parking records in the parking database. As for the LPR database, we select all history trajectory of vehicles that passed through the one or more of the latent 24 detectors. Then for the data publishing of mode 1, we replace each vehicle plate number with a unique random identifier to construct the anonymized LPR database. When it comes to mode 2, the trajectory of

vehicles that parked in one of the five parking lots is first segmented according to the actual parking record in parking database. Then a unique random identifier is allocated to each trajectory segment for data anonymization in mode 2.

Applying the adversary model proposed in section 4, we conduct linkage attack for each parking record in parking database under two different patterns of LPR data publishing. We set the calibration rate α to 0.35 and the fluctuation rate β to 0.4, respectively. For mode 1, we get a match set of 141 matched records, with 101 records successfully linked (i.e., the target parking record and the matched anonymized LPR record belong to the same vehicle) and 40 records incorrectly matched. As for mode 2, we get a match set with all 145 parking records matched. There are 53 records successfully linked to anonymized LPR records. To quantify the privacy risk more comprehensively, we introduce two classical criteria in the field of information retrieval, namely precision and recall. We define precision in our linkage attack scenario as the number of parking records successfully matched in our match set divided by the total number of match results in our match set (i.e., the size of match set). Recall is defined as the number of parking records successfully matched divided by the total number of parking records in parking database. Then the precision and recall under mode 1 are $101/141 = 71.63\%$ and $101/145 = 69.66\%$, respectively. The precision and recall under mode 2 are $53/145 = 36.55\%$ and $53/145 = 36.55\%$, respectively. We can see from the results that individual with anonymized LPR data published in mode 1 have a high probability of being re-identified by our adversary model (over 70% parking records in match set can be linked successfully to anonymized LPR database). In addition, the privacy disclosure risk drops significantly by almost a half (from 71.3% to 36.55%) from mode 1 to mode 2, which provide a practical solution for privacy preservation in LPR data publishing. This is because in mode 1, our model conducts linkage attack between two consecutive spatiotemporal points in one's trajectory. The linkage attack is successful when the matched anonymized LPR record and target parking record belong to the same vehicle. While in mode 2, we need to link the target parking record to two trajectory segments, i.e., two anonymized records in the published database. And only when these three records belong to the same individual, the linkage attack is judged successful.

Besides selecting a single candidate (one LPR record under mode 1 and two under mode 2) from the candidate set for each target parking record, an adversary may choose to output multiple candidates to improve the probability of linking successfully to the target parking record. Instead of selecting the candidate with minimal time error, we output the top 3 candidates with minimal error in the candidate set to our match set and compare it with our original results (denoted as top 1). With the selection of top 3 candidates, in mode 1, the precision and recall are $130/257 = 50.58\%$ and $130/145 = 89.66\%$, respectively. In mode 2, the precision and recall are $98/362 = 27.07\%$ and $98/145 = 67.59\%$, respectively. The overall results are compared in Figure 2. It is clear from the bar graph that, when top 3 candidates are selected, more parking records in parking database are linked successfully. But the precision decline since more candidates are added into match set. This means that an adversary can get more individual re-identified at the cost of precision.

CONCLUSION AND DISCUSSION

In this paper, we quantitatively measure the privacy disclosure risk under linkage attack across multi-source individual-based mobility data sets. Performing a case study on LPR data and parking data, we propose a traffic-knowledge-driven adversary model for linkage attack conducting among LPR data and parking data. Two common modes of LPR data publishing/anonymization

