Transportation Research Part C 104 (2019) 78–94

Contents lists available at ScienceDirect

Transportation Research Part C

journal homepage: www.elsevier.com/locate/trc

Quantifying privacy vulnerability of individual mobility traces: A case study of license plate recognition data

Jing Gao[a], Lijun Sun[b], Ming Cai[a,*]

[a] School of Intelligent Systems Engineering, Sun Yat-sen University, Shenzhen, Guangdong 518107, China
[b] Department of Civil Engineering and Applied Mechanics, McGill University, Montreal, Quebec H3A 0C3, Canada

ARTICLE INFO

Keywords:
Mobility traces
Privacy protection
License plate recognition data
Disclosure risk
k-anonymity
Data utility

ABSTRACT

Emerging smart transportation applications are calling for publishing and sharing individual-based mobility trace data sets to researchers and practitioners; in the meanwhile, however, privacy issues have become a major concern given that true identities of individuals can be easily revealed from these data sets. Data synthesis In this paper, we quantitatively measure the risk of privacy disclosure in mobility trace data set caused by re-identification attacks based on the concept of *k*-anonymity. Using a one-month license plate recognition (LPR) data set collected in Guangzhou, China, we examine a variety of factors determining the degree of anonymity of an individual, including the temporal granularity and the size of the published data, local v.s. non-local vehicles, and continuous v.s. non-continuous observations. We find that five spatiotemporal records are enough to uniquely identify about 90% of individuals, even when the temporal granularity is set to be half a day. To publish LPR data without compromising privacy, we propose a suppression solution and a generalization solution and quantify the privacy-and-utility trade-off of them. Our results show that the suppression solution, which removes sensitive records, have a notable performance on privacy protection. The average individual anonymity identified by three spatiotemporal records increases by more than 20% at the cost of losing less than 8% of the data. We also propose a bintree-based adaptive time interval cloaking algorithm as a generalization solution. To meet a specific anonymity constraint, this algorithm adjusts the temporal resolution adaptively based on traffic counts under the principle of minimal information loss. We find that the generalization algorithm performs extremely well in satisfying different user-specified anonymity constraints and it is more flexible and reliable than the traditional uniform time interval cloaking method. We also find a strong correlation between the resulting temporal accuracy of data anonymized by the algorithm and the traffic condition. This study serves as a reminder to relevant agencies and data owners about the privacy vulnerability in individual-based mobility trace data sets and provides methodological guidance when publishing and sharing such sensitive data set.

1. Introduction

With the advances in sensor technology, identity detection-based intelligent transportation systems—such as license plate recognition (LPR)—have become widely applied in urban transportation, generating large quantities of individual-based mobility trace data set (Herrera et al., 2010). Given the high frequency, great precision, and extensive coverage, these mobility trace data sets have

* Corresponding author.
E-mail addresses: lijun.sun@mcgill.ca (L. Sun), caiming@mail.sysu.edu.cn (M. Cai).

https://doi.org/10.1016/j.trc.2019.04.022
Received 29 September 2018; Received in revised form 18 April 2019; Accepted 23 April 2019
Available online 07 May 2019

provided new opportunities to a wide range of transportation research questions, from transportation planning, to traffic control and management, to individual mobility pattern profiling. Essentially, the LPR system is a network of cameras that can take pictures of every passing vehicle and transform the image into a detailed spatiotemporal record automatically, capturing vehicle data in real-time with high precision and wide coverage. The information in an LPR record includes license plate number (vehicle ID), the timestamp when vehicle captured by the camera, detector ID (representing different camera gantries), and driving direction of the vehicle. Therefore, LPR system collects the location information of each vehicle and it allows us to reconstruct the trajectories of each individual vehicle by summarizing a series of spatiotemporal records. Apart from individual mobility, the LPR also play an essential role in estimating aggregated traffic flow variables (e.g., flow, speed, and density) at a network scale, which is critical to a wide range of intelligent transportation system (ITS) applications such as travel demand prediction, trip planning, travel time estimation, route planning, ride sharing, transit service scheduling, signal control, and disruption management. Due to these advantages, the system has attracted considerable attention in the past few years (see e.g. Shao and Chen, 2018; Zhan et al., 2015; Chen et al., 2017). Despite the increasing interest in using LPR data for research purposes, their applications are still very limited mainly due to privacy issues/concerns[1,2] that restrict agencies/data owners to publish and share the raw data.

Although one can replace license plate numbers by random identifiers as a simple anonymization strategy, many previous studies have demonstrated its vulnerability in many data sets across different fields/applications: most individuals can be easily re-identified from only a few observations. For example, Golle (2006) found that 87% of the populations in the US can be uniquely identified by the combination zip-code, gender, and date of birth. Location privacy disclosure occurs when individuals are re-identified by linkage attack from an anonymized travel behavior data sets (spatiotemporal mobility traces). Ma et al. (2013) indicated that an adversary is able to identify the trace of 30%-50% of the victims when she has collected 10 pieces of side information about a victim. Sun et al. (2013) proposed the traffic-knowledge-based adversary model which provided a high probability of successfully linking different anonymized traces between several consecutive intersections. de Montjoye et al. (2015) showed that, in a credit card metadata set, 90% users can be uniquely re-identified by using only four spatiotemporal points with the temporal resolution of one day, and the rate can be even higher if the transaction costs are included. Zang and Bolot (2011) examined a large-scale data set of call details records and showed that the top two visited locations are highly likely to be home and work locations, suggesting that individual's personal location information such as home/work location can be easily inferred from historical trajectory data. Therefore, developing new solutions to provide privacy protection before publishing and sharing individual-based data sets is an essential research question. Suppression and generalization are the most popular two anonymization approaches for privacy protection (Sweeney, 2002a). Suppression involves removing extreme cases from the original data set to avoid re-identification, while the idea of generalization is to replace or recode a value with a less specific but semantically consistent value. However, for both suppression and generalization, privacy is preserved at the cost of sacrificing data utility and analytical validity. An excellent data publishing/sharing solution should take into account the balance between the degree of privacy protection and the utility/usefulness of the data set itself. Therefore, it is critical to exploring possible solutions to enhance privacy protection in individual-based mobility trace data set and to investigating the privacy-and-utility trade-off of the publishing solutions.

In this paper, we first quantitatively measure to what extent an individual can be re-identified in mobility trace data sets. In doing so, we use a large LPR data set collected from Guangzhou, China, consisting of 260 million spatiotemporal transactions generated by 14 million vehicles in one month. We quantify the privacy disclosure risk of this data set based on the concept of k-anonymity (Sweeney, 2002b). Then, we examine a variety of factors that may impact an individual's degree of anonymity, including the temporal granularity, the size of the published data, the anonymity differences between local and non-local cars and the anonymity identified by continuous records. Finally, both a suppression solution and a generalization solution are proposed to design privacy-preserved publishing/sharing scheme for LPR data set, and we also explore the privacy-utility trade-off in these two solutions. The suppression solution removes sensitive records from individual's trace set to meet a specified anonymity constraint. For the generalization solution, we proposed an adaptive bintree-based cloaking algorithm that increases the temporal granularity of location information under the principle of minimum information loss to meet a specified anonymity constraint. To the best of our knowledge, this is the first study that (1) empirically measures the disclosure risk of a large-scale LPR data set from urban transportation systems, and (2) provides privacy-preserved LPR data publishing/sharing/processing solutions with the utility/value of data taken into account.

The remainder of this paper is structured as follows. Section 2 reviews the application of LPR data in transportation research and gives an overview of previous research on privacy protection in individual-based data sets. The notations used throughout this paper are summarized in Section 3. Then, Section 4 introduces the LPR data set and the data preprocessing. Section 5 introduces the k-anonymity model and gives a detailed description of the threat model. We explore the effects of different factors determining individual anonymity in Section 6, and propose a suppression solution and a generalization solution for publishing/sharing LPR data in a privacy-preserved way in Section 7. In this section, we also examine the privacy-and-utility trade-off of those two proposed solutions. Finally, Section 8 presents concluding remarks and possible future research directions.

2. Literature review

In this section, we enumerate some applications of LPR data in transportation research and provide background on the classical

[1] https://www.theatlantic.com/politics/archive/2016/01/vigilant-solutions-surveillance/427047/.
[2] https://www.wired.com/2015/05/even-fbi-privacy-concerns-license-plate-readers/.

fundamental privacy model and framework. The privacy issue of the publication of time-series trajectory data has also raised major concern in recent years. In addition, LPR data is essentially one of the individual-based spatiotemporal mobility trace data. Therefore works on the related issue of time-series trajectory sanitization thereof are also reviewed.

2.1. LPR data in transportation research

As an emerging ITS framework, LPR systems generate large quantities of spatiotemporal data in daily transport operations. These data sets provide the transportation research community with new opportunities to solve many real-world problems. In the following, we briefly review some representative work founded on LPR data. Shao and Chen (2018) applied tensor decomposition on spatio-temporal LPR data to estimate traffic volume. This method effectively solved the sparsity problem in traffic volume data. A lane-based real-time queue length estimation model using LPR data was proposed in (Zhan et al., 2015), which shows good performance on both precision and immediacy. Chen et al. (2017) applied K-mean clustering algorithms on LPR data to discover the underlying vehicle travel patterns. Sun et al. (2019) developed a two-dimensional laten Dirichlet allocation (LDA) model to discover underlying patterns and detect anomalies in individual travel behavior and demonstrate the effectiveness of the proposed modeling framework on a license plate recognition data set.

