Achieving Guaranteed Anonymity in GPS Traces via Uncertainty-Aware Path Cloaking

Baik Hoh, Marco Gruteser, Hui Xiong, *Senior Member*, *IEEE*, and Ansaf Alrabady

Abstract—The integration of Global Positioning System (GPS) receivers and sensors into mobile devices has enabled collaborative sensing applications, which monitor the dynamics of environments through opportunistic collection of data from many users' devices. One example that motivates this paper is a probe-vehicle-based automotive traffic monitoring system, which estimates traffic congestion from GPS velocity measurements reported from many drivers. This paper considers the problem of achieving guaranteed anonymity in a locational data set that includes location traces from many users, while maintaining high data accuracy. We consider two methods to reidentify anonymous location traces, target tracking, and home identification, and observe that known privacy algorithms cannot achieve high application accuracy requirements or fail to provide privacy guarantees for drivers in low-density areas. To overcome these challenges, we derive a novel time-to-confusion criterion to characterize privacy in a locational data set and propose a disclosure control algorithm (called *uncertainty-aware path cloaking algorithm*) that selectively reveals GPS samples to limit the maximum *time-to-confusion* for all vehicles. Through trace-driven simulations using real GPS traces from 312 vehicles, we demonstrate that this algorithm effectively limits tracking risks, in particular, by eliminating tracking outliers. It also achieves significant data accuracy improvements compared to known algorithms. We then present two enhancements to the algorithm. First, it also addresses the home identification risk by reducing location information revealed at the start and end of trips. Second, it also considers heading information reported by users in the tracking model. This version can thus protect users who are moving in dense areas but in a different direction from the majority.

Index Terms—Privacy, GPS, traffic monitoring, uncertainty, anonymity, cloaking.

1 Introduction

Collaborative sensing networks (e.g., [31], [1], [3], [9]) anonymously aggregate location-tagged sensing information from a large number of users to monitor their environment. However, sharing anonymous location-tagged sensing information from users can raise serious privacy concerns. At first glance, rendering the location traces anonymous (i.e., removing identity information) before sharing them with application service providers or third parties appears to be a suitable solution. However, a time-series trace of anonymous location data exhibits spatiotemporal correlation between successive updates, potentially allowing an adversary to follow anonymous location updates and eventually reidentifying the users. Reidentification is possible, for example, because Global Satellite Navigation System (GNSS) location readings are often precise enough to identify a driver's home or workplace. This allows linking homeowner, telephone, or employee records with the anonymous traces.[1]

Motivated by an increasing number of data breaches and the potential for these reidentification attacks at network/application service providers, we consider the challenge of designing disclosure control algorithms for location traces. Prior privacy techniques for location data such as spatial cloaking techniques based on k-anonymity [20], [19] and best-effort algorithms (e.g., [39], [11], [22]) do not simultaneously meet both the privacy and the data accuracy requirements for collaborative sensing applications. This raises the problem of guaranteeing anonymity in a data set of location traces while maintaining high data accuracy and integrity.

To overcome these challenges, we develop a novel privacy metric, called *Time-To-Confusion*, to characterize the privacy implication of anonymous location traces and propose an uncertainty-aware path cloaking algorithm that guarantees a maximum time-to-confusion and provides high data accuracy. *Time-To-Confusion* effectively captures how long an adversary can follow an anonymous user at a specified level of confidence and depends on parameters such as sampling frequency and user density. The uncertainty-aware path cloaking algorithm then determines which location samples from a set of users can be revealed anonymously given a maximum allowable time-to-confusion parameter. As in other location privacy solutions [20], [22], this algorithm relies on a trustworthy privacy server, which receives location updates from all participating users and controls disclosure of samples to third-party applications. It can operate online—it does not require all traces to be complete. The algorithm also protects against home identification risks, and can handle location traces that also report directional (heading) information with each location update by taking the direction information into consideration when computing tracking uncertainty. We evaluate our proposed solution using an automotive traffic monitoring case study. The evaluation uses real GPS traces from 312 vehicles collected mostly over suburban areas of a large city in the United States.

[1] Basic anonymization is known to not fully protect against reidentification in many data sets. Examples are census database [36], search engine logs [10], and movie rating [35].

