Transportation Research Part B 56 (2013) 50–69

Contents lists available at ScienceDirect

Transportation Research Part B

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

Privacy protection method for fine-grained urban traffic modeling using mobile sensors

Zhanbo Sun [a], Bin Zan [b], Xuegang (Jeff) Ban [a,*], Marco Gruteser [b]

[a] *Department of Civil and Environmental Engineering, Rensselaer Polytechnic Institute (RPI), 110 Eighth Street, Room JEC 4034, Troy, NY 12180-3590, United States*
[b] *WINLAB, Rutgers University, 671 Route 1 South, North Brunswick, NJ 08902-3390, United States*

ARTICLE INFO

Article history:
Received 5 November 2012
Received in revised form 18 July 2013
Accepted 19 July 2013

Keywords:
Mobile traffic sensors
Fine-grained urban traffic modeling
Privacy protection
Privacy by design
Adversary model

ABSTRACT

With the ubiquitous nature of mobile sensing technologies, privacy issues are becoming increasingly important, and need to be carefully addressed. Data needs for transportation modeling and privacy protection should be deliberately balanced for different applications. This paper focuses on developing privacy mechanisms that would simultaneously satisfy privacy protection and data needs for fine-grained urban traffic modeling applications using mobile sensors. To accomplish this, a virtual trip lines (VTLs) zone-based system and related filtering approaches are developed. Traffic-knowledge-based adversary models are proposed and tested to evaluate the effectiveness of such a privacy protection system by making privacy attacks. The results show that in addition to ensuring an acceptable level of privacy, the released datasets from the privacy-enhancing system can also be applied to urban traffic modeling with satisfactory results. Albeit application-specific, such a "Privacy-by-Design" approach would hopefully shed some light on other transportation applications using mobile sensors.

1. Introduction and motivation

Advances in wireless communications have prompted the rapid deployment of *mobile traffic sensors* that are able to move along with the flow they are monitoring. They provide an alternative to fixed-location sensors, such as loop detectors, that currently dominate the traffic detection systems. Broadly speaking, mobile traffic sensors include any monitoring or data collection system with a device that can move with the traffic flow. They include probe vehicles (such as those equipped with Electric Toll Collection (ETC) tags), cellular phones, portable global positioning system (GPS) devices (like GPS-enabled smart phones or navigation systems), Bluetooth Mac Address Matching (BMAM, see Wasson et al. (2008)), and vehicles in Connected Vehicles (NHTSA, 2013), among others. Most mobile sensors need to communicate with satellites (GPS), cellular towers (cell phones), or dedicated roadside infrastructure (ETC, Connected Vehicles, BMAM) to derive the position of the mobile component, its speed, and other relevant information. Here we focus on mobile sensors such as GPS, which can provide detailed tracking capabilities, including detailed location traces of individuals or vehicles. These capabilities, and the data they provide, may promise great advances in science and engineering.

In transportation, mobile sensing data have been used to extract *city-scale* urban knowledge on land use (Toole et al., 2012), human mobility patterns (Gonzalez et al., 2008), and urban congestion patterns such as travel times (Yuan et al., 2010). Such large-scale information is important since it reveals the "big picture" of urban traffic. In this paper, we focus on an equally important problem, the so-called *fine-grained* urban traffic modeling using mobile sensors (Ban and Gruteser,

* Corresponding author. Tel.: +1 (518) 276 8043; fax: +1 (518) 276 4833.
E-mail addresses: sunz2@rpi.edu (Z. Sun), zanb@winlab.rutgers.edu (B. Zan), banx@rpi.edu (Xuegang (Jeff) Ban), gruteser@winlab.rutgers.edu (M. Gruteser).

0191-2615/$ - see front matter © 2013 Elsevier Ltd. All rights reserved.
http://dx.doi.org/10.1016/j.trb.2013.07.010

2012), which concerns the detailed modeling of urban traffic states and performance measures, such as the real-time performance of urban traffic signals. This is critical for daily traffic operations and control, which previously has been studied using mainly fixed-location sensor data. Mobile data can provide alternative perspectives which, however, also impose great challenges.

One of these challenges is addressed in this paper: the selection of which mobile data elements to be collected and used. To this end, two important issues are considered. The first is how to satisfy the need for information extraction, i.e., data for transportation and especially traffic modeling purposes. Mobile data are fundamentally different from data collected via traditional means: they are more detailed spatially, but usually only provide a sample of the entire flow (Ban et al., 2011). As a result, choosing the form of mobile data to be collected and used will have profound implications on the development of new modeling techniques (Ban et al., 2009, 2011; Hao et al., 2012; Hofleitner et al., 2012). The second concern is how to address the privacy issues evoked by collecting mobile data, such as location traces from individual drivers. Such concerns can slow down or impede the adoption of new technologies, as experienced by Google Street View in parts of Europe (Claburn, 2009). These two sometimes conflicting needs, data for modeling and privacy protection, are related to many mobile-sensor-based applications and need to be addressed in a holistic manner. Here we focus on urban traffic modeling applications.

To date, transportation modeling and privacy protection are largely disconnected. On the one hand, with the primary goal of extracting as much information as possible, transportation modeling researchers have traditionally sought the greatest and most finely detailed data available. This has been done by either ignoring privacy issues completely, or by hoping that primitive privacy schemes such as simple anonymization will be sufficient to protect privacy. However, simple anonymization is not enough to protect privacy, as shown in Hoh et al. (2007). Recently proposed or deployed mobile sensing-based (or similar) systems emphasize privacy more, but mainly from a policy perspective (Jacobson, 2007) or by applying a limited set of privacy techniques (He et al., 2002; Demers et al., 2006). On the other hand, privacy experts (Kargupta et al., 2003; Hoh and Gruteser, 2005) have been focusing on designing privacy algorithms to protect individuals' privacy, without paying much attention to the real-world transportation applications. As a result, a large proportion of location data are either hidden or perturbed by these privacy protection algorithms. The dataset released after applying such privacy algorithms can rarely be used for fine-grained urban traffic modeling. In a nutshell, the full needs of modeling and privacy protection cannot be satisfied simultaneously by the current practice. As mobile data becomes more widespread, this issue is becoming increasingly critical.

Recently there are trends to simultaneously consider privacy protection and traffic modeling needs (Hoh et al., 2008; Ban and Gruteser, 2010). This is achieved when researchers are conscious of the effects of applying privacy schemes to data when developing modeling methods, and conscious of data needs when designing privacy preserving mechanisms. Privacy methods need to be application-specific. Different types of applications (e.g., transportation planning, traffic operations, safety, etc.) may need different types of data, and the applicable privacy algorithms need to be designed accordingly. However, the concept of "co-designing" privacy algorithms with modeling methods to simultaneously satisfy both privacy protection and data needs should apply generally, to different applications. This actually follows the "Privacy-by-Design" concept in Cavoukian (2009): instead of applying policies and techniques to relieve privacy concerns in an existing system (i.e., systems already designed and built), privacy mechanisms should be integrated deliberately and consistently into the system design (e.g., system structure, hardware design, data processing, applications, etc.); as such, they should be co-developed. One example of this is the virtual trip line (VTL) concept proposed in Hoh et al. (2008). VTLs can be used to regulate where and when mobile data should be collected to satisfy the needs of both traffic modeling and privacy protection. The effectiveness of VTL has been tested for both freeway (Herrera et al., 2010) and urban arterial modeling (Ban et al., 2009, 2011). For urban traffic modeling, however, VTL was only tested for applications related to isolated intersections, such as delay or queue length estimation. The question is: when considering an urban corridor or network, how should the VTL method be enhanced to ensure both privacy and data needs for fine-grained urban traffic modeling? This paper focuses on answering this question.

A VTL zone system is proposed in this paper for privacy protection in fine-grained urban traffic modeling applications. The VTL zone system combines access control and privacy protection techniques, which we believe is the most suitable system for urban fine-grained applications. To obtain higher levels of privacy, different filtering approaches are proposed in the VTL zone system. Traffic-knowledge-based adversary models are also developed, which are then used to attack the privacy-aware datasets. The overall performance of the system and different filtering approaches are evaluated, with respect to privacy protection and data needs for traffic modeling. The results indicate that the proposed system and related filtering approaches (especially those based on entropy and individual tracking probability) can properly balance the needs for privacy protection and traffic modeling.

2. Literature review

In this section, the current literature on privacy in transportation is summarized. We also briefly summarize the research in the closely related field of *location privacy,* which is concerned with the collection and use of location traces.

2.1. Privacy research in transportation

Mobile sensing is considered as one of the Intelligent Transportation System (ITS) technologies. With the ubiquitous applications of ITS, privacy issues are becoming increasingly important and need to be addressed carefully in transportation

(Garfinkel, 1996). Current privacy research in transportation is mainly policy-oriented; a summary of the policy-oriented privacy studies and efforts is provided in Table 1. Kokotovich and Munnich (2007) specified five dimensions of the privacy web. They stated that ITS-related privacy concerns fall under the informational and behavioral dimensions. Douma et al. (2008) analyzed existing federal laws and legal doctrines applicable to privacy in transportation systems and technology, and they claimed that the existing laws do little to protect individuals' privacy against the violations from ITS technologies. According to Cottrill (2009), privacy methods can be broadly categorized as technique-based and policy-based, who then concluded that "there is little consistency or certainty when it comes to the place of privacy in relation to ITS applications." The National Vehicle Infrastructure Integration (VII) coalition (now called Connected Vehicles) proposed a policy framework to address privacy issues in VII. However, as Cottrill (2009) pointed out, "These limits are, much like the privacy principles, fairly vague…" Therefore, while these privacy policies are important and provide insightful guidance on privacy protection, they usually lack detail, and need to be further materialized by specific (technical) schemes. In transportation modeling and land use, there is a well-studied privacy-preserving method called *population synthesis* (Beckman et al., 1996; Muller and Axhausen, 2011), which combines different data sources to produce synthetic representations of individuals. However, such synthetic data are not sufficient for fine-grained urban traffic modeling, for which data regarding *true* individual behavior (such as individual travel times) are needed.

Privacy issues become more critical when dealing with mobile data. For every object involved in a transportation activity (e.g., vehicle, driver, goods, etc.), there are signatures (e.g., ID, license plate, driving behavior, etc.) associated with it. Notice that these signatures do not have to be unique, as long as they can be applied to identify one object, or a group of objects. Now consider a person who wants to identify this object and discover some privacy information for malicious purposes (referred to as the *adversary* hereafter in the paper; a more precise definition can be found in Shokri et al. (2011)). If an adversary has access to the trace (or a part of the trace) of this object obtained from mobile data, it is possible to link and track the signatures for a significant distance/time period. If the adversary finds out some sensitive locations (e.g., gas station, office building, residential area, warehouse, etc.) along the location trace, it is then not hard to identify the object. For fear of being identified or re-identified (even if the traces are anonymous), there are usually great privacy concerns associated with applications that collect and use location trace information. In this regard, privacy protection using purely policy-based methods may not work well; advanced technical approaches also need to be considered.

2.2. Location privacy

The field of location privacy is concerned with technical approaches to address the privacy issues associated with collecting/processing certain location (mobile) data elements. These technical approaches play an important role in supporting the developments of privacy policy and regulation, which is a fast growing research area (Duckham and Kulik, 2006; Krumm, 2009). An overview of privacy protection techniques is provided in Table 2.

One of the important privacy techniques is *anonymization* (Sweeney, 2002; Rass et al., 2008; Stenneth and Yu, 2010), which guarantees the anonymity of an object, including static (using one *pseudonym*, i.e., a randomly generated ID, throughout the dataset) or dynamic pseudonyms (periodically updating the current ID with randomly generated pseudonym), and pure anonymity (removing the IDs for all the data points completely). However, pseudonyms are subject to privacy breaches with hidden information and domain knowledge. For example, Machanavajjhala et al. (2006) pointed out when the sensitive attributes in a dataset are of little diversity, or when the adversaries have access to external data sources, pseudonyms can be easily breached. Hoh et al. (2007) showed, using a dataset of a week-long anonymous GPS traces from 239 drivers, that they were able to find home locations of 85% of a subset of 65 drivers. Therefore, using pseudonyms alone is not sufficient for proper privacy protection. On the other hand, pure anonymity, when the ID of each data point is completely removed, is not suitable for the data needs of fine-grained urban traffic modeling, because location trace information is lost.