Despite its evident utility, the privacy issue is still the main obstacle preventing agencies from publishing and sharing the LPR data. Recently, the privacy issues in mobility trace data publishing and location information gathering under specific transportation applications have gained more and more attention in the literature (Cottrill and Vonu Thakuriah, 2015; Antoniou and Polydoropoulou, 2015; Zangui et al., 2013; Hu et al., 2019; Belletti and Bayen, 2018). Both Cottrill and Vonu Thakuriah (2015) and Antoniou and Polydoropoulou (2015) conducted a stated-preference survey to obtain the perception of consumers towards privacy risk. Cottrill and Vonu Thakuriah (2015) applied Principal Components Analysis (PCA) to investigate factors (both personal and contextual considerations) affecting privacy preferences (e.g., benefits, willing to trade and compensation related to mobile and locational technologies). Based on the survey data, Antoniou and Polydoropoulou (2015) adopted some economic concepts, i.e., indifference curve and marginal rate of substitution, to quantify the value of privacy. The analysis results show that the mean of the value of privacy is 2.87 €. Although their findings intuitively reflected users' perception of privacy risk, both of them may not be of much help to the fundamental privacy preservation of location information publishing/sharing, as the value of privacy was only explored from the perspective of consumer/user in general terms. Noting the raising privacy concerns of travelers in the application of congestion pricing, Zangui et al. (2013) proposed a new type of congestion pricing considering users' specific travel characteristics and attributes. The privacy costs of users choosing location-specific pricing schemes, and consequently, providing their location information were integrated into their generalized OD travel costs and the total toll revenue in road network was minimized to find the optimum pricing scheme. Similarly, Hu et al. (2019) considered different levels of privacy cost among different location-based service (LBS) user groups and proposed an incentive mechanism to encourage users to provide their location/trajectory information. The incentive model with user stochastic equilibrium was then developed and solved to capture the mixed behavior of groups with different privacy levels. Zangui et al. (2013) and Hu et al. (2019) provided great demonstrations of addressing privacy concern under specific transportation applications. Regrettably, however, both of them only conducted experiments on the theoretical level and didn't verify their frameworks on actual mobility trace data. Moreover, a vital part of their works is how to quantify the privacy risk in specific location and link to obtain the corresponding privacy cost, which is the problem we aim to address in this paper. Recently, focus has been also increasingly centered on the privacy preservation in location-based services (LBS). Avodji et al. (2016) proposed a decentralized architecture to compute meeting points in ridesharing service while protecting the privacy of location data of users. Belletti and Bayen (2018) devised a constrained integer quadratic program (CIQP) to solve the on-demand traffic fleet optimization problem in Mobility as a Service (MaaS) and strong privacy standards were guaranteed by using the dual splitting approach in the optimization process. These works provided vigorous technical guidance for location-based service (LBS) operators on preserving their user's privacy.

Recently, there are a number of studies focusing on privacy issues in the context of LPR data (Laine et al., 2009; Hier and Walby, 2011; Warren et al., 2013; Newell, 2014). Laine et al. (2009) summarized several types of privacy risks surrounding the use of LPR system and made recommendations for governing an agency's operation of a LPR system. Hier and Walby (2011) assessed the policy setting of public-area streetscape video surveillance system in Canada. Warren et al. (2013) examined privacy as a technique for governing road traffic surveillance using case studies of LPR in Canada and Australia. Newell (2014) investigated the trade-off between the personal information privacy and the efficacy of law enforcement in automated LPR system. However, they only describe the privacy risk of LPR data in a qualitatively way and present some solutions at the policy and administrative level.

2.2. Fundamental privacy framework

In the general field of privacy research, there is a large body of literature committed to designing privacy-preserved data publishing and sharing strategies. Sweeney (2002b) proposed the most famous k-anonymity privacy protection model. k-anonymity model is one of the most initial attempts in privacy protection and is used as a base for most of the privacy protection models. We call a data set having k-anonymity if each individual is indistinguishable from at least $k-1$ others in the data set. Consequently, a larger anonymity requirement k guarantees a higher level of privacy. To achieve k-anonymity protection of an anonymized data set, suppression and generalization techniques are usually applied (Sweeney, 2002a). k-anonymity ensures that anonymized individuals cannot be re-identified by linking attacks. But it fails to provide strong privacy protection and have some limitations (Domingo-Ferrer and Torra, 2008). For instance, the vulnerabilities of k-anonymity model are exposed when encountering two typical attacks: the

homogeneity attack and background knowledge attack, respectively. Homogeneity attack occurs when there is little diversity in the sensitive attributes in a *k*-anonymous table. An attacker can consequently discover the values of sensitive attributes of anonymized subjects. Background knowledge attack happens in the situation that the data recipients possess much specific background knowledge. The anonymized individual can still have high probabilities of being re-identified by combining the anonymous table with these background knowledge. Considering the shortcomings of it, many researchers devoted to improve the original *k*-anonymity property and proposed a number of sophistications of *k*-anonymity (Truta and Vinay, 2006; Wong et al., 2006; Machanavajjhala et al., 2006; Li et al., 2007). Machanavajjhala et al. (2006) proposed the *l*-diversity to defend against the homogeneity attack and the background knowledge attack on *k*-anonymity. Li et al. (2007) further improved the *k*-anonymity and proposed *t*-closeness to solve the attribute disclosure vulnerabilities inherent to *l*-diversity. Although many efforts were made to further enhance the *k*-anonymity model, neither *k*-anonymity nor its evolutions are entirely successful in ensuring that no privacy leakage occurs while keeping a reasonable data utility level (Domingo-Ferrer and Torra, 2008). The differential privacy protection model proposed by Dwork (2008) overcame the defects of the traditional privacy protection model and became a research hotspot in the field of privacy protection. Differential privacy is a statistical property about the behavior of the privacy mechanism and therefore is independent of the computational power and auxiliary information available to the adversary/attacker. One common technique to achieve differential privacy is by adding an amount of noise (Laplacian noise/Guassian noise) that grows with the complexity of the query sequence applied to the database. Although differential privacy was proved to provide rigorous privacy guarantee and immune to the background knowledge attack, most works on it focus on traditional relational data and are still limited to a very few primitive data application scenarios. Furthermore, the randomized mechanisms (by adding noise) cannot maintain data truthfulness, which is critical for some specific data mining tasks (e.g., auditing, data interpretation, and visual data mining).

2.3. Trajectory and sequential data sanitization

In recent years, the privacy issue has become the major concern of publishing and sharing trajectory data as a consequence of the increasing demand of it in extensive research in trajectory data mining. To date, several studies have investigated the anonymization of sequential locations under *k*-anonymity. Many researchers have utilized suppression or generalization techniques to realize anonymization of trajectory data while maintaining data utility as much as possible (Gruteser and Grunwald, 2003; Terrovitis and Mamoulis, 2008; Shin et al., 2010; Chen et al., 2013). Gruteser and Grunwald (2003) proposed an adaptive quadtree-based algorithm that decreases the spatial resolution of location information to meet a specified anonymity constraint. Shin et al. (2010) split user's trajectory and generalized the segmented trajectory to a spatiotemporal region to ensure trajectory *k*-anonymity. But both of them were focused on the anonymization of trajectory in the context of location-based services (LBS) in which the anonymity requirement is a local constraint (over a specific time duration and in a specific region) instead of a global one. While our work dedicates to realize *k*-anonymity protection of the whole LPR data set. The adversary may possess external information about any detectors over the released time period of data, which requires a global anonymization of individual's trajectory. Furthermore, spatial generalization is not applicable in the sanitization of LPR data set in which individual's location is represented by stationary detector's location. Terrovitis and Mamoulis (2008) devised a data suppression technique to prevent privacy leakage in the publication of trajectory database. Chen et al. (2013) first introduced local suppression to trajectory data anonymization to prevent individual's privacy against both identity linkage attack and attribute linkage attack under the framework of $(K, C)_L$-privacy. The $(K, C)_L$-privacy model assume that the adversary's prior knowledge is limited to at most L attributes in quasi-identifier (Mohammed et al., 2009). While in our assumption the adversary probably possess the knowledge of any spatiotemporal information of every vehicle in the LPR data set. Local suppression is examined to be superior to global suppression at the aspect of preserving data utility. However, a major problem with it is the expensive computational cost of finding valid suppression items, especially for the large trajectory database.

As the emerging privacy standard providing rigorous privacy guarantee against arbitrary background knowledge, differential privacy has been recently adopted to trajectory data sanitization (Chen et al., 2012, 2013; Hua et al., 2015). Chen et al. (2013) proposed a differentially private data sanitization approach based on a hybrid-granularity prefix tree structure for the transit data publication. The inherent consistency constraints of the prefix tree were utilized to conduct constrained inferences for better data utility. They extended the proposed approach by using variable-length n-gram model to sanitize general sequential data (Chen et al., 2012). There is no denying that the approach proposed by Chen et al. (2013, 2012) provides a new direction for sequential data sanitization under the framework of differential privacy. Nevertheless, as a consequence of the sanitization mechanism that adding noises to the real counts of prefixes or n-grams to guarantee differential privacy, an implicit assumption of it is that the raw trajectories to be published contain a large volume of common prefixes or n-grams. If these counts are very small, the added noises become relatively large and can compromise the truthfulness and utility of data. Unfortunately, although LPR data is much like transit data in terms of data type (i.e., both of them are collected by stationary detectors), LPR data does not follow this assumption. This is because the connectivity of road network is much higher than that of the public transit network. For instance, there are usually only a few of bus/metro lines between one's origin and destinations station. But she has multiple route choices when driving to her destination. Therefore, the frequencies of individual trajectories, prefixes and n-gram are extremely small. Hua et al. (2015) removed this assumption and proposed a differentially private publishing mechanism for more general time-series trajectories by generalizing the trajectories, i.e., dividing locations into several groups and replacing all locations within the same group with their centroid. However, as mentioned before, the spatial/location generalization is not applicable to LPR data in which individuals' locations are represented by locations of detectors. Since the location information of LPR data at detector level is of vital importance in many transportation researches, e.g., traffic volume estimation and traffic signal timing optimization, it is unacceptable to further generalize individuals' locations in LPR data set. Overall, although several studies have investigated trajectory and time-series data

Table 1
Summary of notations.