- *B. Hoh is with Nokia Research Center, 955 Page Mill Road, Palo Alto, CA 94304-1003. E-mail: baik.hoh@nokia.com.*
- *M. Gruteser is with WINLAB, Electrical and Computer Engineering Department, Rutgers University, Technology Centre of New Jersey, 671 Route 1 South, North Brunswick, NJ 08902-3390. E-mail: gruteser@winlab.rutgers.edu.*
- *H. Xiong is with the Management Science and Information Systems Department, Rutgers University, 1 Washington Park, 1 Washington Street, Newark, NJ 07102. E-mail: hxiong@rutgers.edu.*
- *A. Alrabady is with General Motors, 20058 Edgewood, Livonia, MI 48152. E-mail: ansaf.alrabady@gm.com.*

Manuscript received 1 June 2009; revised 12 Nov. 2009; accepted 3 Dec. 2009; published online 19 Mar. 2010.
For information on obtaining reprints of this article, please send e-mail to: tmc@computer.org, and reference IEEECS Log Number TMC-2009-06-0201.
Digital Object Identifier no. 10.1109/TMC.2010.62.

Contributions. In summary, this paper offers the following specific contributions:

- Formal definition of a novel time-to-confusion metric to evaluate privacy in a set of location traces. This metric describes how long an individual user can be tracked.
- Development of an uncertainty-aware path cloaking algorithm that can guarantee a specified maximum time-to-confusion and protect against home identification risks.
- Demonstration through experiments on real-world GPS traces that this algorithm limits maximum time-to-confusion while providing more accurate location data than a random sampling baseline algorithm. In particular, it offers guaranteed protection for users that move into low-density areas.

This work extends our earlier paper [25] by providing a formal definition for the time-to-confusion concept, by including heading information in the tracking uncertainty model, by addressing the additional home identification risk model, and by evaluating these extensions using the real-world GPS traces. It also includes an expanded discussion of privacy risks in anonymous GPS traces.

Overview. The remainder of the paper is structured as follows: Section 2 briefly introduces collaborative sensing applications and our specific case study, the automotive traffic monitoring system. Section 3 presents possible inference attacks on anonymous location databases and evaluates known privacy algorithms. In Section 4, we describe the threat model and introduce the time-to-confusion metric that captures the time an adversary can track with high confidence. After enumerating several existing algorithms in Section 5, we present the uncertainty-aware privacy algorithm in Section 6. Sections 7 and 8 present the experimental results obtained with real-world location traces, which demonstrate the privacy and data accuracy advantages. We then discuss limitations, extensions, and future directions in Section 9. Section 10 reviews related work, and finally, Section 11 concludes the paper.

2 Collaborative Sensing Applications

Collaborative sensing applications rely on the availability of periodic location updates provided by ever more cost-effective GNSS chips. The applications that actively use GNSS location traces span the intelligent transportation domain (e.g., [42], [26], [24]), pollution monitoring (e.g., [7], [9]), pedestrian flow monitoring for marketing [4], and urban planning [3]. In this paper, we select automotive traffic monitoring as a case study.

2.1 Traffic Monitoring with Probe Vehicles

Automotive traffic monitoring application aims to provide estimates of current travel time for routes using real-time traffic-flow information. Traffic-flow information is derived from probe vehicle speed readings on road segments.

The probe vehicles use on-board GPS receivers [42] (or GPS-enabled mobile devices) and cellular communications (or WiFi [26]) to periodically report records with the following parameters to traffic information systems: latitude, longitude, time, and speed (with heading). A central traffic monitoring system stores them in a database for real-time and historical traffic analysis. From this information, the system can estimate current mean vehicle speed, which can be used to build a real-time congestion map (e.g., a congestion index). Estimated traffic information can then be broadcasted to subscribers or made available through a Web interface, where drivers can access it through their navigation systems or from home or office computers. Fig. 1 illustrates this architecture.

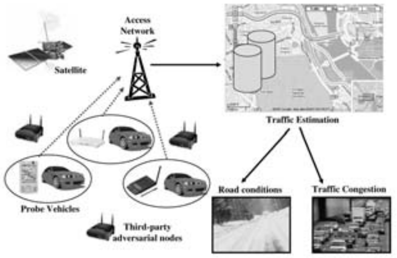


Fig. 1. Traffic monitoring architecture comprising three entities: probe vehicles, communication service provider, and traffic monitoring service provider. Traffic monitoring builds a real-time congestion map from vehicle speed and position reports.

In this approach, probe vehicles replace infrastructure sensors such as loop detectors and cameras, thereby reducing the installation/maintenance cost. Moreover, it can more cost-effectively achieve wide road coverage since the system can use existing devices such as navigation systems or even GPS smart phones carried by drivers to collect the speed and position measurements [23], [13]. Using such aftermarket or handheld devices, the necessary penetration rate (fraction of vehicles equipped with sensing devices) for reliable traffic status estimation can be achieved relatively quickly. This rate is estimated at 5 percent [17].