More sophisticated approaches have been developed to enhance anonymity, mainly by perturbing data accuracy or restricting the release of certain location information data points; called *obfuscation* (Ardagna et al., 2007). For example, *location perturbation* methods (Agrawal and Srikant, 2000; Kargupta et al., 2003; Gruteser and Grunwald, 2003; Gedik and Liu, 2005) try to preserve privacy by perturbing or reducing the accuracy of either spatial or temporal information in order to satisfy the k-anonymity (see the definition in Appendix A). Location data perturbed by such methods, however, can rarely

Table 1
Policy-oriented privacy studies and efforts.

Policy-oriented privacy studies and efforts	Contributions
Briggs and Walton (2000)	Develop guidelines and models for the management of sensitive data collected through ITS
ITSA (2001)	Adopt Fair Information and Privacy Principles
Clarke (2001)	Discuss the technologies which evoke privacy concerns and their implications
Douma et al. (2008)	Study how ITS technologies fit into U.S. privacy law
Kokotovich and Munnich (2007)	Specify five dimensions of the privacy web; present case studies of ITS related applications
Jacobson (2007)	Propose the privacy policy framework for VII (now Connected Vehicles)
Cottrill (2009)	Review related policy and techniques for privacy protection in ITS applications
ISO (2009)	Give general guidelines to develop ITS standards and systems on data privacy aspects

Table 2
Technical approaches for privacy protection.

Category	Related work	Contributions	Limitations
Static pseudonym	Stenneth and Yu (2010)	Mode homogeneity anonymization (*k*-anonymity) for location traces	For location trace data, such method is subject to privacy breaches with external data sources
Various pseudonyms	Rass et al. (2008)	Break long traces into multiple short traces, using different pseudonyms for short traces	Subject to privacy breaches under certain conditions
Pure anonymity	Sweeney (2002)	*k*-Anonymity for privacy protection. The identifier of each data point should be removed. Not suitable for location privacy	Pure anonymity is expensive to implement
Location perturbation	Agrawal and Srikant (2000)	Adding random noise to the sensitive data, Not suitable for location privacy	Location Information can get severely degraded; such datasets cannot be applied for fine-grained urban traffic applications
	Kargupta et al. (2003)	Show that random data distortion are subject to attacks using spectral filters	
	Gruteser and Grunwald (2003)	Propose spatial cloaking (and temporal) cloaking for anonymous usage of Location-Based Services	
	Gedik and Liu (2005)	Perturb location information by replacing it with a spatial range	
Reduce sampling frequency	Tang et al. (2006)	Suggest to use larger intervals between two location reports	Not enough data to support fine-grained urban traffic applications
Location hiding	Beresford and Stajano (2004)	Propose the concept of mix zone, in which different pseudonyms are used when a user entering or leaving the network	Cause a loss of data; Traffic modeling needs are not considered; The effectiveness of such algorithms need to be justified on real traffic networks
	Freudiger et al. (2007)	Further extend the mix zone approach to vehicular networks	
	Hoh et al. (2007)	Propose uncertainty-aware path cloaking algorithm, where the uncertainty is measured by entropy	
	Hoh et al. (2008)	Propose to use virtual trip line (VTL) to regulate speed and location reports	
Dummy traces	Kido et al. (2005) Lu et al. (2008)	Generate false location data, and send the false data along with the true data to the location-based service provider	Lead to unrealistic traffic estimation
	Nergiz et al. (2009)	Generate dummy traces for trajectory anonymization	

be used for applications that require fine-grained location traces. Similarly, *reducing sampling frequency* (Tang et al., 2006) or *using dummies* (Kido et al., 2005; Lu et al., 2008; Nergiz et al., 2009) may also severely degrade location information. It turns out that these approaches either cannot effectively guarantee privacy, or may filter out too much information to satisfy data needs from a modeling perspective.

Another frequently used approach is called *location hiding* (Beresford and Stajano, 2004; Freudiger et al., 2007; Hoh et al., 2007). From a transportation point of view, this approach is very appealing since many transportation applications have no problem removing location traces at less important places. However, real-world road topology is seldom considered in previous research. More importantly, privacy experts are mainly focusing on designing privacy protection mechanisms, and transportation/traffic modeling needs are usually overlooked. Recently, Hoh et al. (2008) proposed the idea of Virtual Trip Lines (VTLs). VTLs are geographic markers that indicate where vehicles should provide location updates. Ban and Gruteser (2010) further showed that by using VTLs to regulate location and speed reports, the data needs for intersection modeling such as signal performance measurement can be satisfied, while privacy can be simultaneously protected. However, in the context of network-wide urban traffic modeling, the effectiveness of VTL methods need to be further justified.

3. VTL zone system for fine-grained urban traffic modeling

3.1. Privacy definition for fine-grained urban traffic modeling

It is generally understood that the longer an object can be tracked, the more likely it can be identified. Therefore in this paper, *privacy* is defined as the untrackability of a moving object for a certain distance/time. Notice that the actual criteria for such distance or time can be user- adaptive: users can make this decision based on the level of privacy they are comfortable with. Since urban arterial traffic modeling is our focus here, the concept of privacy protection for arterial traffic modeling can be more specifically defined as follows.

For urban environments, we assume that to satisfy modeling needs, some short traces of vehicles (say a few hundred feet) around an intersection are available. Here we focus on intersections, especially signalized intersections, since they are the most critical for urban traffic. We then define *privacy* as the unlinkability of the short traces of the same vehicle over multiple (say *N*) intersections. Again, *N* is a parameter that implies different levels of privacy concerns and should be user-specific. In this paper, *N* is set to 2, which provides the highest level of privacy protection. This means that one can collect short vehicle traces around one intersection, but such short traces should not be linked together for the same vehicle for 2 or more intersections.

With respect to the metrics of privacy (unlinkability in this paper), *k-anonymity* and *entropy* have been frequently used in the location privacy field (see definitions in Appendix A). Both metrics are used to measure the level of unlinkability among a group of users. From an individual's viewpoint, however, they are not very intuitive. An individual user cares more about the probability that he/she could be successfully tracked, rather than the privacy on a system-wide metric. In this paper, to measure unlinkability, we propose to use the *individual tracking probability* as the metric for individuals, and *entropy* as the metric for a group of users. A lower value of individual tracking probability, or a larger entropy value indicates a higher level of privacy (unlinkability). More detailed explanations will be provided in Section 4.

3.2. VTL-zone system for fine-grained urban traffic modeling

In general, privacy protection approaches can be grouped into two categories (Briggs and Walton, 2000; Anderson, 2008): (i) controlling the access to privacy information, for example, password, software/hardware enforcement, and privacy agreement, etc.; (ii) Giving access to processed information only after privacy-protection techniques (e.g., anonymization and obfuscation) have been applied. Both approaches may work for some specific applications. However, there is no single approach that can solve all privacy problems for all applications. For the first approach, to ensure privacy, sophisticated secure systems (in terms of both hardware and software) usually need to be deployed, which can result in very high costs and/or overhead. With respect to the second approach, we showed earlier that some privacy techniques (e.g., simple anonymization) could be subject to privacy breaches under certain conditions.

In this paper, we propose the VTL zone system for privacy protection in fine-grained urban traffic modeling applications. The VTL zone system is a combination of access control and sophisticated privacy protection techniques. The system consists of VTLs that are pre-defined, *virtual* geographic markers on roadways; see the dashed lines in Fig. 1. When crossing a VTL, the mobile sensor equipped in a vehicle will report its location and speed information. A VTL zone is the area between two VTLs, one upstream and one downstream of a signalized intersection; see Fig. 1. These two VTLs are specifically designed to include vehicle deceleration and acceleration processes due to traffic signals. Vehicle trace data are only collected within VTL zones; starting from the location sample right before a vehicle enters a VTL zone, and ending with the location sample right before a vehicle leaves a VTL zone. The traces of the same vehicle at different VTL zones will be assigned different random IDs (pseudonyms). Thus, due to the discontinuity of location traces between VTL zones, it would not be a trivial task for an adversary to track a vehicle across multiple zones (intersections).

As shown in Fig. 1, the system structure is comprised of a trusted location proxy server, and an application server, similar to that of the VTL system in Hoh et al. (2008). Location data are anonymized and filtered (if necessary, to ensure higher levels of privacy) in the location proxy server, and then transmitted to the application server. The access control and privacy techniques are both applied in the location proxy server. The application server does not have to be trustworthy, and we can always assume that the adversary has access to the application server. Since the privacy-sensitive information in such a system is only stored and processed at the proxy server, which is not accessible to outside users, the risk of it being breached can be dramatically reduced. As a result, the secure system for the proxy server does not need to be that sophisticated, compared to when only pure access control (i.e., to combine the proxy and application servers as a single server) is applied.

The system in Fig. 1 does not require a dedicated location proxy server for each intersection. Mobile sensors can transmit the raw data (from different intersections) to a trusted central location proxy server, where centralized anonymization and other privacy techniques can be implemented before the processed data are sent to the application server.

In summary, the VTL zone system collects location traces only within VTL zones. Among the collected location traces, additional privacy protection techniques can be applied in the location proxy server. For example, trace identifiers can be

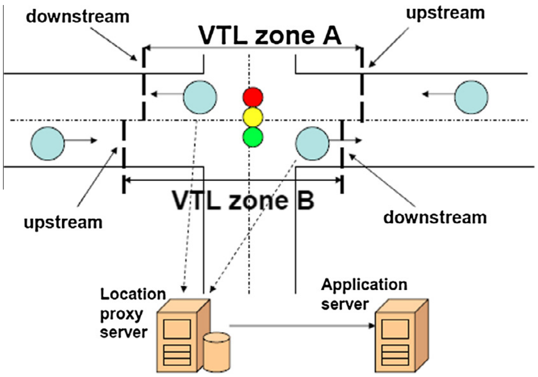

Fig. 1. System structure of the VTL-zone based method.

removed and random IDs assigned, filtering approaches can be applied (see Section 4 below), among others. As a result, only a subset of the location traces will be released to the application server for each zone. The released dataset offers a fixed level of privacy defined either by the individual tracking probability or entropy, yet the dataset can still satisfy the data needs for traffic modeling as shown later, in Section 7.

4. Filtering approaches

In this section, several filtering approaches are proposed. The *baseline* approach refers to releasing all of the traces collected in a VTL zone, which provides the least privacy (while the VTL zone system is applied), but the most usable data. To enhance privacy, part of the location traces need to be filtered out in the VTL zone. The filtering approaches include *random sampling*, *individual probability based* and *entropy based* approaches.

4.1. Baseline approach

In the baseline approach, all vehicle traces within a VTL zone are released. Due to the discontinuity of vehicle traces on the link between two VTL zones, it is not a trivial task for the adversary to keep track of the same vehicle between multiple zones. In fact, the baseline approach can be treated as the opposite of the *mix-zone* algorithm (Beresford and Stajano, 2004). A mix zone is defined as an area in which location traces cannot be released, so that it would be hard for an adversary to link upstream traces with downstream traces (which have different pseudonyms). Related works (Li et al., 2006; Freudiger et al., 2007; Buttyan et al., 2007; Dahl et al., 2010; Carianha et al., 2011) study the mix-zone approach in vehicular networks. Their focus is finding the optimal sizes and locations of mix zones to suppress traces so that privacy can be best protected, without much consideration of the application, or whether the released traces are sufficient for transportation modeling purposes. Our baseline approach, by contrast, defines areas where data should be collected. In particular, in order to satisfy the data needs for fine-grained urban traffic modeling, the baseline approach (and the VTL-zone method in general) releases location data near an intersection, which is exactly where mix zones are always deployed to suppress data. The authors believe that our proposed approach, by focusing on where data should be collected (instead of suppressed), minimizes the data collection effort and can better satisfy the data needs for applications. However, this approach also imposes challenges. For the baseline approach, as shown later in Section 7, a large proportion of linkage can be successfully built using certain adversary models. Therefore, we need to develop specific filtering methods to release only portions of the traces in a VTL-zone to guarantee privacy.

4.2. Random sampling

On top of the baseline approach, a random sampling approach can be applied to enhance the level of privacy protection. In particular, only a portion of the traces (say 50%) is randomly selected and released at each VTL zone. It is thus even harder for the adversary to continuously track traces of the same vehicle across VTL zones. To some extent, this approach is fairly naïve since the tracking probability of different location traces are not taken into consideration.