Notations	Description
About LPR dataset	
T	Time span of the LPR data
t	Temporal granularity of the LPR data
V	Set of vehicles in the published data set
$N = \|V\|$	Number of vehicles in published data set
V_i $(i = 1, ...,N)$	Vehicle i in V
R_i^j $(j = 1, ...,n_i)$	A spatiotemporal tuple record of V_i
$R_i = \{R_i^j \| j = 1, ...,n_i\}$	Trajectory (set of all spatiotemporal records) of V_i
$n_i = \|R_i\|$	Number of spatiotemporal records in R_i
$TR = \{R_i \| i = 1, ...,N\}$	Set of all vehicle trajectories R_i $(i = 1, ...,N)$
$TR_{N=100k}$	The experimental data set with 100,000 vehicles
$TR_{N=50k}$	The experimental data set with 50,000 vehicles
$TR_{N=10k}$	The experimental data set with 10,000 vehicles
$TR_{T=2wk}$	The experimental data set over first two weeks with 12,431 vehicles with monthly records larger than 1000
$TR_{T=4wk}$	The experimental data set over first four weeks with the same 12,431 vehicles in $TR_{T=2wk}$
About k-anonymity model	
QI_i^l	Quasi-identifier set of V_i with l records
l	Number of spatiotemporal records in QI_i^l
$A_i = \{V_u \| QI_i^l \subseteq R_u, V_u \in V\}$	Anonymity set of V_i identified by QI_i^l.
$K_i = \|A_i\|$	Anonymity value of V_i.
Two possible solutions	
SR_ε	Sensitive records with anonymity set less than ε
AGR_ε	Anonymity gain ratio after removing SR_ε
DLR_ε	Data loss ratio caused by removing SR_ε
$k_{\min}$	Uniform minimal anonymity constraint
t_{init}	Initialized time interval to aggregate data records (given in minutes)
$\delta_1, \delta_2, ...,\delta_n$	Initialized time periods with identical time interval t_{init}
$q_1, q_2, ...,q_n$	Traffic counts of $\delta_1, \delta_2, ...,\delta_n$
$\bar{\delta}_1, \bar{\delta}_2, ...,\bar{\delta}_m$	Aggregated time periods by our algorithm
$\bar{q}_1, \bar{q}_1. ...,\bar{q}_m$	Traffic counts of $\bar{\delta}_1, \bar{\delta}_2, ...,\bar{\delta}_m$
$\Gamma_{\text{opt}} = \{\tau_1, \tau_2, ...,\tau_{m+1}\}$	Optimal splitting scheme
D	Set of data records generated from a specific location in a specific day
$\Pi(D, \Gamma_{\text{opt}})$	Information loss of D caused by splitting scheme Γ_{opt}

sanitization approaches under differential privacy model, there is still a dearth of an efficient data-dependent differentially private sanitization method for LPR data.

We aim at quantifying privacy vulnerability of individual mobility trace data and provide efficient and practical technique guidances for data holder to publish and share this type of data (take the example of LPR data) in this paper. Consequently, we adopt the concept of anonymity to quantify the privacy disclosure risk of each individual in the LPR data set and propose practical solutions for publishing and sharing LPR data under the framework of k-anonymity. Although there are some flaws in the k-anonymity privacy protection model, it still becomes the most favorite indicator due to its natural accessibility and strong interpretability. In particular, k-anonymity is found to be an intuitive and effective indicator to evaluate privacy vulnerability in recent studies (de Montjoye et al., 2015; Zang and Bolot, 2011; de Montjoye et al., 2013). Based on the concept of k-anonymity, this work is dedicated to providing a quantitative framework to reveal the privacy disclosure risk in LPR data sets. More importantly, to provide technical guidance for privacy protection of LPR data set. Both suppression and generalization solutions are proposed to achieve k-anonymity privacy protection, and the trade-off between utility and privacy protection are explored.

3. Notations

We use the notations listed in Table 1 throughout this paper.

4. License plate recognition data

We use the anonymized LPR data collected from Guangzhou, China in March 2017. In this data set, each anonymized ID represents a unique license plate number. Overall, the data set contains 260 million records from 14 million vehicles. In total, there are 516 unique stationary gantries (detectors). Fig. 1 shows the road network of Guangzhou, in which the lines in different colors show different categories of roads and the colored dots represent detectors. Table 2 lists the primary fields in an example LPR record.

As can be seen from Table 2, the original passing timestamp is coded in a high resolution of one second, while in practice one cannot release the data at such a high temporal resolution. Such a high resolution is not necessary for traffic information retrieval/

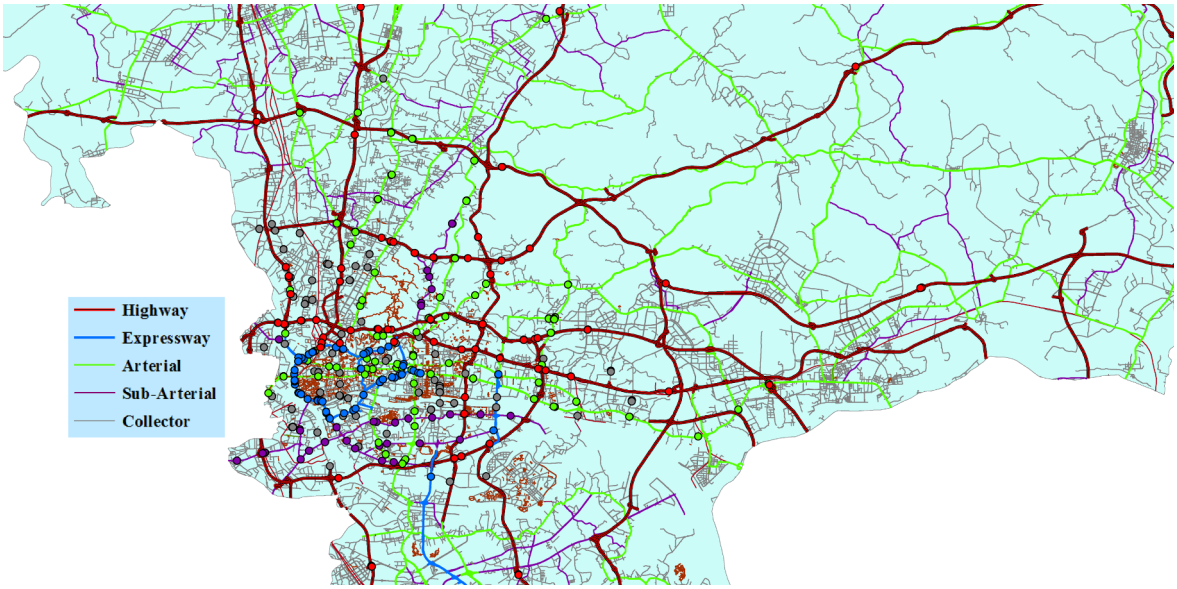

Fig. 1. Road network and distribution of LPR system in Guangzhou.

Table 2
Primary fields in the LPR data set.

Field	Example value
Vehicle ID	5ac0bd6239f8b9ac (anonymized)
Passing timestamp	2017-03-01 03:59:07
Address	Inner Ring Road, Meizhou Building (East to West)
Detector ID	17068
Drive direction	0
Local vehicle or non-local vehicle	1 (0: local car or 1: non-local car)

estimation. In addition, the one-second resolution is sufficient to uniquely identify most vehicles. Therefore, we divide a day into several slots (windows) of identical duration t, and convert the passing time to slot/window ID accordingly. For instance, if t is set to 2 h, the records generated from 00:00:01 to 02:00:00 and from 02:00:01 to 04:00:00 will be coded as "1" and "2", respectively. We experiment with five different temporal granularity values in this study, namely 30 min, 1 h, 3 h, 6 h, and 12 h.

In the anonymized LPR data set, we denote R_i^j as one spatiotemporal observation from vehicle V_i. And all R_i^j ($j = 1,...,n_i$) over a period of time T constitute the trajectory of V_i, which is denoted by R_i. A quasi-identifier QI_i^l is generated by sampling l spatiotemporal records (R_i^j) without replacement from its trajectory set R_i. In the following, we study the impact of different factors such as temporal resolution t, size of vehicle set N, and the total time span of data T on individuals' anonymity in Section 6.

5. *k*-anonymity and adversary model

We adopt the concept of k-anonymity to quantify the degree of uniqueness of each individual in the LPR data set. As mentioned, an anonymized data set with k-anonymity means each individual is indistinguishable from at least $k-1$ others. Specifically, the anonymity set of V_i given quasi-identifier QI_i^l is defined as the set of vehicles (including V_i) whose trajectory sets include QI_i^l. Thus, k-anonymity guarantees the size of anonymity set of each individual is not less than k. The larger the individual's anonymity is, the lower the privacy disclosure risk becomes. A smaller individual's anonymity means a larger privacy disclosure risk of her. However, an individual's anonymity equals to 1, in other words, one's anonymity set only contains herself, isn't equivalent to the leakage of her privacy. An adversary use information/knowledge from an external source and link to the released anonymized data set to identify the identity and sensitive information of an individual.