2.2 Real-World GPS Trace Collection

We have offline collected a data set containing GPS traces from 312 volunteer drivers driving in a large US city and its suburban area for a week. The collected traces, which are similar to a data set of real deployments (e.g., [5], [6]), covered the 70 km × 70 km region as depicted in Fig. 2a. To protect drivers' privacy, no specific information about the vehicles or drivers is known to the authors. Each GPS sample comprises vehicle ID, time stamp, longitude, latitude, velocity, and heading information. Samples are recorded every minute, while the vehicle's ignition is on. The collected traces contain temporal gaps in the following cases: when the vehicle is parked with its ignition switched off, when the GPS reception is lost (e.g., due to obstruction from high-rise buildings), or when the receiver is still in the process of acquiring the satellite fix. Because the traces do not contain information about ignition and GPS receiver status, we assume that a gap longer than 10 min indicates that the vehicle was parked. Fig. 2b illustrates the distribution of gaps in the traces of around 312 vehicles. Each dot represents a received data sample. We refer to the parts of a trace between two gaps longer than 10 min as a *trip*.

2.3 Data Quality Metrics and Requirements

There exists a trade-off between data quality (or its utility) and the degree of privacy in data privacy algorithms because each algorithm introduces unavoidable data modifications such as omission, perturbation, or generalization of a datum to increase privacy. To evaluate privacy algorithms meaningfully, we first discuss data quality requirements and metrics for the traffic monitoring application.

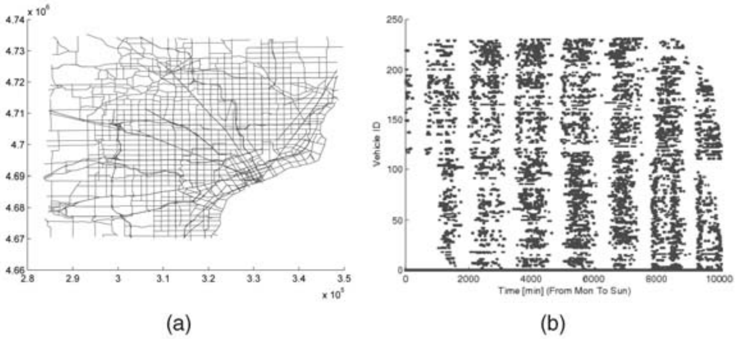


Fig. 2. Spatiotemporal distribution of location reports in real-world data set. (a) $70\ \mathrm{km} \times 70\ \mathrm{km}$ road network with cell weights indicating the busiest areas. (b) Temporal distribution of GPS traces for 312 vehicles.

The application represents a road map as a graph comprising a set of road segments, where each road segment describes a stretch of road between two intersections. Generating the congestion map then proceeds in three steps: Mapping new GPS samples to road segments, computing mean road segment speed, and inferring a congestion index (e.g., by comparing current mean speed on a road segment of interest with its free-flow speed).

Mapping GPS samples onto road segments requires high *spatial precision and accuracy*. Consider that two different parallel road segments (with traffic flow in same direction) may be only about 10 m apart, as on the New Jersey Turnpike, for example. Cayford and Johnson [13] showed, however, that using tracking algorithms, the correct road can be determined in 98.4 percent of all surface streets and 98.9 percent of all freeways if the location system provides a spatial accuracy of 100 m and updates in 1 s intervals. When reducing the update interval from 1 to 45 s, the correctly determined roads drop from 99.5 to 98 percent (at 50 m spatial accuracy). Therefore, to maintain high road mapping accuracy at the 1 min sample interval for our data traces, we can assume that a minimum spatial accuracy of 100 m is needed.

Another important data quality requirement is *road coverage*, which describes the fraction of road segments from which speeds updates were received in a given time interval. It primarily depends on the distribution of vehicles and the penetration rate, the percentage of vehicles carrying the traffic monitoring equipment. To achieve high coverage, these systems aim at a minimum penetration rate of 3 (for freeways) to 5 percent (for surface streets) [17], but during the initial deployment phase, penetration rates may be much lower. Thus, privacy algorithms must offer protection even under low deployment densities.