4.3. Individual tracking probability based filtering

Compared with random sampling, the individual tracking probability based approach uses the individual tracking probability as the privacy metric. It can thus release traces that are less likely to be tracked (smaller tracking probability) and suppress traces that are more likely to be tracked (higher tracking probability). For a released dataset that guarantees a 0.2 individual tracking probability, the statistical implication is that no more than one out of five vehicles can be successfully tracked. Consider the vehicle tracking problem in Fig. 2: for one vehicle trace at the downstream VTL zone c, and given a set of previously released traces in each upstream VTL zone, what is the probability that one vehicle trace at an upstream VTL zone (say v_1) will be for the same vehicle as the target trace in downstream VTL zone c? Below is a probabilistic interpretation of how the individual tracking probability is formulated.

Define discrete random variables C, V to capture these random events: a vehicle arrives at a downstream VTL zone, and a vehicle passes an upstream VTL zone, respectively. The continuous random variable T denotes the travel time. $P(C)$, $P(V)$ and $P(T)$ are the probabilities of the three random variables. We use lower case c, v and t to denote the instantiations of those variables. Now consider a vehicle that passes the downstream VTL zone C. We are interested in the probability that this vehicle also passed an upstream VTL zone V, taking travel time T from V to C. This conditional probability, denoted as $P(T, V|C)$, is referred to as the *individual tracking probability* in this paper.

Using the Bayesian Theorem, Eq. (1) can be easily derived:

$$P(T,V|C) = \frac{P(T,V,C)}{P(C)} = \frac{P(T,V,C)}{\int \sum_V P(T,V,C)dT} \tag{1}$$

The key term in Eq. (1) is the joint probability $P(T, V, C)$, which can be expressed as the product of three (conditional) probabilities in an Eq. (2), using the Chain Rule:

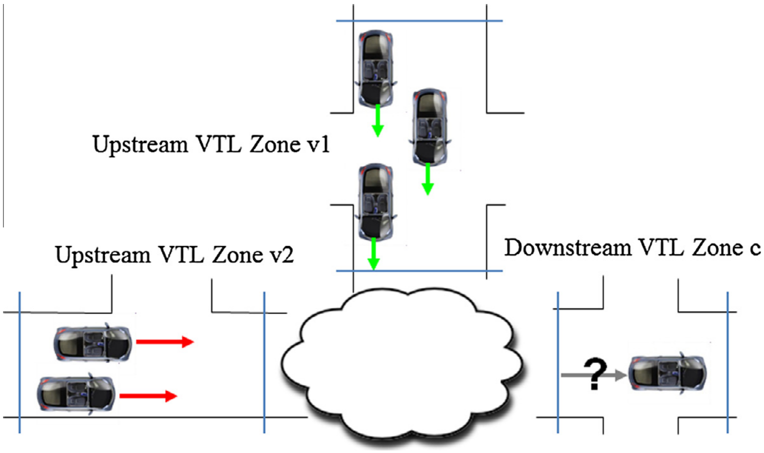

Fig. 2. Vehicle tracking between upstream and downstream VTL zones.

$$P(T,V,C) = P(T|V,C)P(C|V)P(V) \tag{2}$$

For the right side of Eq. (2), the first term is the probability of travel time, given the fact that the vehicle passes an upstream VTL zone *V*, and a downstream VTL zone *C*, in sequence, which follows a *travel time distribution* $p_T^{v\to c}(t)$, as defined later in this section. The second term is the probability of passing a downstream VTL zone, given the fact that this vehicle passes an upstream VTL zone, which is defined as the *path likelihood* $\rho_{v\to c}$, later in this paper. The third term is the prior probability of passing an upstream VTL zone. Eq. (2) does not specify any mathematical form of the probabilistic distributions. In theory, any form of distribution would work in this probabilistic model, as long as it provides a meaningful reflection of the real-world traffic situation.

4.3.1. Path likelihood

Among the location traces that pass an upstream VTL zone (denoted as v), the path likelihood from v to a downstream VTL zone c is defined as the proportion of the traces that go through both v and c, as represented in Eq. (3).

$$\rho_{v\to c} = \frac{\sum_{d\in D_v} k_{v\to c}^d}{\sum_{d\in D_v} k_v^d} \tag{3}$$

In Eq. (3), d is the vehicle pseudonym; D_v is the set containing all the pseudonyms of vehicles passing v; $k_v^d$is the number of times vehicle d goes through v (e.g., in the historical dataset); and $k_{v\to c}^d$ is the number of times vehicle d goes through both v and c in sequence. Statistically, if one vehicle passes the upstream VTL zone (v), the path likelihood $\rho_{v\to c}$ indicates the probability that this vehicle will also go through the downstream VTL zone (c), among other possible choices.

4.3.2. Travel time distribution

There have been some previous studies about arterial travel time distributions along a route (Kwong et al., 2009; Hofleitner et al., 2012). However, here we are concerned with the travel time distribution, i.e., $p_T^{v\to c}(t)$, between two urban intersections, v and c. Since there may be multiple routes between two urban locations (especially when they are far away from each other), the distribution actually contains travel times among all possible *used* routes. To the best of the authors' knowledge, the research of quantifying such travel time distributions (i.e., between two urban locations) is rather sparse in the literature. Since this is not the focus of this paper, we simply assume that the vehicle travel time probability between two VTL zones (e.g., VTL zone v to c) follows a three-parameter log-normal distribution, as shown in Fig. 3 (statistical results are drawn from a VTL zone pair from the NGSIM dataset; see Cambridge Systematics (2007)).

The mathematical form of the 3-parameter lognormal distribution is given in Eq. (4). Here θ is the shift parameter, corresponding to the free flow travel time between the two VTL zones; σ is the shape parameter and μ is the scale parameter. The travel time distribution can be estimated by Least Squares Estimation (LSE) (see Zan et al. (2011) for more details). Future research on better quantifying the travel time distributions between two urban locations is needed, as noted in Section 8.

$$p_T^{v\to c}(t) = \begin{cases} \frac{1}{\sigma\sqrt{2\pi}(t-\theta)} e^{-\frac{(\log(t-\theta)-\mu)^2}{2\sigma^2}} & \mathit{for}\ t > \theta \\ 0 & \mathit{for}\ t \leqslant \theta \end{cases} \tag{4}$$

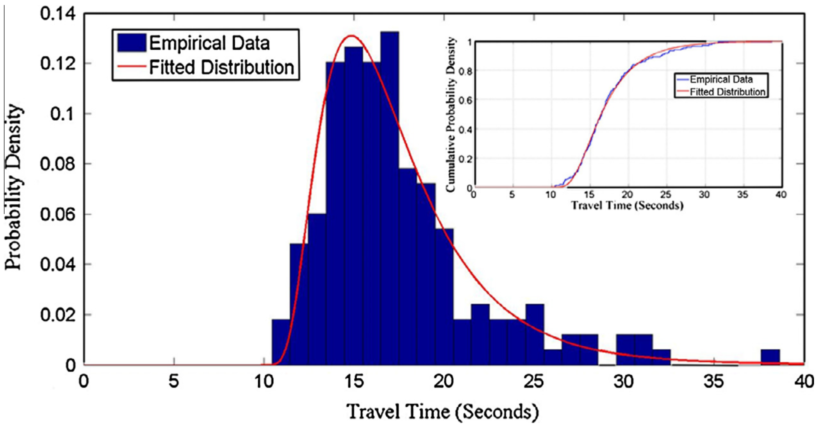

Fig. 3. 3-Parameter log-normal travel time distribution.

4.3.3. Individual tracking probability

By incorporating Eq. (2) into Eq. (1), the individual tracking probability (density) of any instantiation ($T = t$, $V = v$, $C = c$) can now be expressed as Eq. (5), in which $P(T = t|V = v, C = c)$ is the travel time probability and $P(C = c|V = v)$ is the path likelihood as defined previously.

$$P(T=t,V=v|C=c)=\frac{P(T=t|V=v,C=c)P(C=c|V=v)P(V=v)}{\int\sum_{V}P(T|V,C=c)P(C=c|V)P(V)dT}=\frac{p_T^{v\to c}(T=t)\rho_{v\to c}P(V=v)}{\int\sum_{v'\in V\backslash c}p_T^{v'\to c}(T=t)\rho_{v'\to c}P(V=v')dT}$$
$$=\frac{p_T^{v\to c}(T=t)\rho_{v\to c}P(V=v)}{\sum_{v'\in V\backslash c}\rho_{v'\to c}P(V=v')\int p_T^{v'\to c}(T=t)dT} \quad (5)$$

Note that in the above derivation, the probability distribution of travel time is continuous. In practice, however, the observations of travel time are discrete. It is therefore necessary to discretize the continuous travel time into a finite number of time intervals, and use the probability of a discrete time interval to represent the travel time probability of an instance that falls into this interval. In this paper, an equal-width discretization scheme (Cios et al., 1998) is applied. The number of time intervals can be calculated by the rule of thumb formula, as shown in Eq. (6). Here n_T is the number of discrete time intervals; M is the number of travel time observations in the historical data (which are used to estimate the travel time distributions); C is the number of VTL zone pairs, i.e., for any $v' \in V/c$ and c.

$$n_T=\frac{M}{3C} \quad (6)$$

After the number of time intervals is specified, the width of discrete intervals (W), and the corresponding cut points of each interval, can be determined easily. Based on such a discretization scheme, the discrete form of the individual tracking probability for a given instance d from an upstream VTL zone v to the current VTL zone c is now defined as $p^d_{v\to c}$, which can be approximated by the following equation:

$$p^d_{v\to c}\triangleq\frac{p_T^{v\to c}(T=t^d_{v\to c})W\rho_{v\to c}P(V=v)}{\sum_{v'\in V\backslash c}\rho_{v'\to c}P(V=v')\sum_{d'\in D_{v'\to c}}p_T^{v'\to c}(T=t^{d'}_{v'\to c})W}=\frac{p_T^{v\to c}(T=t^d_{v\to c})\rho_{v\to c}P(V=v)}{\sum_{v'\in V\backslash c}\rho_{v'\to c}P(V=v')\sum_{d'\in D_{v'\to c}}p_T^{v'\to c}(T=t^{d'}_{v'\to c})} \quad (7)$$

Here $t^d_{v\to c}$ is the (observed) travel time for vehicle d from zone v to c (c is the current VTL zone). And $D_{v'\to c}$ is the set containing all of the pseudonyms of vehicles going from zone v' to zone c. Note that here we use the probability density of a travel time instance (i.e., $p_T^{v\to c}(T=t^d_{v\to c})$) to approximate the average probability density of its corresponding discrete interval. Therefore, the discrete form of travel time probability (for a given instance $(T=t^d_{v\to c})$ is essentially $p_T^{v\to c}(T=t^d_{v\to c})W$, which is approximately the area under the travel time probability density function for the discrete time interval. Furthermore, we use the summation of the probabilities of all the travel time instances for a particular VTL zone pair (i.e., $\sum_{d'\in D_{v'\to c}}p_T^{v'\to c}(T=t^{d'}_{v'\to c})W$) to approximate the integral of travel time probability density as shown in Eq. (5), i.e., $\int p_T^{v'\to c}(T=t)dT$. It is worth mentioning that this integral equals 1 in theory; the approximation term $\sum_{d'\in D_{v'\to c}}p_T^{v'\to c}(T=t^{d'}_{v'\to c})W$ should also be 1 if the instances are evenly drawn from the underlying distribution (Cios et al., 1998). However, in reality, the approximation term may not be 1, depending on how these instances are distributed. We find that the approximation term does produce better results since it considers real-time traffic information. In practice, the individual tracking probability is computed in real time (i.e., considering the vehicles travel in the network during the past 1 or 5 min). Therefore the number of travel time instances (for a travel time distribution) is usually less than the number of discrete time intervals. This may lead to an over-estimation of the individual tracking probability, resulting in better privacy performance.