For each vehicle, the quasi-identifier is created by randomly drawing (without replacement) l spatiotemporal tuples R_i^j from its trajectory set, and we denote by $A_i(QI_i^l)$ the set of all vehicles V_u (including V_i) with $QI_i^l \subseteq R_u$ (i.e., QI_i^l is also observed in the trajectory of vehicle V_u). With this, the anonymity value of vehicle V_i is defined as $K_i = |A_i|$. By definition, we have $K_i \geqslant 1$, since at least QI will be observed in its own trajectory. We focus on the scenario that government department publishes or shares LPR data set to third parties (e.g., research institutions and technology enterprises) in this study. It is common that they possess a variety of databases and mature technologies (e.g., triangulation) for acquiring information or background knowledge of individuals. For an global anonymization of LPR data set to provide complete privacy protection, we assume that an adversary probably has the knowledge of any

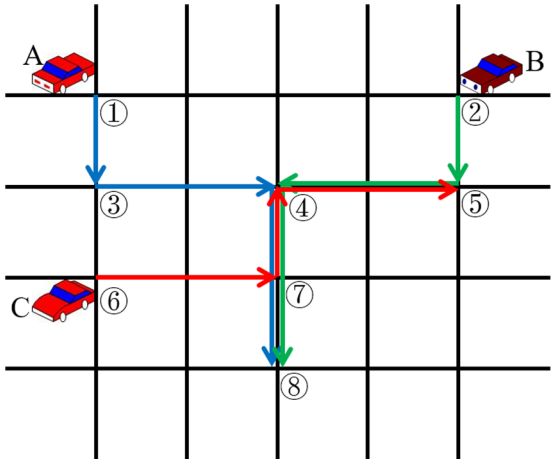

Fig. 2. Schematic illustration of *k*-anonymity.

spatiotemporal information of every vehicle in the LPR data set. The objective of an adversary is to infer the identity of an individual to learn sensitive information of her. Therefore, any spatiotemporal record in the LPR dataset is a potential threat to the privacy of individuals if they are utilized to identify the identity of individuals. It should be noted that the spatiotemporal records selected as quasi-identifier themselves probably not carry sensitive information. For example, suppose a vehicle has been identified by several spatiotemporal records and Jim is recognized as the owner. The adversary can track the history trajectory of Jim and obtain multiple sensitive information of him. For example, one can count the visit frequency of locations of Jim. Then the top two visited locations are highly likely to be his home and work locations (Zang and Bolot, 2011).

Fig. 2 gives a simple example of a road network in a regular square grid, where the serial numbers represent the locations of camera gantries, and the arrow lines represent trajectories of different vehicles. Three vehicles A, B, and C with the same speed departure their respective starting points at timestamp 0 and arrive at their destinations at timestamp 5. Table 3 lists the spatiotemporal points of A, B, and C, respectively. If we select the spatiotemporal tuple $QI = \{(①, 0)\}$ (the first element in a tuple represents the location of detector and the second denotes timestamp) as the quasi-identifier of vehicle A, then the anonymity of A is 1 because only A passed through location ① at time 0. In this case, when an adversary happened to across intersection ① and observed vehicle A (saw the license plate number of A) at time 0, then the identity (license plate number) of A was uniquely identified. If a set of three spatiotemporal tuples $QI = \{(④, 3), (⑦, 4), (⑧, 5)\}$ is selected, the anonymity of car A becomes 2 because both A and B went through location ④, ⑦, and ⑧ in sequence at time 3, 4 and 5. Similarly, the anonymity of A becomes 3 if the spatiotemporal point $QI = \{(④, 3)\}$ is selected as its quasi-identifier, because all of A, B, and C simultaneously passed ④ at time 3. In the situation of individual's anonymity larger than 1, the adversary cannot uniquely identify the identities of the vehicles although he had the observation of partial trajectory of them. Consequently, we can find that the concept of anonymity is an intuitive indicator for quantitatively measuring the degree of privacy disclosure risk in individual-based mobility trace data set.

6. Factors affecting anonymity

In this section, we quantify the disclosure risk in the LPR data set by computing anonymity measures and investigate different factors that affect the size of anonymity set. We conduct three experiments here on the LPR data. First, we quantify how anonymity measures interact with the temporal granularity and the size of the published data set. Next, we measure whether there is a significant difference in the anonymity of local and non-local vehicles. Finally, we test whether the results are consistent if we use temporally continuous records to create quasi-identifiers for each individual.

6.1. Temporal granularity

Different from Zang and Bolot (2011) that creates quasi-identifier using the most preferential locations, we randomly select l

Table 3
Spatiotemporal points of vehicles.

Vehicle ID	Timestamp					
	0	1	2	3	4	5
A	①	③	–	④	⑦	⑧
B	②	⑤	–	④	⑦	⑧
C	⑥	–	⑦	④	–	⑤

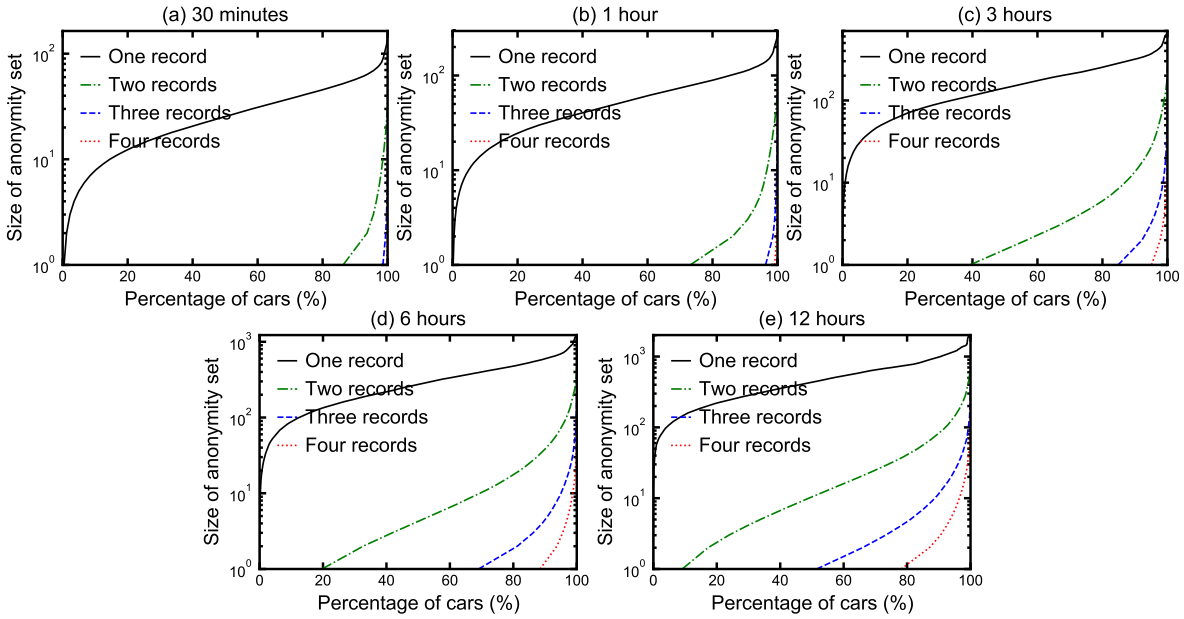

Fig. 3. Size of anonymity set when 1, 2, 3 and 4 (l) records are selected at different temporal granularity levels.

spatiotemporal records without replacement from individual trajectory to build QI (see Algorithm 1). Although we have 14 million vehicles in the LPR data set, computing anonymity on all of them will be infeasible given the $O(n^2)$ complexity. For this experiment, we create an experimental data set by randomly selecting 100,000 vehicles from the whole LPR data, which is denoted by $TR_{N=100k}$.

We examine the size of anonymity set identified by different numbers of spatiotemporal records at different levels of temporal granularity. In doing so, we sort individuals based on increasing anonymity and plot the cumulative curve of individual anonymity over different temporal granularity in Fig. 3. As can be seen, the larger the temporal granularity t is, the larger the average anonymity set will be. This is because a large t value leads to more vehicles passing during the time interval, and thus the individual anonymity increase.

Algorithm 1. Computing anonymity

Inputs:
TR
Outputs:
K_i, A_i, $i = 1, 2, ...,N$
Initialize:
$A_i \leftarrow \emptyset$, $i = 1, 2, ...,N$
for $i = 1$ to N **do**
$QI_i^l \leftarrow$ randomly sample l records from R_i
for $j = 1$ to N **do**
if $QI_i^l \subseteq TR_j$ **then**
$A_i \leftarrow A_i \cup V_j$
end if
end for
$K_i \leftarrow |A_i|$
end for

We also find that the individual anonymity decreases substantially with the increasing of the number of spatiotemporal records selected to identify individual's anonymity. We calculate the proportion of uniquely identified individuals whose anonymity set only contains itself at different temporal granularity level. Fig. 4 plots a histogram of the proportions of individuals uniquely identified by 2, 3, 4 or 5 spatiotemporal records, respectively. Error bars denote the 99% confidence interval on the mean. The result corresponds to our intuition that a larger set of quasi-identifier leads to higher privacy risk (a lower K_i). Surprisingly, even when we set the temporal granularity to half a day (12 h), there are still about 90% of individuals can be uniquely identified by only 5 spatiotemporal records in the LPR data set. This result is in line with previous studies based on call detailed records and credit card transactions (de Montjoye et al., 2013, 2015). The LPR data further confirms that the individual mobility traces are highly unique, which enables the re-identification of individuals without using external information. In other words, there is substantial privacy disclosure risk in releasing mobility trace data sets.

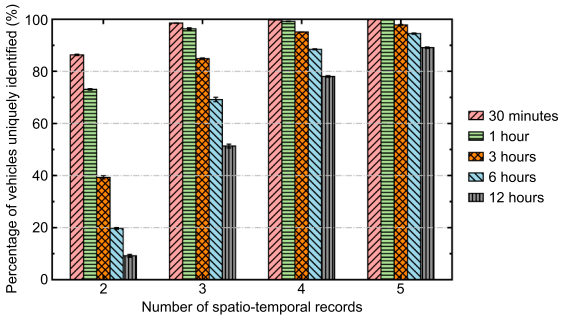

Fig. 4. Percentage of uniquely identified vehicles at different temporal granularity level.

In practice, agencies often only provide a small set of samples, such as PUMS (public use micro samples) in census. We next investigate the impact of the size of released data set on individual anonymity. We examine two factors affecting the size of data set: (1) number of released vehicles, and (2) length of covered time period.