Since road coverage is often reduced by the privacy algorithms, for example, if the algorithm omits all location samples from a specific road segment, we select road coverage as a key evaluation metric. In particular, we measure *relative weighted coverage metric*, which is based on the following rationale: First, it is weighted by traffic volume. While probe-vehicle-based traffic monitoring aims to extend traffic monitoring beyond a few key routes, information from busier roadways is certainly more important than from low traffic routes. Second, it is measured relative to the coverage from the original data set before privacy algorithms are applied because road coverage is fundamentally limited by the number of probe vehicles on roads.

To measure the effect of removed samples on road coverage, relative weighted coverage first assigns each location sample a weight, depending on how busy the area around this sample is. Then, it divides the sum of weighted location samples from modified (after privacy algorithm processing) traces by the sum of weighted location samples from the original traces. To estimate these weights for our data set, we divide the area into $1\ \mathrm{km} \times 1\ \mathrm{km}$ grid cells and count the number of location samples n_i emanating from each cell i over one day in the original traces. The resulting weights for each cell are overlaid on the road map in Fig. 2a. The weights are normalized with the sum of weights over all samples so that the relative weighted road coverage for the original data set is equal to 1. More precisely, the weight for all samples in cell i equals

$$w_i = \frac{n_i}{\sum_j n_j^2}.$$

With these weights, relative weighted road coverage for a set of location samples L is then defined as $\sum_{l \in L} w_{c(l)}$, where the function c returns the cell index in which the specified location sample lies.

In summary, we can measure data quality for a traffic monitoring application through the relative weighted road coverage, where we consider a road segment covered if a data sample with sub-100 m accuracy is available. Table 1 summarizes the key system parameters and requirements that we will assume in the following sections.

3 Adversary Model

Monitoring a vehicle's movements can reveal driver's sensitive information particularly in the United States, where peoples' lifestyle heavily relies on automobiles. Knowing either the origin or the destination of a trip can reveal information about a driver's health, lifestyle, or political associations if it is associated with driver identity. While privacy is only compromised if both sensitive information

TABLE 1
Traffic Monitoring System Data Requirements

Parameter	Requirement
Spatial Accuracy	100m
Sample Interval	1min
Delay	few minutes

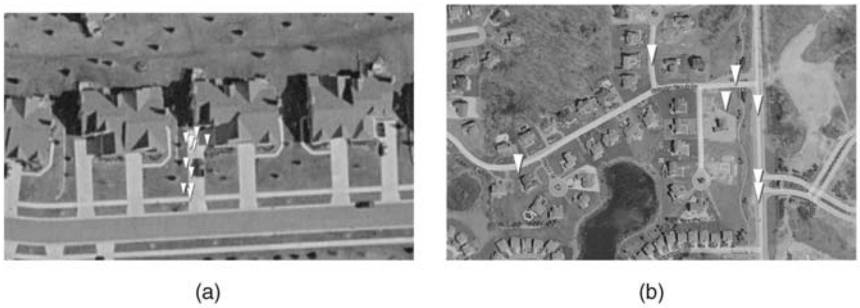


Fig. 3. Place identification example. Determining which building a driver visited is possible in left scenario because trip endpoints (shown by the markers) cluster denser than nearby homes. (a) Well-clustered destinations (b) Noisy originations due to GPS signal acquisition delay.

TABLE 2
Adaptive k-Means Clustering for Home Identification

1. Drop location samples with too high speed ($> 1m/s$) from all vehicles (i.e., remaining samples contain the candidate trip endpoints).
2. Select a target region of interest to improve computational efficiency, and drop samples outside this region.
3. Apply pair-wise clustering algorithm to samples in target region and store the returned cluster centroids.
4. Filter the candidate home locations out of all centroids using two heuristics (A:arrival time and B:zoning information).

and identity of data subjects are known, naive anonymization is not sufficient to protect privacy. Identities can often be reconstructed through inference attacks that combine the anonymous data set with other external data sources [35].

Two such inference techniques for location traces are tracking and home identification. They together allow reconstructing likely identities of some drivers and allow to follow them to potentially sensitive locations. This privacy leakage scenario assumes that an adversary has gained access to anonymous location traces, for example, by accidental or intentional disclosure of location traces by insiders.[2] This data set of anonymous location samples has the form

$$M = \{\langle m_{1,t_1}, \ldots, m_{k_1,t_1}\rangle, \langle m_{1,t_2}, \ldots, m_{k_2,t_2}\rangle, \ldots,\\ \langle m_{1,t_n}, \ldots, m_{k_n,t_n}\rangle\},$$