Now assume that $P(V)$ is uniformly distributed (in a discrete form), implying that the prior probability of passing an upstream VTL zone is the same as passing others. Eq. (8) can then be obtained.

$$p^d_{v\to c} = \frac{p_T^{v\to c}(T = t^d_{v\to c})\rho_{v\to c}}{\sum_{v'\in V\backslash c, d'\in D_{v'\to c}} p_T^{v'\to c}(T = t^{d'}_{v'\to c})\rho_{v'\to c}} \tag{8}$$

In Eq. (8), the individual tracking probability of vehicle $d \in D_c$ (i.e., $p^d_{v\to c}$) is now formulated as the product of the path likelihood and travel time probability, normalized over all of the suspect vehicles, i.e., any vehicle d' that has arrived at the current VTL zone c from an upstream VTL zone v' other than c. If the individual tracking probability of vehicle d is smaller than a pre-defined level (e.g., 0.2), the location trace of this vehicle can be released; otherwise, the trace should not be released. In practice, the path likelihood and travel time distribution can be updated in real time (e.g., they can be calculated in the location proxy server, based on data collected during the previous time period), which can incorporate the traffic dynamics at different times of a day.

4.4. Entropy based filtering

The entropy value of a specific location trace (now treated as a random variable) in the downstream VTL zone c, can be calculated using Eq. (9). Here v is any upstream VTL zone that is not c, and d is any vehicle at v. Eq. (9) indicates that the entropy of the specific location trace at c is the summation of $-p^d_{v\to c}\log_2 p^d_{v\to c}$ for all possible vehicles that previously passed some upstream VTL zones. The specific vehicle trace (in VTL zone c) is safe to be released if the entropy value H is larger than α (e.g. $\alpha = 2.0$), where α is a predefined confusion level that characterizes the desired degree of privacy. A higher α indicates more uncertainty (or level of confusion) in terms of identifying the same vehicle, thus providing better privacy.

$$H = -\sum_{v\in V\backslash c, d\in D_{v\to c}} (p^d_{v\to c} log_2 p^d_{v\to c}) \tag{9}$$

We know from Eq. (9) that the metrics of entropy and individual tracking probability are mathematically related. As an illustrative example, when the entropy value is 1, this is a way to illustrate that we have two possible outcomes (corresponding to two vehicle traces from an upstream VTL zone to the downstream VTL zone) both having 0.5 individual probability. However, in extreme cases, entropy and individual tracking probability may have different privacy implications. For example, consider the scenario of four vehicles, with individual tracking probabilities: 0.2, 0.2, 0.2 and 0.4. The entropy value of this original set is therefore $3 * 0.2 * \log_2 0.2 + 0.4 * \log_2 0.4 = 1.92$. Now consider another scenario with only three vehicles, all of them having 0.33 individual tracking probabilities. From an individual's perspective, the second scenario provides better privacy (since 0.33 is smaller than 0.40). However, from the entropy perspective, the entropy value for the second scenario is 1.58, implying a poorer system-wide privacy level. In this paper, we therefore use both the individual tracking probability and entropy as privacy metrics.

5. Traffic-knowledge-based adversary model

To evaluate the filtering approaches in the previous section, a traffic-knowledge-based adversary model is proposed. This adversary model takes into account real-time travel time information and signal timing information, which can better capture traffic dynamics compared with statistical models. Here *deterministic* attacks are made to the released traces by attempts to link traces from two VTL zones. Two cases are considered.

5.1. Tracking between two neighboring VTL zones (Case 1)

Consider two neighboring VTL zones (Z_1 and Z_2), which cover two consecutive intersections, with one link (L_{12}) in between, as shown in Fig. 4. On L_{12}, since vehicles are usually proceeding at a speed close to the free-flow speed, the travel time on L_{12} is relatively stable. In this paper, the estimated travel time (referred to as T_{12}) can be estimated by the length of L_{12} divided by the estimated average speed on this link. The estimated average speed is calculated by taking the average of the speed of the last sample in Z_1, and the first sample in Z_2. Note that the estimated average speed (called "average" in the sense that this speed is averaged over two speed reports of two traces, rather than as an average speed for the whole population) may be different for a different pair of traces released at the upstream and downstream. It is thus able to capture

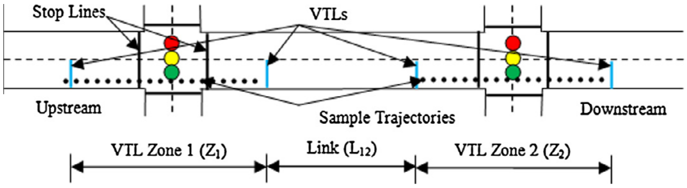

Fig. 4. VTL zones (Case 1).

instances when vehicles are not traveling at the free-flow speed (e.g., aggressive drivers, road constructions, recurrent traffic congestions, etc.). Hereafter in the paper, if the trace of vehicle n is released at Z_1, and the trace of vehicle m is released at Z_2, we use $T_{12}^{n,m}$ to represent the estimated travel time for this pair of released traces, between Z_1 and Z_2.

The next step is to look at a set of released traces (Ω_1), which go through Z_1 and a set of released traces (Ω_2)that arrive at Z_2. For a trace $n \in \Omega_1$, it is easy to tell when this vehicle leaves Z_1, referred to as T_1^n. For a trace $m \in \Omega_2$, if n and m belong to the same vehicle, the time that this vehicle enters Z_2 can then be approximated as $T_1^n + T_{12}^{n,m}$, referred as $T_2^{n,m}$. This is the estimated arrival time for this vehicle, based on the trace n and trace m. See Eq. (10).

$$T_2^{n,m} \approx T_1^n + T_{12}^{n,m} \tag{10}$$

If the travel time estimation is slightly relaxed with a threshold T_t, and trace m enters Z_2 within the time period $[T_2^{n,m} - T_t, T_2^{n,m} + T_t]$, this trace may belong to the same vehicle as trace n. In other words, for vehicle $m \in \Omega_2$, with T_2^m as the actual time entering Z_2, if $T_2^{n,m} - T_t \leqslant T_2^m \leqslant T_2^{n,m} + T_t$, m is added into a suspect list for trace n (denoted as S^n). This means that m may belong to the same vehicle as n. If S^n is not empty, the trace $\hat{k}$ can be chosen, which satisfies Eq. (11) as an inference. An *inference* is defined as an attack made by the adversary, by claiming one location trace in Z_2 belongs to the same vehicle as a trace released at Z_1.

$$\hat{k} = \underset{k\in S^n}{\operatorname{argmin}} |T_2^k - T_2^{n,k}| \tag{11}$$

Eq. (11) indicates that $\hat{k}$ and n belong to the same vehicle if the arrival time (of $\hat{k}$) at Z_2 is the closest to the estimated arrival time of vehicle n. If the inference is correct, i.e., $\hat{k}$ and n indeed belong to the same vehicle, the vehicle traces at the two neighboring VTL zones are successfully linked, which violates privacy.

Notice that there are many factors that can potentially impact the cardinality (size) of the suspect list S^n: for example, the adversary model (i.e., the travel time estimation method), the threshold T_t, and the number of users in the system. If the adversary model is very accurate, it is more likely to find some suspect vehicles whose estimated travel times are close to the revealed travel times (S^n is not empty in this case); if the threshold T_t is large, the cardinality of S^n may increase since a larger number of estimated travel times are considered to be "close enough" to the revealed travel times; if the number of users in the system is large or the dataset is dense, the cardinality of S^n may also increase.

In the numerical section (Section 7), the threshold T_t is set to 3 s, and the typical cardinality of S^n is less than 10. Of course, the actual distribution of the cardinality of S^n is subject to change for different scenarios. In particular, if S^n is empty, the estimated travel times will be too far away from the revealed travel times, implying that no inference can be made in this case. This means that the adversary cannot make attacks to the released mobile dataset. To avoid having an empty S^n, the adversary may increase the threshold T_t so that the set S^n is not empty. However, numerical experiments show that such adjustment have no direct impacts on the privacy performance. In other words, even though the adversary can manipulate the threshold to get larger suspect lists, and to avoid the scenario in which S^n is empty, doing so cannot guarantee more effective privacy attacks. Details of the numerical experiments are omitted here due to space limitations.

5.2. Tracking between two non-neighboring VTL zones (Case 2)

Consider two VTL zones that are not adjacent to each other (e.g., Z_1, Z_2 and Z_3, which cover a corridor with three intersections, and the adversary model is trying to link the vehicle traces from Z_1 to those in Z_3, as shown in Fig. 5). Following the same logic as in Case 1, the travel times on the links (L_{12},L_{23}) between two consecutive VTL zones are stable, which can be estimated as T_{12} and T_{23}, using the same approach as in Case 1.

Next, link the traces from a set of released traces (Ω_1) that go through Z_1 to a set of released traces (Ω_3) that go through Z_3. For trace $n \in \Omega_1$ and trace $m \in \Omega_3$, if n and m belong to the same vehicle, the time that this vehicle enters Z_3 can be approximated as Eq. (12).

$$T_3^{n,m} \approx T_1^n + T_{12}^{n,m} + T_{2,D}^{n,m} + T_{23}^{n,m} \tag{12}$$

$T_3^{n,m}$ is the estimated time that this vehicle enters Z_3; T_1^n is the time that this vehicle leaves Z_1; $T_{12}^{n,m}$ and $T_{23}^{n,m}$ are the estimated travel times for this pair of released traces (i.e., $n \in \Omega_1$ and $m \in \Omega_3$), on the link segments between Z_1 and Z_2, Z_2 and Z_3, respectively; and $T_{2,D}^{n,m}$ is the travel time of this vehicle within Z_2, which can be estimated via the delay pattern of Z_2; see Ban et al. (2009) for more details. Notice that the delay pattern of Z_2 is reconstructed using a set of released traces (Ω_2) that go through

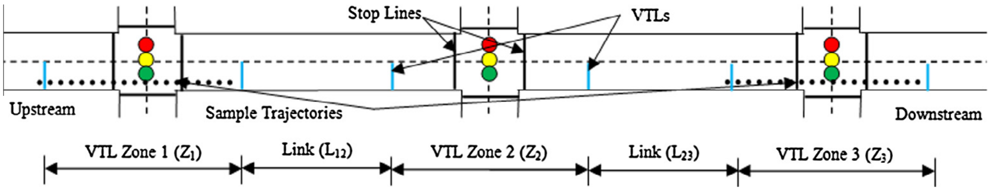

Fig. 5. VTL zones (Case 2).

Z_2, and the signal timing information. This model can therefore provide deterministic estimation of travel time for any imaginary vehicle entering a signalized intersection. For vehicle $m \in \Omega_3$, if $T_3^{n,m} - T_t \leqslant T_3^m \leqslant T_3^{n,m} + T_t$, we add m into a suspect list (S^n). If S^n is not empty, we can choose the vehicle $\hat{k}$ that satisfies Eq. (13) as an inference.

$$\hat{k} = \underset{k \in S^n}{\operatorname{argmin}} |T_3^k - T_3^{n,k}| \tag{13}$$

Note that if the two zones Z_1 and Z_3 are far away from each other, there might be multiple routes that a vehicle could take between these two zones. In this case, the adversary needs to figure out which is the most likely route, and then apply the method described in this subsection (which makes the tracking problem even harder). Details are not presented in this paper.

6. Evaluation criteria

In this section, the criteria used to evaluate the performances of the VTL zone method (especially with different filtering approaches) are described, with respect to both privacy protection and the data needs for fine-grained urban traffic modeling.

6.1. Privacy protection

In terms of privacy protection, the performance of the privacy models can be evaluated by applying adversary models. In this paper, two criteria are proposed for privacy evaluation purposes, namely, the percentage of correctly tracked traces (P_1), and the percentage of correct inferences (P_2). P_1 indicates the probability that the traces of one vehicle can be successfully linked at the two VTL zones. It is obtained by using the number of correct inferences divided by the total number of traces (which do not necessarily need to be released) going through both the VTL zones (e.g., both Z_1 and Z_2 in Case 1; Z_1 and Z_3 in Case 2). P_2 is obtained using the number of correct inferences divided by the total number of inferences, which indicates how accurate the inferences are (corresponding to the correctness concept in Shokri et al. (2011)). From the perspective of individuals, P_1 may seem more important, since it reveals the potential risk that one vehicle trace can be tracked. However, P_2 is also important, because it directly measures the accuracy of the inferences.

Notice that in our adversary models, one released vehicle trace in the first VTL zone (e.g., Z_1 in Case 1) can at most correspond to one inference (selected from the suspect list) based on Eq. (11) or Eq. (13). There are some situations in which the suspect list is empty, and therefore no inferences can be made.