6.2. Number of released vehicles

To examine how the number of released vehicles interacts with anonymity values, we focus on the case of $l = 1$ (only one spatiotemporal record selected as *QI*) and the temporal granularity $t = 3$ h. We randomly select two subsets of 50,000 vehicles and 10,000 vehicles from $TR_{N=100k}$, and denote them by $TR_{N=50k}$ and $TR_{N=10k}$, respectively. We calculate the anonymity of vehicles in the three experimental data sets according to Algorithm 1. Fig. 5 shows the distribution of individual anonymity in the corresponding data set. In order to more intuitively reveal the relationship of individual anonymity in data sets with different number of released vehicles, we expand the anonymity of each individual in the data set with 10,000 vehicles by 5 and 10 times and the data set with 50,000 vehicles by 2 times, respectively, and denote them by $TR_{N=10k\times5}$, $TR_{10k\times10}$ and $TR_{N=50k\times2}$. As a comparison, we also plot the distribution of the expanded anonymity in Fig. 5. We see that the size of anonymity set increases with the number of vehicles in the released data set and the anonymity distribution of $TR_{N=10k\times5}$, $TR_{N=10k\times10}$ and $TR_{N=50k\times2}$ are almost identical to the anonymity distribution of $TR_{N=50k}$ and $TR_{N=100k}$, respectively. This analysis shows that the individual anonymity is basically in proportion to the number of vehicles in the released data set.

6.3. Length of released time period

To study the anonymity of individuals in published datasets over different released time length, we focus on the case with temporal granularity $t = 3$ h. To ensure the consistency of vehicle set in the published data sets over different released time length, we select the trajectories of individuals with more than 1000 monthly records from original data set to create a new data set of 12,431 vehicles. Then we use the records from the first two weeks and the records from all the four weeks as two experimental data sets and denote them as $TR_{T=2wk}$ and $TR_{T=4wk}$. Again, we calculate the anonymity of individuals in these two experimental data sets based on Algorithm 1. Fig. 6 shows the distributions of individual anonymity identified by 1, 2, and 3 spatiotemporal records in the experimental data sets. The box plot highlights that individual anonymity distributions on published data sets over for different released time spans are almost the identical, which suggests that individual anonymity is essentially invariant to the length of released time period.

To explain the high similarity of individual anonymity over for different time spans (two weeks and four weeks), we examine the probability that a vehicle passes through different locations over different time periods of a day in two days and calculate the cosine

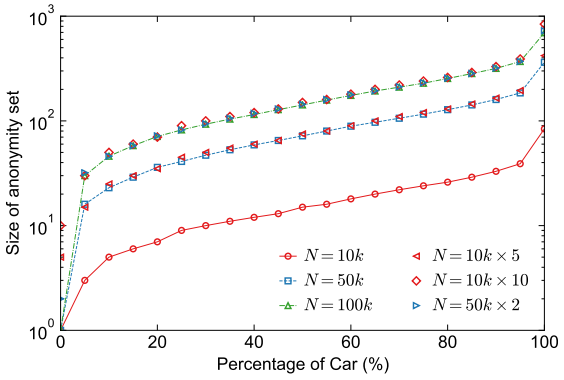

Fig. 5. Anonymity distribution in published datasets with different number of vehicles.

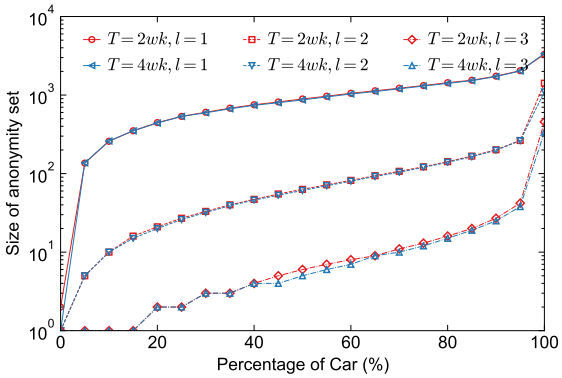

Fig. 6. Anonymity distribution of published datasets over different period of time.

similarity of the distribution of the two days. The length of the vector is the number of time intervals (depending on the time interval t) and each element in the vector corresponds to the fraction of vehicles passing through a particular location/detector at each time interval of a day. Fig. 7 shows the cumulative distribution function (cdf) of the cosine similarity between March 15 and March 16, and between March 15 and March 22, respectively. We find that the cosine similarity between March 15 and March 22 is slightly smaller than between March 15 and March 16. However, we also notice that over 90% and over 95% of the locations have cosine similarity greater than 0.8 which means that the traffic volume patterns of most locations have strong periodicity and the travel routes of most individuals are very similar from day to day (see Cao et al., 2007). Overall, the results suggest that the high similarity of day-to-day travel behaviors of individuals is the factor that leads to the identical anonymity distributions in published data sets covering different time length.

6.4. Difference between local and non-local vehicles

We also investigate the anonymity difference between local vehicles and non-local vehicles. According to the license plate number, there are 48,489 local vehicles and 51,511 non-local vehicles in $TR_{N=100k}$. We count the mean anonymity of local vehicles and non-local vehicles according to the calculated anonymity results of $TR_{N=100k}$ in the previous section and calculate the difference between mean anonymity of local vehicles and non-local vehicles. Fig. 8 shows the difference of the mean anonymity of local vehicles and non-local vehicles identified by 1, 2, 3 and 4 spatiotemporal records. Error bars denote the 95% confidence interval on the mean. We find that the anonymity of local vehicles are greater than the anonymity of non-local vehicles when only one spatiotemporal record is selected to identify the anonymity set. When multiple spatiotemporal records are selected, the result is the opposite.

The anonymity identified by only one spatiotemporal record is actually the number of vehicles passing a specific location over a specific period of time. Obviously, the smaller the traffic volume of the location over the time period is, the smaller anonymity set identified by the single record is. The non-local vehicles enter and leave Guangzhou driving along the inter-city highway built in the suburbs and the local vehicles are mainly driven along urban roads in urban areas. The traffic volumes of inter-city highways are generally smaller than the road traffic volumes in urban areas of Guangzhou. So the anonymity values of local vehicles are greater than the anonymity of non-local vehicles when only one record is used. However, when it comes to the situation with multiple spatiotemporal records, the anonymity set is actually the intersection of several anonymity sets identified by several single records. The locations of camera gantries in these inter-city highways coexist in the trajectories of non-local vehicles entering or leaving Guangzhou. When multiple spatiotemporal records are selected to identify the anonymity set, the common trajectories in inter-city highways of non-local vehicles entering/leaving Guangzhou increase the anonymity of those vehicles.

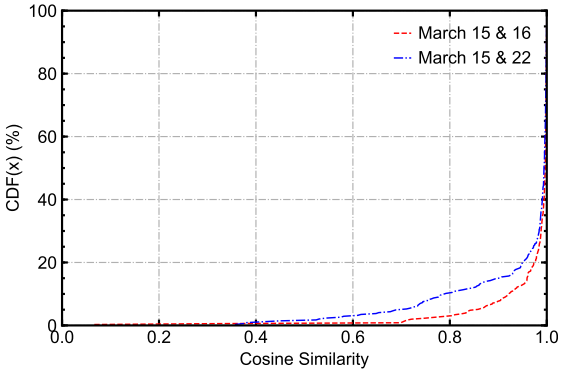

Fig. 7. Cosine similarity between two different days.

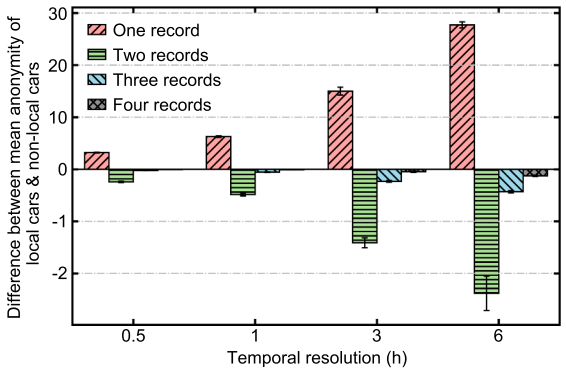

Fig. 8. The anonymity differences between local vehicles and non-local vehicles.

6.5. Quasi-identifier using continuous records

To investigate the anonymity identified by temporal continuous records, we randomly selected several temporal continuous records from individual trajectory to identify its anonymity set (see Algorithm 2). We focus on the case with temporal granularity $t = 6h$ and calculate the individual anonymity of $TR_{N=100k}$ identified by 1, 2, 3 and 4 records according to Algorithm 1 and Algorithm 2, respectively. Fig. 9 shows the anonymity results.

As shown in Fig. 9, there is almost no difference between the individual anonymity calculated by Algorithm 1 and Algorithm 2 when only one spatiotemporal record is selected, which is expected. When multiple records are selected, the individual anonymity calculated by Algorithm 2 becomes much larger than the anonymity calculated by Algorithm 1, indicating that individuals are more likely to be uniquely identified by several discontinuous spatiotemporal points of trajectory. This is consistent with our intuition. It is common for different individuals to have a common continuous trajectory: for example when vehicles traveling in a group, they will more likely to be in the anonymity set of each other's.

Algorithm 2. Computing anonymity with continuous quasi-identifiers

Inputs:
TR
Outputs:
$K_i, A_i, i = 1, 2, ...,N$
Initialize:
$A_i \leftarrow \varnothing, i = 1, 2, ...,N$
for $i = 1$ to N **do**
$QI_i^l \leftarrow$ randomly sample temporal continuous l R_i^j from R_i
for $j = 1$ to N **do**
if $QI_i^l \subseteq TR_j$ **then**
$A_i \leftarrow A_i \cup V_j$
end if
end for
$K_i \leftarrow |A_i|$
end for

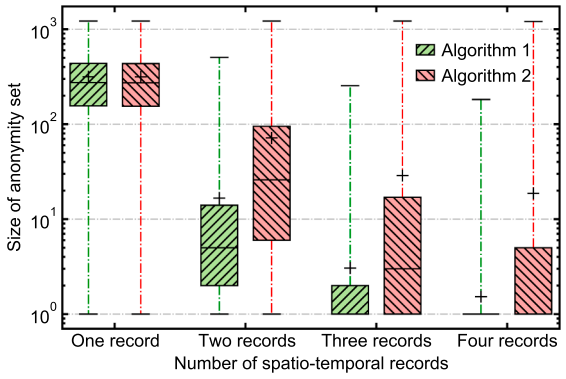

Fig. 9. Anonymity results identified by Algorithm 1 and Algorithm 2.