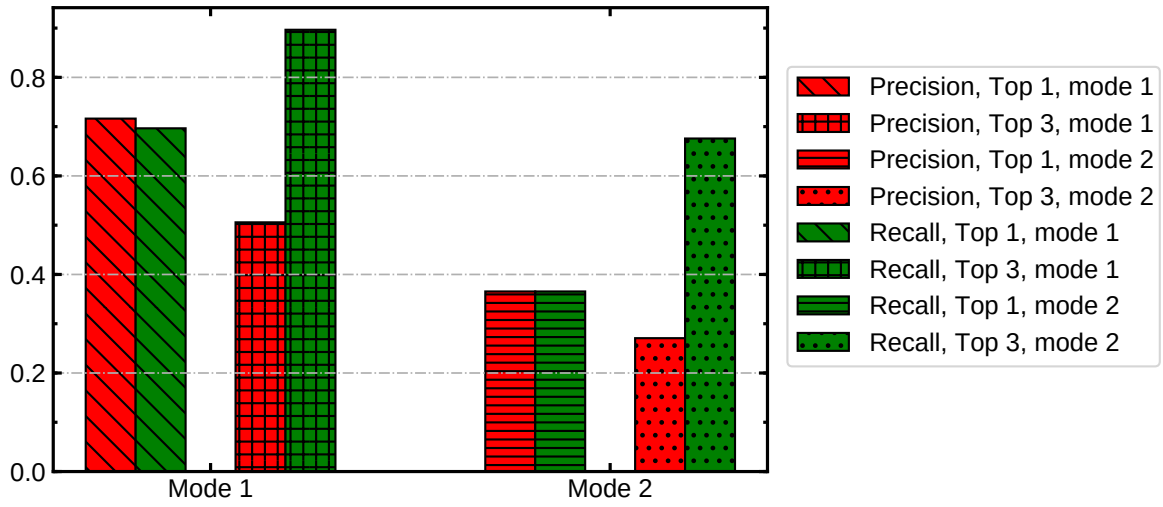

FIGURE 2 : Detectors and parkings considered in the scenario

FIGURE 3 : Detectors and parkings considered in the scenario

are examined by applying the proposed model. Two quantitative evaluation criteria, precision and recall are introduced to investigate the privacy disclosure risk under linkage attack. Empirical experiments are conducted in actual scenario in Guangzhou, China. The result shows that when LPR data is published in mode 1, our adversary model can achieve a precision of 71.61% and a recall of 69.66% which illustrates that individuals in anonymized LPR database still face a high probability of being re-identified. However, in mode 2, the precision and recall decline to 36.55%. This means that LPR data published in mode 2 has a high level of privacy-preserving. When output top 3 candidates, the recall of the adversary model in mode 1 increases to 89.66%, and that of mode 2 increases to 67.59%, respectively. It means, when top 3 candidates are selected, there will have more parking records been linked successfully to anonymized LPR records.

This current work focuses on revealing the great privacy vulnerability caused by linkage

attack across multi-source individual-based mobility data sets from a quantitative perspective. Future extensions will explore the data utility in specific transportation application under two different modes of LPR data publishing and discuss the privacy-and-utility trace-off of different data publishing modes. In addition, we will investigate possible solutions to preserve privacy against linkage attack across multi-source individual mobility data (e.g., reducing the probability of being linked successfully). We hope this work could stimulate more discussion to address the privacy issues in the context of multi-source individual-based mobility data sets.

ACKNOWLEDGEMENT

This research is mainly supported by the National Natural Science Foundation of China (No. 11574407) and the Science and Technology Planning Project of Guangzhou City, China (No. 201704020142).

AUTHOR CONTRIBUTIONS

J.G., Q.L. and M.C. designed the research; J.G. and Q.L. performed the research; J.G. and Q.L. analyzed the data; J.G., Q.L. and M.C. wrote the paper.

REFERENCES

[1] Sweeney, L., k -anonymity: a model for protecting privacy. *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems*, Vol. 10, No. 05, 2002, pp. 557–570.

[2] Gao, J., L. Sun, and M. Cai, Measuring privacy vulnerability of individual mobility traces: a case study on license plate recognition data. In *Transportation Research Board (TRB) Annual Meeting*, 2019.

[3] Gao, J., L. Sun, and M. Cai, Quantifying privacy vulnerability of individual mobility traces: A case study of license plate recognition data. *Transportation Research Part C Emerging Technologies*, Vol. 104, 2019, pp. 78–94.

[4] Hoh, B., M. Gruteser, H. Xiong, and A. Alrabady, Enhancing security and privacy in traffic-monitoring systems. *IEEE Pervasive Computing*, Vol. 5, No. 4, 2006, pp. 38–46.

[5] Liu, X., H. Zhao, M. Pan, H. Yue, X. Li, and Y. Fang, Traffic-aware multiple mix zone placement for protecting location privacy. In *2012 Proceedings IEEE INFOCOM*, IEEE, 2012, pp. 972–980.

[6] Li, M., L. Zhu, and X. Lin, Privacy-Preserving Traffic Monitoring with False Report Filtering via Fog-assisted Vehicular Crowdsensing. *IEEE Transactions on Services Computing*, 2019.

[7] de Montjoye, Y. A., L. Radaelli, V. K. Singh, and A. S. Pentland, Identity and privacy. Unique in the shopping mall: on the reidentifiability of credit card metadata. *Science*, Vol. 347, No. 6221, 2015, pp. 536–9.

[8] Ma, C. Y., D. K. Yau, N. K. Yip, and N. S. Rao, Privacy vulnerability of published anonymous mobility traces. *IEEE/ACM transactions on networking (TON)*, Vol. 21, No. 3, 2013, pp. 720–733.

[9] Gao, J., L. Sun, and M. Cai, Quantifying privacy vulnerability of individual mobility traces: A case study of license plate recognition data. *Transportation Research Part C: Emerging Technologies*, Vol. 104, 2019, pp. 78–94.

[10] Aïvodji, U. M., S. Gambs, M.-J. Huguet, and M.-O. Killijian, Meeting points in ridesharing: A privacy-preserving approach. *Transportation Research Part C: Emerging Technologies*, Vol. 72, 2016, pp. 239–253.

[11] Hoh, B., M. Gruteser, R. Herring, J. Ban, D. Work, J.-C. Herrera, A. M. Bayen, M. Annavaram, and Q. Jacobson, Virtual trip lines for distributed privacy-preserving traffic monitoring. In *Proceedings of the 6th international conference on Mobile systems, applications, and services*, ACM, 2008, pp. 15–28.

[12] Sun, Z., B. Zan, X. J. Ban, and M. Gruteser, Privacy protection method for fine-grained urban traffic modeling using mobile sensors. *Transportation Research Part B: Methodological*, Vol. 56, 2013, pp. 50–69.