6.2. Data needs for arterial traffic modeling

A tradeoff sometimes exists between privacy protection and data needs for traffic applications. To achieve a high level of privacy, transportation researchers may, to some extent, have to sacrifice the ease of traffic modeling. It is therefore important to make sure that after applying the privacy schemes, the released datasets can still be used for traffic modeling purposes, especially fine-grained urban modeling. Two criteria are used here: the percentage of the number of released traces in each VTL zone (compared with the total number of traces in the baseline approach); and the percentage of the number of cycles (out of the total number of cycles in the dataset) for which queue length estimation using mobile data (Ban et al., 2011) can be successfully performed. This is defined as the *success rate* (Ban et al., 2011). Note that the success rate of queue length estimation is used here as an indicator of traffic modeling application performance, though other measures could also be used for the same purpose.

7. Experiment and numerical results

In this section, the VTL zone system and the filtering approaches are evaluated, using both the privacy and modeling criteria defined in the previous section. The evaluation was done using NGSIM data collected at Peachtree St. in Atlanta, Georgia (Cambridge Systematics, 2007). Vehicle traces were collected using recognition techniques via video images, which provide an (almost) continuous tracking (with a 10-Hz sampling frequency) and 100% penetration of the real traffic flow. The geometry of the network is shown in Fig. 6. The network includes four signalized intersections, and one intersection controlled by stop signs. The data collected between 4:00 pm and 4:15 pm are used in this experiment. For the intersection controlled by stop signs (Peachtree St. & 13th Street), since the side street traffic made only marginal impacts on the main road traffic, this intersection is not considered in the scope of this experiment. Its two adjacent intersections (Peachtree St. & 12th Street, and Peachtree St. & 14th Street) are considered as a neighboring intersection pair. The signals are fixed-timed for both directions, with the same cycle length of 100 s. To protect privacy, VTL zones are deployed around the signalized intersections. The original long traces are divided into small segments, and the traces between two neighboring VTL zones are deleted. The baseline dataset can then be obtained. On top of that, the filtering algorithms remove a proportion of the traces. Based on the proposed system structure of the VTL zone method, it is assumed here that the adversary has access to the released vehicle traces (i.e., a sequence of time, location and speed information) within each VTL zone. This kind of data could be in real time,

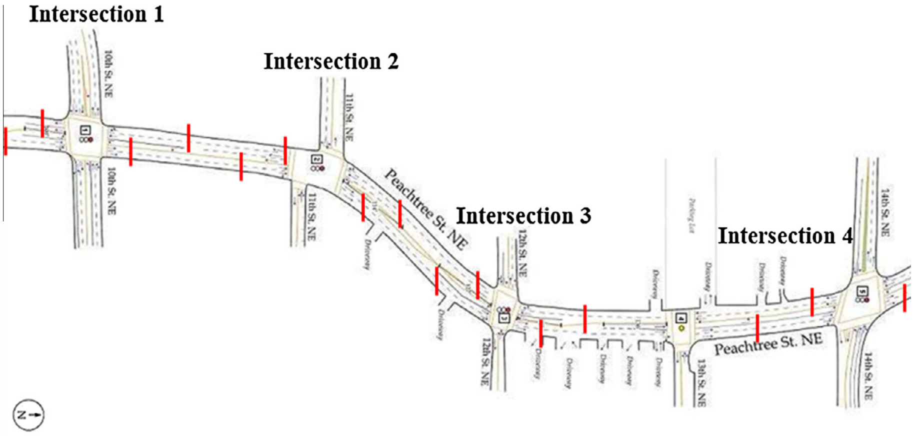

Fig. 6. Network geometry (the Peachtree St. Atlanta, Georgia).

or it could be archived historical data. Using the traffic-knowledge-based adversary model to attack the released datasets, the performances of the filtering algorithms are evaluated with respect to the privacy and modeling criteria. It is worth mentioning that the authors also conducted similar experiments on a much larger network in micro-simulation. The results and conclusions are similar to those presented here. One can refer to Sun et al. (2011) for further details.

7.1. Privacy performance evaluation of Case 1

In Case 1, privacy attacks are made to track the traces between two neighboring VTL zones. Table 3 shows the performance of the baseline. Column 2 of the table indicates the intersection pairs where privacy attacks are made. Column 3 is the number (percentage in the parenthesis) of released traces for vehicles going through both Z_1 and Z_2. Column 4 and Column 5 are the number (percentage in the parenthesis) of released traces within Z_1 and Z_2, respectively. Column 6 is the number of inferences made by the adversary model. Column 7 is the number of correct inferences. Column 8 corresponds to P_1 defined in Section 6, obtained by using Column 7 divided by Column 3 in the table. Column 9 corresponds to P_2 defined in Section 6, obtained by using Column 7 divided by Column 6.

Table 3 illustrates that the idea of releasing all of the traces within each VTL zone, i.e., the baseline, is somewhat useful to preserve privacy. However, this is not sufficient since there is still a large proportion of vehicle traces (ranging from 30.1% to 86.5%) that can be successfully tracked. The general performance of the adversary model however varies from one case (VTL zone pair) to another, depending on the travel time estimation between the upstream VTL zone and downstream VTL zone. In this regard, a smarter adversary model (i.e., for better travel time estimation) may lead to higher values of P_1 and P_2, making the system more vulnerable to privacy breaches; more discussions on this can be found in Section 8.

Note that the adversary model assumes that each released trace at Z_1 may also pass Z_2, although this may not always be the case due to vehicles getting off the road at minor intersections, or ending their trips in the middle of the VTL zone pair. Therefore, the number of inferences for a VTL zone pair depends on the number of released traces at Z_1. For example, if the number of released traces at Z_1 is 138 (see column 4 of pair 2_3 in Table 3), it should correspond to at most 138 inferences.

Table 4 shows the results for the 50% random sampling approach. As indicated in Column 4 and Column 5, about 50% of the total traces are released at each VTL zone. Compared with the baseline, random sampling is able to better protect privacy by filtering out some sample traces at each VTL zone, so that fewer traces can be tracked. However, with respect to the accuracy of the inference (P_2), the observed decrement is quite marginal.

Tables 5 and 6 indicate the results of the 2.0 entropy and the 0.2 individual probability filtering approaches for both the northbound and southbound traffic. Compared with Table 4, one can tell that while the released number of traces is similar here to the random sampling dataset (see Columns 4 and 5), the entropy and individual probability based filtering ap-

Table 3
Privacy performance of the baseline dataset (Case 1).

1 Traffic direction	2 VTL zone pair	3 No. of released traces (both Z_1 and Z_2)	4 No. of released traces (Z_1)	5 No. of released traces (Z_2)	6 No. of inferences	7 No. of correct inferences	8 Percentage of tracked traces (P_1) (%)	9 Percentage of correct inferences (P_2) (%)
Northbound	1_2	111 (100%)	119 (100%)	140 (100%)	107	49	44.1	45.8
	2_3	131 (100%)	138 (100%)	134 (100%)	136	98	74.8	72.1
	3_4	129 (100%)	138 (100%)	152 (100%)	133	62	48.1	46.6
Southbound	4_3	166 (100%)	286 (100%)	188 (100%)	220	50	30.1	22.7
	3_2	167 (100%)	173 (100%)	175 (100%)	170	128	77.1	75.3
	2_1	207 (100%)	221 (100%)	208 (100%)	215	179	86.5	83.3

Table 4
Privacy performance of the 50% random sampling dataset (Case 1).

1	2	3	4	5	6	7	8	9
Traffic Direction	VTL zone pair	No. of released traces (both Z_1 and Z_2)	No. of released traces (Z_1)	No. of released traces (Z_2)	No. of inferences	No. of correct inferences	Percentage of tracked traces (P_1) (%)	Percentage of correct inferences (P_2) (%)
Northbound	1_2	30 (27.0%)	62 (52.1%)	76 (54.3%)	38	11	9.9	28.9
	2_3	40 (30.5%)	74 (53.6%)	70 (52.2%)	62	33	25.2	53.2
	3_4	39 (30.2%)	72 (52.2%)	73 (48.0%)	62	39	30.2	62.9
Southbound	4_3	51 (30.7%)	144 (50.3%)	106 (56.4%)	88	22	13.3	25.0
	3_2	39 (23.4%)	82 (47.4%)	98 (56.0%)	72	33	19.8	45.8
	2_1	51 (24.6%)	100 (45.2%)	113 (54.3%)	85	47	22.7	55.3

Table 5
Privacy performance of the 2.0 entropy dataset (Case 1).

1	2	3	4	5	6	7	8	9
Traffic direction	VTL zone pair	No. of released traces (both Z_1 and Z_2)	No. of released traces (Z_1)	No. of released traces (Z_2)	No. of inferences	No. of correct inferences	Percentage of tracked traces (P_1) (%)	Percentage of correct inferences (P_2) (%)
Northbound	1_2	30 (27.0%)	119 (100%)	59 (41.2%)	70	22	19.8	31.4
	2_3	10 (7.6%)	59 (42.8%)	71 (53.0%)	42	8	6.1	19.0
	3_4	58 (45.0%)	81 (58.7%)	126 (82.9%)	72	29	22.5	40.3
Southbound	4_3	106 (63.9%)	286 (100%)	128 (68.1%)	185	35	21.1	18.9
	3_2	26 (15.6%)	167 (96.5%)	34 (19.4%)	76	24	14.4	31.6
	2_1	10 (4.8%)	70 (31.7%)	149 (71.6%)	19	9	4.3	47.4

Table 6
Privacy performance of the 0.2 individual probability dataset (Case 1).

1	2	3	4	5	6	7	8	9
Traffic direction	VTL zone pair	No. of released traces (both Z_1 and Z_2)	No. of released traces (Z_1)	No. of released traces (Z_2)	No. of inferences	No. of correct inferences	Percentage of tracked traces (P_1) (%)	Percentage of correct inferences (P_2) (%)
Northbound	1_2	32 (28.8%)	119 (100%)	61 (43.6%)	64	15	13.5	23.4
	2_3	9 (6.9%)	61 (44.2%)	67 (50.0%)	36	6	4.6	16.7
	3_4	12 (9.3%)	77 (55.8%)	92 (60.5%)	62	6	4.7	9.7
Southbound	4_3	87 (52.4%)	286 (100%)	109 (58.0%)	160	22	13.3	13.8
	3_2	19 (11.4%)	102 (59.0%)	77 (44.0%)	70	14	8.4	20.0
	2_1	12 (5.8%)	125 (56.6%)	102 (49.0%)	75	12	5.8	16.0

proaches achieve overall higher levels of privacy, i.e., reduced P_1 and P_2. The only exceptions are for the VTL zone pair 1_2, and the VTL zone pair 4_3. Since there is no upstream VTL zone for the starting zone of the network (i.e., zone 1 for the northbound traffic and zone 4 for the southbound traffic), the entropy and individual probability based filtering algorithms tend to release most of the traces entering the network (e.g., in Tables 5 and 6, see Column 4 of the VTL zone pair 1_2). This will result in reduced privacy performance, as shown in Tables 5 and 6.

For the VTL zone pair 2_3, the privacy performances of different filtering approaches (baseline, random sampling, entropy and individual probability) with different privacy metrics are illustrated in Fig. 7. Compared with the baseline approach, all of the other approaches can provide higher levels of privacy. A stronger privacy metric (e.g., a 0.2 individual probability rather than a 0.8 individual probability) can lead to a correspondingly higher level of privacy. The results also indicate that entropy and individual probability based filtering approaches are more effective since better privacy performance can be achieved, while releasing similar numbers of traces compared with the random sampling dataset; see Figs. 9 and 10 later.

In reality, Case 1, i.e., tracking between two (close) neighboring zones hardly imposes any privacy threat. However it is still presented in detail here for two reasons. First, as shown in Section 5, Case 1 provides the basis for developing the Case 2 adversary model. Second and more importantly, if an adversary can correctly track the traces between two neighboring zones with high probability, he or she may successively track the traces for a number of zones (intersections). This is a simple adversary model based on tracking traces between neighboring zones. As shown in Table 3, since the percentage of correctly tracked traces could be as high as 86.5%, an adversary may successively track more than 20% of the entire traffic for 10 intersections, which can be the entire trip length of some urban trips. This could impose major privacy threats to urban traffic. However if filtering approaches are applied, as shown in Tables 4–6, the level of privacy can be guaranteed for two neighboring zones. As a result, the unlinkability of traffic can certainly be guaranteed at least for the above simple adversary model based on tracking neighboring VTL zones. In more realistic scenarios, if the actual penetration of mobile data is considered (which should be much less than 100%) or if the traffic density is low, or if smarter and more effective adversary models are applied, the privacy threat can be much more significant and the importance of the filtering algorithms will be more evident.