7. Two possible solutions

In this section, inspired by the empirical results in Section 6, we propose possible solutions from perspectives of both suppression and generalization to achieve *k*-anonymity privacy protection of LPR data set. More specifically, we remove sensitive records in the suppression solution and pay close attention to the loss of data caused by it. As for the generalization solution, we propose an adaptive time interval cloaking algorithm that adaptively adjust the temporal resolution of location information based on traffic counts and the principle of minimal information loss.

7.1. Suppression solution

As can be seen from Algorithm 1 and Algorithm 2, the anonymity identified by only one spatiotemporal record is actually the number of vehicles passing a specific location over a specific period of time and the anonymity set identified by multiple spatiotemporal records is actually the intersection of several anonymity sets identified by several single records. Obviously, the smaller the traffic volume of the location over the time period is, the smaller anonymity set identified by the single record is. The anonymity identified by a quasi-identifier depends on the anonymity sets identified by several single records in the quasi-identifier.

Based on the observations above, we propose a suppression solution for privacy protection of LPR data publishing. Suppression refers to remove extreme cases from original data set, which is the most direct way to achieve *k*-anonymity protection. When it is applied to our LPR data set, we define sensitive records as the record whose anonymity set is smaller than ε and denote it as SR_ε. In other words, a sensitive record means the traffic count of a specific location over a specific period of time is less than ε. It is obvious that the presence of extreme sensitive records in individual's trajectory brings down the overall anonymity of individuals. Therefore, we remove sensitive records from individual's trajectory and investigate the effects on privacy protection and data of it. It should be noted that the suppression solution here is a kind of global suppression (comparing to the local suppression proposed by Chen et al. (2013)) since we remove sensitive records from the whole LPR data set. Actually, the suppression solution here is a special case of the anonymization method in (Chen et al., 2013). The sensitive record is essentially a MVS (Minimal violating sequence) with a sequence length of 1. Since we aim at achieving *k*-anonymity privacy guarantee when a single spatiotemporal record is selected to identify individual's anonymity, the global suppression is necessary for the global anonymization of trajectory of each individual. To quantify the impact of the suppression solution on privacy-preserving and usefulness of data, we define anonymity gain ratio as the ratio of mean anonymity gain to the mean anonymity of initial data set ($\varepsilon = 0$) after removing of sensitive records and denote it as AGR_ε. We define data loss ratio as the ratio of the number of removed data records to the total number of records in initial data set and denote it as DLR_ε. We remove SR_ε from our experimental data set $TR_{N=100k}$ and explore the impact on AGR_ε and DLR_ε over different ε. We focus on the case with temporal granularity $t = 6h$ and the individual anonymity is identified by 1, 2, and 3 spatiotemporal records according to Algorithm 1. Fig. 10 shows AGR_ε (left y-axis scale) and DLR_ε (right y-axis scale) after removing SR_ε over different ε. Error bars demote the 95% confidence interval on the mean. We find that individual anonymity has a significant improvement after removing sensitive records. The average individual anonymity identified by three spatiotemporal records increases by more than 20% at the cost of less than 8% loss of data. We also notice that both AGR_ε and DLR_ε increase with ε, which demonstrate the proposed suppression solution is virtually a game between privacy preserving and utility of data.

Except for the overall improvement of individual anonymity, we also concern about the impact of the proposed suppression solution on minimum individual anonymity in data set. Fig. 11 plots the minimum individual anonymity identified by 1, 2, and 3 spatiotemporal records after removing SR_ε over different ε. Error bars denote the 95% confidence interval on the mean. Comparing to the case of $\varepsilon = 0$ (unprocessed data set), we find that the minimum anonymity identified by one or two spatiotemporal records has a distinct increasement after removing sensitive records, which indicate our solution has a good performance on reducing the risk of being re-identified. Just as we expected, ε-anonymity of released data set is achieved when only one spatiotemporal record is selected to identify the anonymity set. However there is only a slight improvement on the minimum individual anonymity identified by three spatiotemporal records, illustrating the strong uniqueness of individual mobility traces.

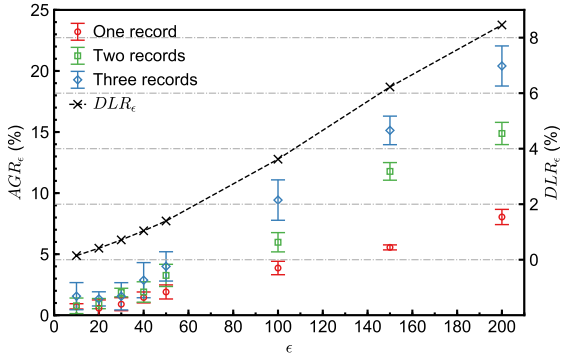

Fig. 10. AGR_ε and DLR_ε after removing SR_ε over different ε.

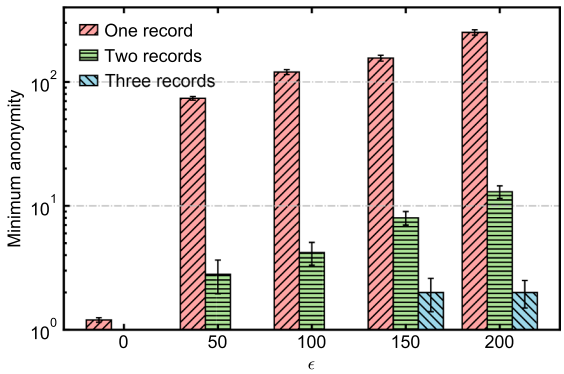

Fig. 11. Minimum anonymity after removing SR_ε over different ε.

7.2. Generalization solution

Generalization is an another common technique to achieve k-anonymization. We coarsen the location information in time domain by dividing a day into several time slots with identical time interval t to realize the generalization of location data in the previous section and the results presented above have shown that the individual anonymity increase with the temporal granularity, which inspires us to publish mobility trace data without compromising privacy by enlarging temporal granularity. However, the coarser the temporal resolution is, the more information the data losses. Taking into account the balance between the degree of privacy protection and the utility/usefulness of the data, we propose an temporal generalization solution to achieve k-anonymity privacy protection of released data set under the principle of minimum information loss.

Inspired by the adaptive quadtree-based spatial cloaking algorithm proposed by Gruteser and Grunwald (2003), we propose an adaptive temporal cloaking algorithm to meet specified anonymity constraints. While Gruteser and Grunwald (2003) used quadtree partitioning, our algorithm is based on the binary tree structure which is more flexible and suitable for temporal cloaking. The key idea underlying our algorithm is that a given specified anonymity constraint $k_{\min}$ can be maintained in any period of time in a day by decreasing the temporal resolution of released data records. To this end, given an anonymity constraint $k_{\min}$, for a specific location of detector, our algorithm subdivide the period of time until the traffic count during this period of time falls bellow the constraint $k_{\min}$. In addition, our algorithm is committed to achieving k-anonymity protection of released data set while maintaining utility/usefulness of data, which is proved to be a NP-hard problem (Gionis and Tassa, 2008). For this purpose, the algorithm select splitting points recursively based on the principle of minimal information loss to find the approximate solution to this problem.

We propose an entropy measure of information loss caused by temporal cloaking. Let D be the collection of data records generated from a specific location within a specific day. As can be seen from Table 2, the original resolution of data records is one second. To enhance the algorithmic efficiency and decrease the calculating time, as we do in previous section, we divide a day into several time slots $\delta_1, \delta_2, ..., \delta_n$ with identical time interval t_{init} (given in minutes, e.g. 5 min, 10 min) and the splitting points are constrained to choose from the beginnings or endings of these time slots, namely, the subscript set $\{1, 2, ..., n+1\}$. We count the number of data records within these time slots, namely, the traffic counts of $\delta_1, \delta_2, ..., \delta_n$ and denote them as $q_1, q_2, ..., q_n$. To meet the given anonymity constraint $k_{\min}$, our algorithm aggregate $\delta_1, \delta_2, ..., \delta_n$ to several coarser period of time $\bar{\delta}_1, \bar{\delta}_1. ..., \bar{\delta}_m$ segmented by splitting points $\Gamma_{\text{opt}} = \{\tau_1, \tau_2, ..., \tau_{m+1}\}$. It is obvious that Γ_{opt} is a subset of $\{1, 2, ..., n+1\}$. Similarly, we count the number of data records within different time periods $\bar{\delta}_1, \bar{\delta}_1. ..., \bar{\delta}_m$ and denote them as $\bar{q}_1, \bar{q}_1. ..., \bar{q}_m$. Define two discrete random variables $\delta \in \{\delta_1, \delta_2, ..., \delta_n\}$ and $\bar{\delta} \in \{\bar{\delta}_1, \bar{\delta}_1. ..., \bar{\delta}_m\}$ to capture these two random events: a data record is generated in the initial time slot δ, and the aggregated time period $\bar{\delta}$, respectively. Then the information loss of data set D due to the temporal generalization scheme Γ_{opt} can be defined as the empirical conditional entropy as the following Eq. (1). The temporal generalization algorithm that achieves k-anonymity protection of released data set with minimal information loss is described in more detail in Algorithm 3. After getting the temporal splitting scheme of a location, we can convert passing time of records generated from this location to ID of time period accordingly.

$$\Pi(D, \Gamma_{\text{opt}}) = H(\delta|\bar{\delta}) = -\sum_{i=1}^{m} \frac{\bar{q}_i}{|D|} \sum_{j=\tau_i}^{\tau_{i+1}-1} \frac{q_j}{\bar{q}_i} \log \frac{q_j}{\bar{q}_i} \tag{1}$$

To verify the performance of our algorithm, we divide $TR_{N=100k}$ into several subdataset according to different dates and locations and implement the temporal generalization algorithm on them. We set the initial temporal resolution t_{init} to 5 min and implement our algorithm to get several anonymized data sets under different anonymity constraint $k_{\min}$. We calculate the individual anonymity identified by 1, 2, 3, and 4 data records of these anonymized data sets according to Algorithm 1. Fig. 12 presents an overview of our results. It illustrates the tradeoff between the overall anonymity and temporal resolution, showing mean anonymity of all individuals (left y-axis scale) and median resolution of all generalized data records (right y-axis scale) over different anonymity constraint $k_{\min}$. As we can see from Fig. 12, the bigger the anonymity constraint $k_{\min}$ is required, the bigger the mean individual anonymity of anonymized data set is, and the coarser the temporal information of data records is. This indicates the privacy-and-utility trade-off of the our temporal generalization algorithm.