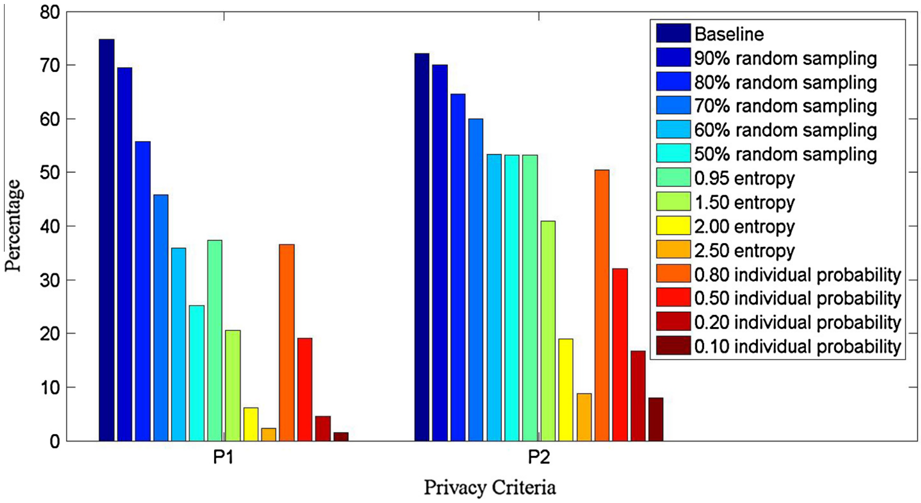

Fig. 7. Privacy performance for VTL zone pair 2_3.

7.2. Privacy performance evaluation for Case 2

In Case 2, privacy attacks are made to track the traces between two non-neighboring VTL zones. The results are shown in Table 7 for the baseline approach. The results indicate that, it is not trivial to track vehicle traces between non-neighboring VTL zones. Table 7 is based on the 100% penetration dataset, with medium traffic volume. In practice, however, the penetration of mobile sensing data is usually much lower and the traffic volume could be lower as well. This may lead to a mobile

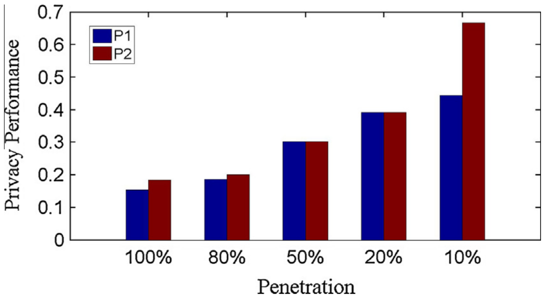

Fig. 8. Privacy performance vs. number of users (Case 2, VTL zone pair 1_2_3).

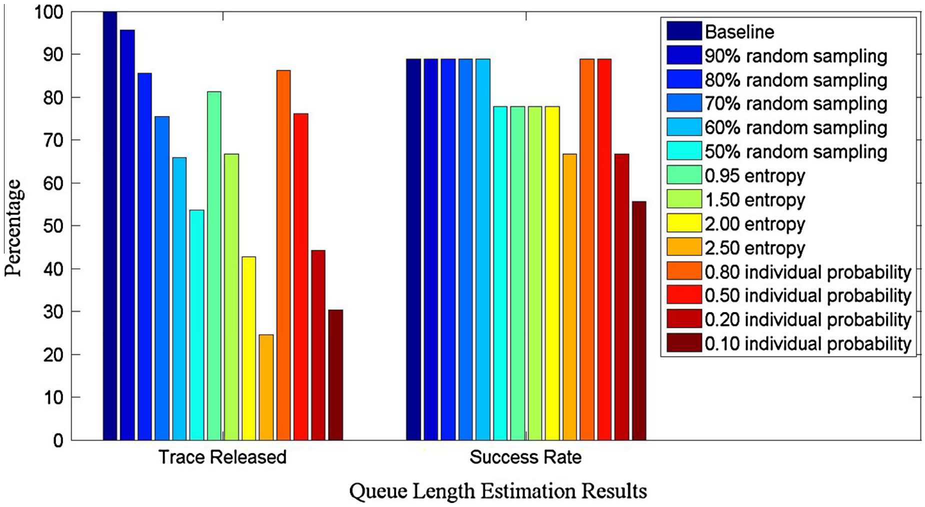

Fig. 9. Queue length estimation results (100% penetration, intersection 2, northbound).

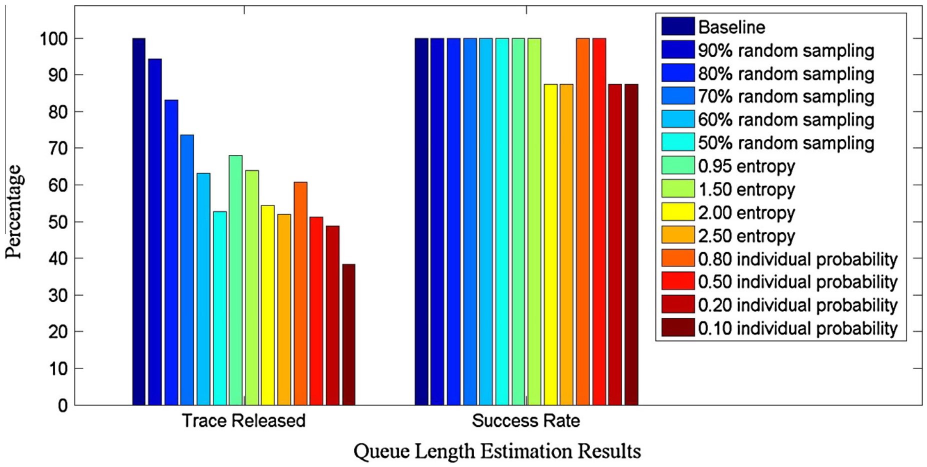

Fig. 10. Queue length estimation results (100% penetration, intersection 3, northbound).

Table 7
Privacy performance of the baseline approach (Case 2, 100% penetration).

1 Traffic direction	2 VTL zone pair	3 No. of released traces (both Z_1 and Z_3)	4 No. of released traces (Z_1)	5 No. of released traces (Z_3)	6 No. of inferences	7 No. of correct inferences	8 Percentage of tracked traces (P_1) (%)	9 Percentage of correct inferences (P_2) (%)
Northbound	1_2_3	104	119	134	87	16	15.4	18.4
	2_3_4	118	138	152	86	12	10.2	14.0
Southbound	4_3_2	151	286	175	263	31	20.5	11.8
	3_2_1	158	173	208	149	18	11.4	12.1

Table 8
Privacy performance of the baseline approach (Case 2, 20% penetration).

1 Traffic direction	2 VTL zone pair	3 No. of released traces (both Z_1 and Z_3)	4 No. of released traces (Z_1)	5 No. of released traces (Z_3)	6 No. of inferences	7 No. of correct inferences	8 Percentage of tracked traces (P_1) (%)	9 Percentage of correct inferences (P_2) (%)
Northbound	1_2_3	28	30	25	23	9	32.1	39.1
	2_3_4	32	33	28	22	7	21.9	31.8
Southbound	4_3_2	25	52	27	40	16	64.0	40.0
	3_2_1	22	26	30	24	9	40.9	37.5

sensing dataset with much smaller number of users, which is more easily to be tracked. Table 8 shows the privacy performance of a much sparser dataset, i.e., the number of users in the system is only about 20% of that in Table 7. Similar results are provided in Fig. 8 for different penetration rates. The results show that, when the number of users in the system gets smaller or the actual penetration rate of mobile data is considered, the level of privacy gets worse. In other words, it is much

Table 9
Privacy performance of the filtering approaches (Case 2, 50% penetration, VTL zone pair 1_2_3).

1 Traffic direction	2 No. of released traces (both Z_1 and Z_2)	3 No. of released traces (Z_1)	4 No. of released traces (Z_2)	5 No. of inferences	6 No. of correct inferences	7 Percentage of tracked traces (P_1) (%)	8 Percentage of correct inferences (P_2) (%)
Baseline	43	52	57	43	13	30.2	30.2
70% Random sampling	23	36	39	29	9	20.9	31.0
0.95 entropy	23	52	32	36	5	11.6	13.9
0.5 Individual probability	23	52	26	38	6	14.0	15.8

easier to make privacy attacks to a system with a small number of users (e.g., due to the fact that not all vehicles are equipped with mobile sensors, i.e., the penetration rate is less than 100%) or lighter traffic. In this context, the level of privacy provided by the baseline approach may not be strong enough; it will be necessary to apply the filtering approaches to ensure higher levels of privacy.

To further illustrate this, we apply the different filtering approaches to the dataset with 50% penetration rate. The results are depicted in Table 9. Compared with the baseline approach (i.e., P_1 and P_2 are both 30.2%), the filtering approaches can provide much stronger privacy guarantees. Compared with the random filtering approach, the individual probability and entropy based approaches, while releasing similar numbers of traces or achieving similar modeling performances (see Figs. 11 and 12 later), can provide higher levels of privacy. As shown in the table, the individual probability and entropy based approaches release the same number of traces passing both zones (Column 2), while the numbers of corrected inferences are much lower (Column 6). This clearly verifies that compared with the random filtering approach (which is purely random without considering whether a vehicle is likely to be tracked or not), the individual probability and the entropy based methods can filter out traces that are more likely to be attacked and to retain traces that are less likely to be tracked. This ensures better privacy protection as shown in Column 7 and Column 8 of Table 9.

7.3. Performance evaluation in terms of data needs for modeling

The VTL zone method must also be evaluated with respect to whether the released dataset is sufficient for traffic modeling. As presented in Section 6.2, one particular application is used to illustrate the concepts: the real time queue length estimation (Ban et al., 2011). In order to perform cycle-by-cycle queue length estimation, the method requires at least two sample travel times at a signalized intersection. In general, a larger number of released traces will lead to a higher success rate of the queue length estimation algorithm.

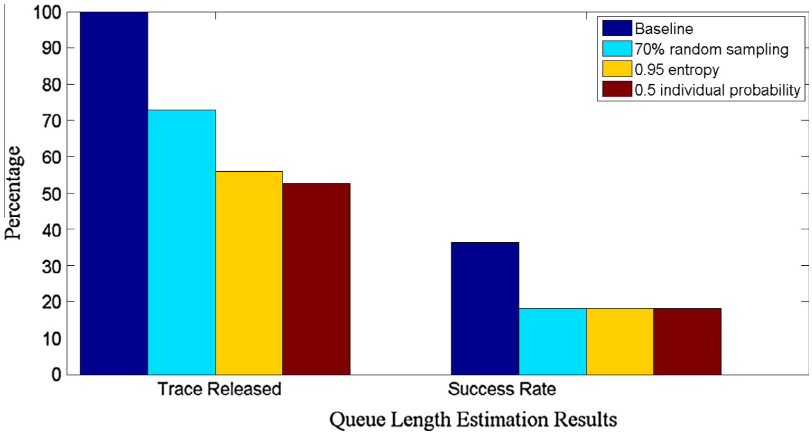

Fig. 11. Queue length estimation results (50% penetration, intersection 2, northbound).

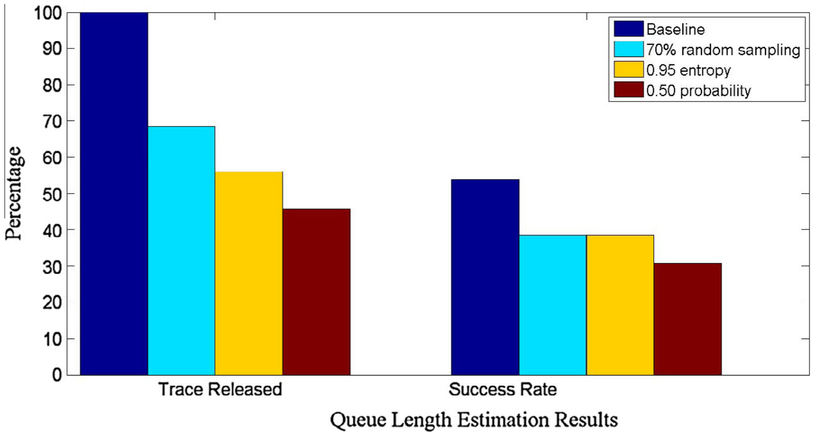

Fig. 12. Queue length estimation results (50% penetration, intersection 3, northbound).

For the 100% penetration case, Figs. 9 and 10 illustrate the queue length estimation results for different filtering approaches, with different privacy metrics. By comparing the 50% random sampling case with the 2.0 entropy and 0.2 individual probability cases, one can tell that they have a similar number of traces released at the VTL zones, as well as similar success rates for queue length estimation. However, the 2.0 entropy and the 0.2 individual probability case can provide much stronger levels of privacy, as shown in Fig. 7. In general, with a reasonable level of privacy, some of the approaches (e.g., the 2.0 entropy and the 0.2 individual probability) do not decrease the success rates much from the baseline results. This implies that by filtering out traces more intelligently (i.e., by considering the tracking probabilities of individual vehicles), the remaining traces can still be properly applied for traffic applications, while guaranteeing the privacy of the released traces.

To consider the actual penetration of mobile data (or for cases with lower number of users or lighter traffic), similar comparisons are made using the dataset with the 50% penetration rate. The results are shown in Figs. 11 and 12. Due to the significant decrement of the number of users, the success rates of queue length estimation for the baseline dataset (i.e., 36.4% for intersection 2, and 53.9% for intersection 3) are not as good as those in Figs. 9 and 10. After applying the filtering approaches, fewer traces are released and the success rates of the queue length estimation also decrease. Nonetheless, compared with the random sampling approach, the entropy and the individual probability approaches can achieve similar success rate, while the latter ones can provide much stronger privacy guarantees as shown in Table 9.

7.4. Discussions

If the proposed VTL-zone system is to be implemented in real-world applications, the choice of the proper value of the privacy metric (i.e., entropy or individual probability) used in the filtering algorithm, and the level of privacy (i.e., P_1 and P_2) should be case- and user-specific. They should be deliberately chosen for different traffic networks, different traffic states, different adversary models, different numbers of users in the system, and different preferred levels of privacy. The authors believe that an iterative process should be taken for each aforementioned scenario before the system is implemented. This process is briefly summarized here.

- *Step 1*: The level of privacy (e.g., in the form of P_1 and P_2) should be specified. Based on this, some tentative (based on experience) value of privacy metric (entropy or individual probability) should be chosen.
- *Step 2*: The privacy algorithm should be implemented based on the value of the privacy metric determined in Step 1, and is applied to some historical mobile dataset. The released traces are generated. Privacy attacks can be made to the released traces. The achieved level of privacy (e.g. P_1 and P_2) can then be evaluated experimentally.
- *Step 3*: The obtained level of privacy in Step 2 should be checked to see if it satisfies the originally specified level of privacy in Step 1. If so, the value of privacy metric is appropriate, and the privacy algorithm can be readily implemented. Otherwise, the value of the privacy metric needs to be adjusted by repeating the above process until the obtained level of privacy satisfies the desired level of privacy.

The authors will continue to study the above iterative scheme for applying the VTL-zone approach to real-world urban traffic systems, if such opportunity appears in the future.

8. Conclusions and future research directions

In this paper, the VTL zone system and related filtering approaches were proposed to protect privacy while satisfying the data needs of fine-grained urban traffic modeling. The idea is an example of the "Privacy-by-Design" approach, which incorporates privacy protection mechanisms into the system design phase. Traffic-knowledge-based adversary models were developed to evaluate the performance of the VTL zone method, especially in terms of the filtering algorithms. The algorithms were evaluated both for privacy protection and for meeting data needs for traffic modeling. It was found that tracking vehicles is usually more difficult for denser traffic. To track vehicles between two neighboring VTL zones, the idea of releasing all traces within VTL zones (i.e., the baseline approach) or random sampling help protect privacy, but not to a satisfactory level. For tracking vehicles between two non-neighboring intersections, the baseline approach might work well in some cases, necessitating no sophisticated filtering algorithms. However, in other cases (especially when the traffic is not very dense or the actual penetration of mobile data is considered), the baseline approach may not work well (see Tables 8 and 9), and sophisticated filtering algorithms do need to be applied. In both cases, filtering approaches based on individual tracking probability and entropy are more effective than pure random sampling in improving the level of privacy. Meanwhile, the released traces of these algorithms can still be applied to traffic applications with satisfactory performance.

The performances of the proposed system and related filtering approaches are based on specific adversary models. The adversary models considered in this paper are straightforward: the models in Section 5 track vehicle traces at different locations using travel time information revealed by mobile sensing data; in Section 7.1, an even simpler adversary model is also presented by tracking traces between neighboring VTL zones. They are also realistic since we assume the adversary only has access to information that is available to the public (in this particular case, short vehicle traces released at each intersection). As shown in Hao et al. (2012), cycle-by-cycle signal timing information can be estimated using intersection travel times. Ban et al. (2009) showed that cycle-by-cycle intersection delay pattern can be estimated using sample travel times and signal timing information. Therefore it is reasonable to assume that the adversary can infer the needed information from short vehicle traces released at each VTL zones. However, using these straightforward adversary models, we are still able to show that (i) privacy issues do exist if nothing is done, e.g., for the baseline case as shown in Tables 3 and 9; and (ii) the proposed VTL-zone system and the filtering approaches are effective to improve the level of privacy of mobile data (see Fig. 7 and Table 9), while at the same time to satisfy the data needs for urban traffic modeling (see Figs. 9–12).

In reality, smarter adversary models can always be developed to make more effective attacks. For example, privacy attacks that incorporate hidden information such as vehicle class and driving behavior could be more threatening (see Zan et al., 2013). If the adversary discovers that some traces have been intentionally removed (e.g., by the filtering algorithms), he or she may be able to do the pairing based on reconstructed vehicle traces (see Sun and Ban, 2013a). If the adversary has access to some prior knowledge (e.g., route choice information) of a particular driver or a group of drivers, he or she may be able to make more pertinent attacks. To deal with these more advanced adversary models, the filtering algorithms proposed in this paper are expected to be more significant, which may also need to be improved to better protect privacy.

For future work, the proposed system needs to be tested and improved using more sophisticated adversary models as discussed previously, and using more traffic modeling applications for urban environments. In terms of data, the NGSIM data have only a few intersections. It would be interesting to test the proposed system using mobile data collected from a large area, such as the traffic network or sub-network of a medium or large-size city. As discussed in Section 7.4, for such real-world applications, the iterative process of determining the value of the privacy metric, based on the selected level of privacy (P_1 or P_2), can also be tested and improved if necessary. Furthermore, as indicated earlier in Section 4.3, more empirical studies are needed to test and validate the assumption that the travel time distribution between two VTL zones in an urban environment follows the log-normal distribution.

This paper focuses on fine-grained urban traffic modeling applications. For applications that need to access full trace information of individuals, e.g., differential pricing (Zangui et al., 2013), other innovative technical methods can be essential such as network modeling methods that can incorporate users' valuation of privacy (Zangui et al., 2013). Furthermore, due to the multi-faceted nature and complexity of privacy issues (Solove, 2008), addressing practical privacy concerns for real-world urban traffic applications may need to involve both policy-oriented and technology-based approaches. Cottrill (2009) suggested this, after pointing out the limitations of both approaches and then recognizing "[the] importance of incorporating both technical and policy approaches to privacy preservation." However, Cottrill (2009) gave no further detail about how to achieve this. Therefore, another interesting future research topic would be to design a comprehensive privacy protection framework for various fine-grained urban traffic/transportation applications using mobile sensors. The framework can help select the most appropriate privacy methods given a particular application. Such a framework will most likely encapsulate both policy and technology-based privacy methods, and will need to simultaneously consider privacy protection and data needs of transportation applications.

Acknowledgements

The authors would like to thank the anonymous referees for their insightful comments and helpful suggestions on earlier versions of the paper. The work of the first and the third authors was supported in part by the National Science Foundation Grant CMMI-1031452. The work of the second and fourth authors was supported in part by the National Science Foundation Grant CMMI-1031400. Any opinions, findings, and conclusions or recommendations expressed in this paper are those of the authors and do not necessarily reflect the views of the National Science Foundation.

Appendix A. Privacy metrics

Here we provide brief definitions of several privacy metrics including *k-anonymity* and *entropy*. An example is first given to show that simple anonymization may not be sufficient to guarantee privacy.

The example is shown in Table A1. Here vehicle *A* (*A* is a pseudonym) passes location *1* at 6:00am, and there are three vehicles (*B, C* and *D*, pseudonyms as well) passing location *2* at time 6:01am, 6:11am and 6:31am, respectively. Assume there is only one route between location *1* and *2* and the adversary wants to tell which vehicle passing location *2* is vehicle *A*. In the table, we describe three scenarios in which the adversary can easily succeed. In the first scenario, the adversary has travel

Table A1
Examples of privacy attacks.

Location	Vehicle ID	Time
1	A	6:00am
2	B	6:01am
	C	6:11am
	D	6:31am

(i) The average travel time from location 1 to 2 is 10 min.
(ii) Vehicle A is a truck, the only truck that passes location 2 is vehicle C.
(iii) Consumption record of vehicle A near location 2, around 6:11 am.

time information (can be obtained from historical data); in the second scenario, the adversary has vehicle class information (vehicle classification using vehicle traces is possible, see Sun and Ban (2013b)); in the third scenario, the adversary has access to the consumption record. In all the three scenarios the adversary can easily figure out that vehicle *C* is indeed the same one as vehicle *A*, which can be considered as a privacy violation.

How to quantify privacy is an important question because it is closely related to how the privacy algorithms are formulated and how different algorithms are evaluated and compared. Apparently there have been some debates (Machanavajjhala et al., 2006; Shokri et al., 2011) regarding which privacy metric is the most appropriate. Unfortunately, to date there is no single privacy metric that can fit all scenarios or applications. Here we briefly describe some of the metrics that are relevant to our concerned applications in this paper.

K-anonymity (Sweeney, 2002) is a well-known privacy metric. One individual is said to be k-anonymous if it cannot be distinguished from other $k-1$ individuals. In this way an adversary cannot claim he or she can track an individual with more than $1/k$ confidence. As an upgraded version of k-anonymity – entropy – a concept borrowed from information theory, is also frequently used as a privacy metric. Entropy measures a system-wide tracking uncertainty, i.e., a privacy metric for a group of users, see Eq. (a1). Here p_i describes the tracking probability for an individual i, which is the probability that individual i can be tracked at other locations.

$$H = -\sum_{i=1}^{n} p_i \log_2 p_i \tag{a1}$$

K-anonymity and entropy are both measures of the privacy among a group of users. However, the k-anonymity treats all the individuals the same, meaning they share the same probability to be tracked at other locations. This appears to be unrealistic for vehicular networks since the probabilities of being tracked may vary significantly among different vehicles, due to e.g., extra information about certain vehicles. For example, they may be large trucks which are only a small portion of the entire traffic flow and thus are easily tracked. As an improvement, entropy provides the capability to incorporate each individual's tracking probability and is thus more suitable for urban fine-grained traffic modeling.

References

Agrawal, R., Srikant R., 2000. Privacy preserving data mining. In: Proceedings of the 2000 ACM SIGMOD International Conference on Management of Data, pp. 439–450.

Anderson, R., 2008. Security Engineering: A Guide to Building Dependable Distributed Systems, second ed. Wiley Publication.

Ardagna, C.A., Cremonini, M., Damiani, De Capitani di Vimercati, E., S., Samarati, P., 2007. Location privacy protection through obfuscation-based techniques. In: Proceedings of the 21st Annual IFIP WG 11.3 Working Conference on Data and Applications, Security, pp. 47–60.

Ban, X., Gruteser, M., 2010. Mobile sensors as traffic probes: addressing transportation modeling and privacy protection in an integrated framework. In: Proceedings of the 7th International Conference on Traffic and Transportation Studies, Kunming, China. <http://www.rpi.edu/~banx/publications/Ban_Privacy_ICTTS2010.pdf> (accessed 30.05.13).