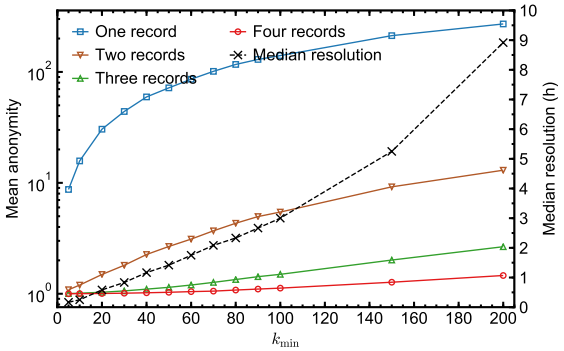

Fig. 12. Dependency of mean anonymity and temporal resolution over different $k_{\min}$.

Algorithm 3. Adaptive time interval cloaking algorithm

Inputs:
D, t_{init}, $k_{\min}$, q_1, q_2, ...,q_n, $n = 1440/t_{\text{init}}$
Outputs:
Γ_{opt}
Initialize:
$\Gamma_{\text{opt}} \leftarrow []$
$\Gamma \leftarrow [1, n+1]$
while $\Gamma_{\text{opt}} \neq \Gamma$ **do**
$\Gamma_{\text{opt}} \leftarrow \Gamma$
$m = |\Gamma|$
for $i = 1$ to $m - 1$ **do**
$\tau_{\text{left}} \leftarrow \min\tau$ **s.t.** $\sum_{j=\tau}^{\tau-1} q_j \geqslant k_{\min}$, $\tau \in \{\Gamma_{[i]}, \Gamma_{[i]}+1, ..., \Gamma_{[i+1]}\}$
$\tau_{\text{right}} \leftarrow \max\tau$ **s.t.** $\sum_{j=\tau}^{\Gamma_{[i+1]}-1} q_j \geqslant k_{\min}$, $\tau \in \{\Gamma_{[i]}, \Gamma_{[i]}+1, ..., \Gamma_{[i+1]}\}$
if $\tau_{\text{left}} \leqslant \tau_{\text{right}}$ **then**
$D' \leftarrow$Data records generated between $\Gamma_{[i]}$ and $\Gamma_{[i+1]}$
$\tau_{\text{opt}} \leftarrow \arg\min_\tau \Pi(D', \tau)$
Add τ_{opt} to Γ
$\Gamma \leftarrow$ sorted(Γ)
end if
end for
end for

To highlight the superiority of our algorithm to the uniform time interval cloaking method, we focus on the case of $l = 1$ (one spatiotemporal record selected). Fig. 13 plots the anonymity distribution of the anonymized data set with a uniform temporal granularity $t = 3h$ and the anonymity distribution of the data set with a median temporal resolution $t_{\text{median}} = 3h$ which is anonymized by Algorithm 3 under the anonymity constraint $k_{\min} = 100$. We find that although these two data set anonymized by different temporal generalization methods have similar temporal granularity level, there are conspicuous differences of the anonymity distribution of them. Although the mean anonymity of the data set anonymized by uniform time interval cloaking method (the green

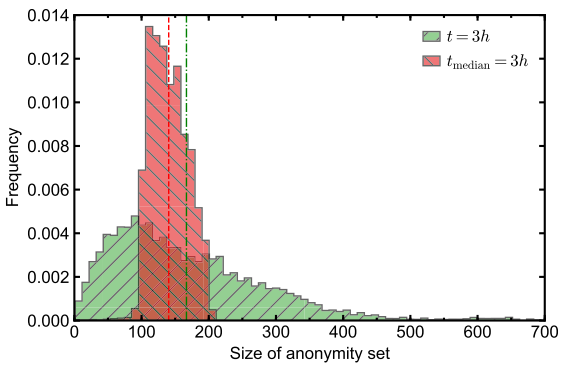

Fig. 13. Anonymity distribution of uniform time interval cloaking method and Algorithm 3.

Table 4
Traffic volume for different road types.

Road type	Traffic volume
Highway	21029
Expressway	33763
Arterial	20329
Sub-Artetial	18997
Collector	6371

vertical dashed line) is larger than the mean anonymity of the data set anonymized by our algorithm (the red vertical dashed line), the individual anonymity generated by our algorithm has a more concentrated distribution than the anonymity result generated by uniform time interval cloaking method. A majority of individuals in the anonymized data set generated by our algorithm have an individual anonymity larger than 100 (except the case that the daily traffic volume is less than 100) which is a direct response to the minimum anonymity constraint $k_{\min} = 100$. At the same time, almost all of individuals have an individual anonymity less than 200 which is the result of pursuing the accurate temporal information of data as much as possible under the minimum anonymity constraint.

Our algorithm adaptively generalize the temporal information of data records according to the traffic volume because it changes heavily between peak and night time off-peak hours. But at the same time, the traffic flow also shows a large imbalance in spatial domain. We classify all detectors into five categories according to the type of road in which they are located. There are five types of roads in the road network of Guangzhou, namely, highway, expressway, arterial, sub-arterial, and collector (see Fig. 1). We average the traffic counts for different road types and show them in Table 4. The traffic volume is calculated as the average 24 h unidirectional traffic counts. Based on the observation above, we investigate the dependency of generalization results of our algorithm on the spatial imbalance of traffic flow. Fig. 14 illustrates the dependency of resulting temporal resolution and information loss under the anonymity constraint $k_{\min} = 100$ on road characteristics and traffic volumes. For each evaluation road type, the figure shows the median resolution (left y-axis scale) and the mean information loss (right y-axis scale) of daily data set generated from all detectors located in the same type of road. We find that the median temporal granularity and the information loss decrease as traffic volume. Under the uniform anonymity constraint, for the expressways with the highest density of vehicles, the temporal accuracy is the 204 min. For the collectors with the lowest traffic volume the resolution decreases to 675 min.

8. Conclusion and discussion

In this paper, we quantitatively measure the privacy disclosure risk in an LPR data set based on the concept of *k*-anonymity. Our study shows that individuals in anonymized LPR data set still have a very high risk of being re-identified. Five spatiotemporal records are enough to uniquely identify about 90% of individuals even the temporal granularity is set to be half of one day. This result shows that replacing license plate number with random identifiers is far from enough to protect privacy, and our study serves as a reminder to relevant agencies and data owners to pay special attention when releasing mobility trace data sets. Moreover, we investigate the effect of multiple factors affecting individual anonymity, including the temporal granularity, the size of the published data, the anonymity differences between local and non-local cars, and the anonymity identified by continuous and non-continuous records. The results provide technical guidance for agencies to publish and share LPR data set.

To publish LPR data without compromising privacy and to maintain the utility of the data, we propose both a suppression solution and a generalization solution in this paper. The suppression solution removes sensitive records from the data set. The results on the trade-off between privacy and utility show that the average individual anonymity with $l = 3$ could increase by more than 20% at the cost of less than 8% loss of data, which suggests the suppression solution has a notable performance on privacy protection without sacrificing much utility. The generalization solution is driven by a bintree-based adaptive time interval cloaking algorithm. The

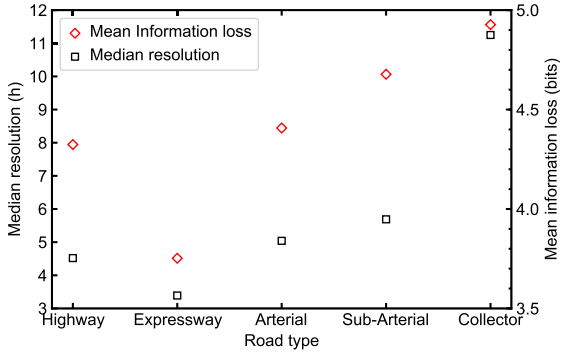

Fig. 14. Dependency of temporal resolution and information loss on road type.

bintree-based temporal generalization algorithm is introduced to achieve *k*-anonymity privacy protection of released data set with minimal loss of information. Our results show that the bintree algorithm is extremely well-designed to satisfy user-specified minimal anonymity constraint and is more flexible and reliable compared to the uniform time interval cloaking method. Furthermore, we show that the temporal accuracy of data depends on traffic condition. For the expressways with the heaviest traffic flow, the median accuracy is 204 min and the value increases to 675 min for collectors with the lowest traffic volume under the uniform anonymity constraint $k_{\min} = 100$.