Ban, X., Gruteser, M., 2012. Towards fine-grained urban traffic knowledge extraction using mobile sensing. In: Proceedings of the ACM SIGKDD International Workshop on Urban Computing (UrbComp 2012), Beijing, China.

Ban, X., Herring, R., Hao, P., Bayen, A., 2009. Delay pattern estimation for signalized intersections using sampled travel times. Transportation Research Record 2130, 109–119.

Ban, X., Hao, P., Sun, Z., 2011. Real time queue length estimation for signalized intersections using travel times from mobile sensors. Transportation Research Part C 19 (6), 1133–1156.

Beckman, R.J., Baggerly, K.A., KcKay, M.D., 1996. Creating synthetic baseline populations. Transportation Research Part A 30 (6), 415–429.

Beresford, A., Stajano, F., 2004. Mix zones: user privacy in location-aware services. In: Proceedings of the Second IEEE Annual Conference on Pervasive Computing and Communication Workshops.

Briggs, V., Walton, C.M., 2000. The Implications of Privacy Issues for Intelligent Transportation Systems (ITS) Data. Southwest Regional University Research Center.

Buttyan, L., Holczer, T., Vajda, I., 2007. On the effectiveness of changing pseudonyms to provide location privacy in Vanets. In: Proceedings of the Workshop on Security and Privacy in Ad hoc and Sensor Networks.

Cambridge Systematics, 2007. Summary Report: NGSIM Peachtree Street (Atlanta) Data Analysis (4:00 p.m. to 4:15 p.m.). <http://ngsim-community.org/> (accessed 30.05.13).

Carianha, A.M., Barreto, L.P., Lima, G., 2011. Improving location privacy in mix-zones for VANETs. In: Proceedings of the 30th International Performance Computing and Communications Conference.

Cavoukian, A., 2009. Privacy by Design, Take the Challenge. Information and Privacy Commissioner of Ontario, Toronto, Ont..

Cios, K.J., Pedrycz, W., Swiniarski, R., 1998. Data Mining Methods for Knowledge Discovery. Kluwer Academic Publishers.

Claburn, T., 2009. Google Fights Street View Ban in Switzerland. InformationWeek. <http://www.informationweek.com/internet/google/google-fights-street-view-ban-in-switzer/219401229> (accessed 30.05.13).

Clarke, R., 2001. Person location and person tracking, technologies, risks and policy implications. Information Technology & People 14 (2), 206–231.

Cottrill, C.D., 2009. An overview of approaches to privacy preservation in intelligent transportation systems and the vehicle infrastructure integration initiative. In: Proceedings of the 88th Annual Meeting of, Transportation Research Board (DVD).

Dahl, M., Delaune, P., Steel, G., 2010. Formal analysis of privacy for vehicular mix-zones. In: Proceedings of the 15th European Conference on Research in Computer Security.

Demers, A., List, G.F., Wallace, W.A., Lee, E.E., Wojtowicz, J.M., 2006. Probes as path seekers: a new paradigm. Transportation Research Record 1944, 107–114.

Douma, F., Frooman, S., Deckenbach, J., 2008. What you need to know – and not know: current and emerging privacy Law for ITS. In: Proceedings of the 87th Annual Meeting of Transportation Research Board (DVD).

Duckham, M., Kulik, L., 2006. Location privacy and location-aware computing. In: Drummond, J., et al. (Eds.), 2007. Dynamic & Mobile GIS: Investigating Change in Space and Time. CRC Press, Boca Raton, FL USA, pp. 34–51.

Freudiger, J., Raya, M., Feleghhazi, M., Papadimitratos, P., Hubaux, J., P., 2007. Mix zones for location privacy in vehicular networks. In: The First International Workshop on Wireless Networking for Intelligent Transportation Systems (WiN-ITS 2007), in Conjunction with QShine 2007, Vancouver, British Columbia.

Garfinkel, S., 1996. Why Driver Privacy Must be a Part of ITS. Converging Infrastructures: Intelligent Transportation and the National Information Infrastructure. MIT Press.

Gedik, B., Liu, L., 2005. Location privacy in mobile systems: a personalized anonymization model. In: Proceedings of the 25th IEEE International Conference on Distributed, Computing Systems, pp. 620–629.

Gonzalez, M., Hidalgo, C., Barabasi, A., 2008. Understanding individual human mobility patterns. Nature 453, 779–782.

Gruteser, M., Grunwald, D., 2003. Anonymous usage of location-based services through spatial and temporal cloaking. In: Proceedings of the First International Conference on Mobile Systems, Applications, and Services.

Hao, P., Ban, X., Bennett, K.P., Ji, Q., Sun, Z., 2012. Signal timing estimation using sample intersection travel times. IEEE Transactions on Intelligent Transportation Systems 13 (2), 792–804.

He, R., Liu, H.X., Kornhauser, A.L., Ran, B., 2002. Study travel time variability from probe vehicle data. In: Proceedings of the Seventh International Conference On: Applications of Advanced Technology in Transportation, Cambridge, MA, United States, pp. 16–23.

Herrera, J.C., Work, D.B., Herring, R., Ban, X., Bayen, A., 2010. Evaluation of traffic data obtained via GPS-enabled mobile phones: the Mobile Century field experiment. Transportation Research Part C 18 (4), 568–583.

Hofleitner, A., Herring, R., Bayen, A., 2012. Arterial travel time forecast with streaming data: a hybrid approach of flow modeling and machine learning. Transportation Research Part B 46 (9), 1097–1122.

Hoh, B., Gruteser, M., 2005. Protecting location privacy through path confusion. In: Proceedings of IEEE/Create-Net SecureComm, Athens, Greece.

Hoh, B., Gruteser, M., Xiong, H., Alrabady, A., 2007. Preserving privacy in GPS traces via uncertainty-aware path cloaking. In: Proceedings of the 14th ACM Conference on Computer and Communications, Security, pp. 161–171.

Hoh, B., Gruteser, M., Herring, R., Ban, X., Work, D., Herrera, J.C., Bayen, A.M., Annavaram, M., Jacobson, Q., 2008. Virtual trip lines for distributed privacy-preserving traffic monitoring. In: Proceedings of the 6th International Conference on Mobile Systems, Applications, and Services, pp. 5–28. <http://doi.acm.org/10.1145/1378600.1378604> (accessed 30.05.13).

International Organization for Standardization (ISO), 2009. Intelligent Transport Systems – System Architecture – Privacy Aspects in ITS Standards and Systems. Report #: TR 12859:2009. <http://www.iso.org/iso/catalogue_detail.htm?csnumber=52052> (accessed 30.05.13).

Intelligent Transportation Society of America (ITSA), 2001. Fair Information and Privacy Principles. <http://www.itsa.org/images/mediacenter/itsaprivacyprinciples.pdf> (accessed 30.05.13).

Jacobson, L., 2007. Vehicle Infrastructure Integration Privacy Policies Framework (Version 1.0.2). Reported of The Institutional Issues Subcommittee of the National VII Coalition.

Kargupta, H., Datta, S., Wang, Q., Sivakumar, K., 2003. Random data perturbation techniques and privacy preserving data mining. In: IEEE ICDM. IEEE Press.

Kido, H., Yanagisawa, Y., Satoh, T., 2005. An anonymous communication technique using dummies for location-based services. In: Proceedings of the 2nd IEEE International Conference on Pervasive Services (ICPS), pp. 88–97.

Kokotovich, A., Munnich, L.W., 2007. Thinking privacy with intelligent transportation systems: policies, tools, and strategies for the transportation professional. In: Proceedings of the 86th Annual Meeting of, Transportation Research Board (DVD).

Krumm, J., 2009. A survey of computational location privacy. Personal and Ubiquitous Computing 13 (6), 391–399.

Kwong, K., Kavaler, R., Rajagopal, R., Varaiya, R., 2009. A practical scheme for arterial travel time estimation based on vehicle re-identification using wireless sensors. Transportation Research Part C 17 (6), 586–606.

Li. M., Sampigethaya, K., Huang, L., Poovendran, R., 2006. Swing & swap: user-centric approaches towards maximizing location privacy. In: Proceedings of the 5th ACM Workshop on Privacy in Electronic Society.

Lu, H., Jensen, C., Yiu M.L., 2008. PAD: privacy-area aware, dummy-based location privacy in mobile services. In: Proceedings of the Seventh ACM International Workshop on Data Engineering for Wireless and Mobile Access, pp. 16–23.

Machanavajjhala, A., Gehrke, J., Kifer, D., 2006. ℓ-Diversity: privacy beyond k-anonymity. In: Proceedings of the 22nd IEEE International Conference on Data Engineering (ICDE), Atlanta Georgia.

Muller, K., Axhausen, K.W., 2011. Population synthesis for micro-simulation: state of the art. In: Proceedings of the 87th Annual Meeting of, Transportation Research Board (DVD).

National Highway Traffic Safety Administration (NHTSA), 2013. Connected Vehicles. <http://icsw.nhtsa.gov/safercar/ConnectedVehicles/> (accessed 30.05.13).

Nergiz, M.E., Atzori, M., Saygin, Y., Guc, B., 2009. Towards trajectory anonymization: a generalization-based approach. Transactions on Data Privacy 2 (1), 47–75.

Rass, S., Fuchs, S., Schaffer, M., 2008. How to protect privacy in floating car data systems. In: Proceedings of the Fifth ACM International Workshop on VehicularAr Inter-NETworking.

Shokri, R., Theodorakopoulos, G., Le Boudec, J., Hubaux, J., 2011. Quantifying location privacy. In: Proceedings of the 2011 IEEE Symposium on Security and Privacy, pp. 247–262.

Solove, D.J., 2008. Understanding Privacy. Harvard University Press, Cambridge, Massachusetts.

Stenneth, L., Yu, P.S., 2010. Global privacy and transportation mode anonymization in location based mobile systems with continuous queries. In: Proceedings of the 6th International Conference on Collaborative Computing: Networking, Applications and Worksharing (CollaborateCom), Chicago, IL, USA.

Sun, Z., Ban, X., 2013a. Vehicle trajectory reconstruction for signalized intersections using mobile traffic sensors. Resubmittal to Transportation Research Part C (3rd revision).

Sun, Z., Ban, X., 2013b. Vehicle classification using mobile traffic sensors. Resubmittal to Transportation Research Part C (2nd revision).

Sun, Z., Zan, B., Ban, X., Gruteser, M., Hao, P., 2011. Evaluation of privacy preserving algorithms using traffic knowledge based adversary models. In: Proceedings of the 14th International IEEE Conference on ITS, pp. 1075–1082.

Sweeney, L., 2002. K-anonymity: a model for protection privacy. International Journal on Uncertainty, Fuzziness and Knowledge-Based Systems 10 (5), 557–570.

Tang, K.P., Keyani, P., Fogarty, J., Hong, J.I., 2006. Putting people in their place: an anonymous and privacy-sensitive approach to collecting sensed data in location-based applications. In: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, pp. 93–102.

Toole, J., Ulm, M., Bauer, D., Gonzalez, M., 2012. Inferring land use from mobile phone activity. In: Proceedings of the ACM SIGKDD International Workshop on Urban Computing, Beijing, China.

Wasson, J.S., Sturdevant, J.R., Bullock, D.M., 2008. Real-time travel time estimates using media access control address matching. ITE Journal 78 (6), 20–23.

Yuan, J., Zheng, Y., Zhang, C., Xie, W., Xie, X., Sun, G., Huang, Y., 2010. T-drive: driving directions based on taxi trajectories. In: Proceedings of the 18th SIGSPATIAL International Conference on Advances in Geographic, Information Systems, pp. 99–108.

Zan, B., Hao, P., Gruteser, M., Ban, X., 2011. VTL zone-based path cloaking algorithm. In: Proceedings of the 14th International IEEE Conference on ITS, pp. 1525–1530.

Zan, B., Sun, Z., Gruteser, M., Ban, X., 2013. Linking anonymous location traces through driving characteristics. In: Proceedings of the Third ACM Conference on Data and Application Security and Privacy (Codaspy), pp. 293–300.

Zangui, M., Yin, Y., Lawphongpanich, S., Chen, S., 2013. Differentiated congestion pricing of urban transportation networks with vehicle-tracking technologies. Transportation Research Part C (in press). http://dx.doi.org/10.1016/j.trc.2013.06.011.