Overall, this paper reveals the high risk of privacy disclosure of LPR data from a quantitative point of view. We introduce possible solutions to provide privacy protection for agencies publishing/sharing mobility trace data and discuss the privacy-and-utility trade-off when releasing such data sets. The results show that our solutions have good performance to improve individual anonymity. In particular, both of proposed solutions can achieve *k*-anonymity privacy protection of released data set when $l = 1$ (one spatio-temporal point as quasi-identifier). But as the size of quasi-identifier increases, the privacy-preserving effect of them gets smaller and smaller, which indicates the strong uniqueness of individual mobility traces. On the other hand, it also shows the drawback of the *k*-anonymity that failing to provide strong privacy guarantee when the adversary possess a large volume of external/background information (e.g., individual's multiple spatiotemporal information in this work). Privacy disclosure risk occurs when the real spatiotemporal information of an individual's mobility traces can be linked to his/her external information. To protect people from linkage attacks and to provide stronger privacy protection, our future work will focus on trajectory synthesis, which generates synthetic trajectories of an individual based on his/her historical mobility trace data using statistical learning models (e.g. He et al., 2015; Sun and Axhausen, 2016; Yin et al., 2018). A good synthesis model is expected to capture the underlying structure and statistical properties of the real data as much as possible. With these solutions, one can reproduce real-world transportation features (e.g., demand, speed, and congestion pattern) without compromising one's privacy. Another future research direction is to identify the best data publishing solutions for different data application scenarios/requirements. Although in this paper we introduce the data loss ratio and information loss to quantify the overall loss of data usefulness/utility, different real-world engineering applications may have different objectives in practice. We hope this work could stimulate more discussion to address the privacy issues in individual-based mobility trace data set.

Acknowledgement

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

Appendix A. Supplementary material

Supplementary data associated with this article can be found, in the online version, at https://doi.org/10.1016/j.trc.2019.04.022.

References

Antoniou, C., Polydoropoulou, A., 2015. The value of privacy: evidence from the use of mobile devices for traveler information systems. J. Intell. Transp. Syst. 19 (2), 167–180.
Avodji, U.M., Gambs, S., Huguet, M.-J., Killijian, M.-O., 2016. Meeting points in ridesharing: a privacy-preserving approach. Transp. Res. Part C: Emerg. Technol. 72, 239–253. http://www.sciencedirect.com/science/article/pii/S0968090X1630184X .
Belletti, F., Bayen, A.M., 2018. Privacy-preserving maas fleet management. Transp. Res. Part C: Emerg. Technol. 94, 270–287.
Cao, H., Mamoulis, N., Cheung, D.W., 2007. Discovery of periodic patterns in spatiotemporal sequences. IEEE Trans. Knowl. Data Eng. 19 (4), 453–467.
Chen, H., Yang, C., Xu, X., 2017. Clustering vehicle temporal and spatial travel behavior using license plate recognition data. J. Adv. Transp. 2017 (7), 1–14.
Chen, R., Fung, B., Desai, B.C., Sossou, N.M., 2012. Differentially private transit data publication: a case study on the montreal transportation system. In: Proceedings of the 18th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. ACM, pp. 213–221.
Chen, R., Fung, B.C., Mohammed, N., Desai, B.C., Wang, K., 2013. Privacy-preserving trajectory data publishing by local suppression. Inf. Sci. 231, 83–97.
Cottrill, C.D., Vonu Thakuriah, P., 2015. Location privacy preferences: a survey-based analysis of consumer awareness, trade-off and decision-making. Transp. Res. Part C: Emerg. Technol. 56, 132–148.
de Montjoye, Y.-A., Hidalgo, C.A., Verleysen, M., Blondel, V.D., 2013. Unique in the crowd: the privacy bounds of human mobility. Sci. Rep. 3, 1376.
de Montjoye, Y.A., Radaelli, L., Singh, V.K., Pentland, A.S., 2015. Identity and privacy. Unique in the shopping mall: on the reidentifiability of credit card metadata. Science 347 (6221), 536–539.
Domingo-Ferrer, J., Torra, V., 2008. A critique of k-anonymity and some of its enhancements. In: International Conference on Availability.
Dwork, C., 2008. Differential privacy: a survey of results. In: International Conference on Theory and Applications of MODELS of Computation. pp. 1–19.
Gionis, A., Tassa, T., 2008. k-anonymization with minimal loss of information. IEEE Trans. Knowl. Data Eng. 21 (2), 206–219.
Golle, P., 2006. Revisiting the uniqueness of simple demographics in the us population. In: ACM Workshop on Privacy in Electronic Society. pp. 77–80.
Gruteser, M., Grunwald, D., 2003. Anonymous usage of location-based services through spatial and temporal cloaking. In: International Conference on Mobile Systems, Applications, and Services. pp. 31–42.
He, X., Cormode, G., Machanavajjhala, A., Procopiuc, C.M., Srivastava, D., 2015. Dpt: differentially private trajectory synthesis using hierarchical reference systems. Proc. VLDB Endowment 8 (11), 1154–1165.
Herrera, J.C., Work, D.B., Herring, R., Ban, X.J., Jacobson, Q., Bayen, A.M., 2010. Evaluation of traffic data obtained via gps-enabled mobile phones: the mobile century field experiment. Transp. Res. Part C: Emerg. Technol. 18 (4), 568–583.
Hier, S.P., Walby, K., 2011. Privacy pragmatism and streetscape video surveillance in canada. Int. Sociol. 26 (6), 844–861.
Hu, H., Sun, Z., Yang, X., 2019. Privacy implication of location-based service: multi-class stochastic user equilibrium and incentive mechanism. In: Transportation Research Board (TRB) Annual Meeting.
Hua, J., Gao, Y., Zhong, S., 2015. Differentially private publication of general time-serial trajectory data. In: Computer Communications (INFOCOM), 2015 IEEE Conference on. IEEE, pp. 549–557.
Laine, R., McMahon, J., Rosenblatt, D., Talucci, V., 2009. Privacy impact assessment report for the utilization of license plate readers. Tech. rep., International Association of Chiefs of Police.

Li, N., Li, T., Venkatasubramanian, S., 2007. t-closeness: privacy beyond k-anonymity and l-diversity. In: IEEE International Conference on Data Engineering. pp. 106–115.
Ma, C.Y., Yau, D.K., Yip, N.K., Rao, N.S., 2013. Privacy vulnerability of published anonymous mobility traces. IEEE/ACM Trans. Netw. (TON) 21 (3), 720–733.
Machanavajjhala, A., Gehrke, J., Kifer, D., Venkitasubramaniam, M., 2006. L-diversity: privacy beyond k-anonymity. In: International Conference on Data Engineering. pp. 24–24.
Mohammed, N., Fung, B., Hung, P.C., Lee, C.-K., 2009. Anonymizing healthcare data: a case study on the blood transfusion service. In: Proceedings of the 15th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. ACM, pp. 1285–1294.
Newell, B.C., 2014. Local law enforcement jumps on the big data bandwagon: automated license plate recognition systems, information privacy, and access to government information. Maine Law Rev. 66 (66), 398.
Shao, W., Chen, L., 2018. License plate recognition data-based traffic volume estimation using collaborative tensor decomposition. IEEE Trans. Intell. Transp. Syst. (99), 1–10.
Shin, H., Vaidya, J., Atluri, V., Choi, S., 2010. Ensuring privacy and security for lbs through trajectory partitioning. In: Mobile Data Management (MDM), 2010 Eleventh International Conference on. IEEE, pp. 224–226.
Sun, L., Axhausen, K.W., 2016. Understanding urban mobility patterns with a probabilistic tensor factorization framework. Transp. Res. Part B: Methodol. 91, 511–524.
Sun, L., Chen, X., He, Z., Miranda-Moreno, L.F., 2019. Pattern discovery and anomaly detection of individual travel behavior using license plate recognition data. In: Transportation Research Board (TRB) Annual Meeting.
Sun, Z., Zan, B., Ban, X., Gruteser, M., 2013. Privacy protection method for fine-grained urban traffic modeling using mobile sensors. Transp. Res. Part B: Methodol. 56 (5), 50–69.
Sweeney, L., 2002a. Achieving k-anonymity privacy protection using generalization and suppression. Int. J. Uncertain. Fuzz. Knowl.-Based Syst. 10 (05), 571–588.
Sweeney, L., 2002b. k-anonymity: a model for protecting privacy. Int. J. Uncertain. Fuzz. Knowl.-Based Syst. 10 (05), 557–570.
Terrovitis, M., Mamoulis, N., 2008. Privacy preservation in the publication of trajectories. In: Mobile Data Management, 2008. MDM'08. 9th International Conference on. IEEE, pp. 65–72.
Truta, T.M., Vinay, B., 2006. Privacy protection: p-sensitive k-anonymity property. In: null. IEEE, pp. 94.
Warren, I., Lippert, R., Walby, K., Palmer, D., 2013. When the profile becomes the population: examining privacy governance and road traffic surveillance in canada and australia. Curr. Issues Crim. Justice 25 (2), 565–584.
Wong, R.C.-W., Li, J., Fu, A.W.-C., Wang, K., 2006. (α, k)-anonymity: an enhanced k-anonymity model for privacy preserving data publishing. In: Proceedings of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. ACM, pp. 754–759.
Yin, M., Sheehan, M., Feygin, S., Paiement, J.-F., Pozdnoukhov, A., 2018. A generative model of urban activities from cellular data. IEEE Trans. Intell. Transp. Syst. 19 (6), 1682–1696.
Zang, H., Bolot, J., 2011. Anonymization of location data does not work:a large-scale measurement study. In: International Conference on Mobile Computing and Networking, MOBICOM 2011, Las Vegas, Nevada, USA, pp. 145–156.
Zangui, M., Yin, Y., Lawphongpanich, S., Chen, S., 2013. Differentiated congestion pricing of urban transportation networks with vehicle-tracking technologies. Transp. Res. Part C: Emerg. Technol. 36, 434–445.
Zhan, X., Li, R., Ukkusuri, S.V., 2015. Lane-based real-time queue length estimation using license plate recognition data. Transp. Res. Part C: Emerg. Technol. 57, 85–102.