where m_{i,t_j} stands for the ith location sample received during time interval t_j, $k_1, \ldots, k_n$ denotes the number of different samples (from different subjects) received during each quantized time interval $t_1, \ldots, t_n$. We assume that each data subject will provide at most one location sample per time interval and the order of the samples is randomized at each time instant (i.e., the index i conveys no information about the data subject that generated the sample, or whether it is the same subject that generated a prior sample). Also note that each sample m_{i,t_j} is a tuple containing both a location and a speed information. We distinguish two different cases: speed-only and speed-with-heading information. In the speed-only case, $m_{i,t_j} = (x_{i,t_j}, y_{i,t_j}, v_{i,t_j})$. In the speed-with-heading information, every sample contains a speed vector $m_{i,t_j} = (x_{i,t_j}, y_{i,t_j}, v^x_{i,t_j}, v^y_{i,t_j})$. Without loss of generality, we assume a two-dimensional space.

The adversary may combine the anonymous location data set M with data obtained from other sources to reidentify drivers. A likely approach is that the adversary will use a data set of identifying locations, these are locations usually only visited by a unique user identity. Therefore, the adversary can identify the originator of sample $m_{i,tj}$ that stems from such an identifying location. The best-known source of such identifying locations is home locations; an adversary could derive a data set of identities and locations by geocoding residential addresses obtained from telephone white pages, for example.

3.1 Home Identification

Home identification seeks to determine users' home positions given the anonymous location data set M. We assume that the number of participating users may be much smaller than the number of people living in the area. Thus, identifying such candidate home locations is important, since simply matching all location samples to all residential addresses would lead to too many false positives.

Recent studies [24], [32] demonstrated that clustering can be an effective tool for home identification. In particular, clustering promises to group a set of anonymous location samples (with low-to-zero speed) that likely belong to the same destination, and the centroid of each cluster provides a good estimate of the destination. Fig. 3a shows a sample scenario, where GPS samples cluster precisely on a single home's driveway. In contrast, we found that trip origins are usually harder to identify, likely because the first GPS samples are farther away from the exact destinations due to the receiver's GPS acquisition delay after power on, as shown in Fig. 3b. Details on home identification algorithm can be found in Section 3.

In this work, we use the clustering-based home identification algorithm presented in [24]. As summarized in Table 2, it first extracts low-speed location samples to concentrate on samples where drivers are likely to approach a home. It then clusters samples from multiple days of data using a maximum cluster diameter parameter, which can be estimated from the population density in the region. The clustering algorithm repeatedly combines the closest clusters until any further combining would create clusters larger than

[2] Such data breaches are not uncommon; a study [2] reports that 217,551,182 records with sensitive personal information have been involved in data breaches since 2005 in the United States.

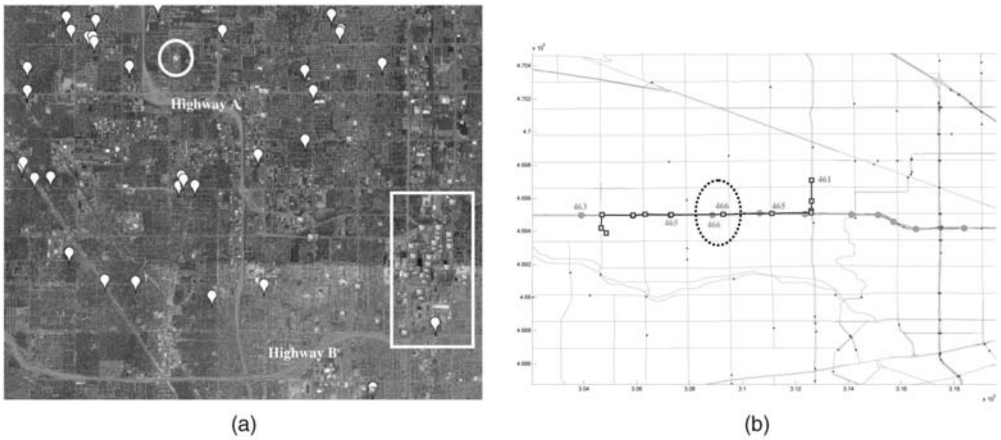


Fig. 4. Heading information helps tracking performance. White symbols denote origins of trips and white rectangles represent destinations of the corresponding trips (a). Two samples at $T = 466$ constitute a candidate set with high tracking uncertainty when an adversary tracks location samples, but one of them drops out, considering their driving directions. (a) Trips in the morning peak time of the day. (b) Likelihood computation based on direction and distance.

the specified maximum diameter. In our simulations, we use a value of 100 m for this threshold which we derive from the actual home density in the region. The centroids of these clusters are now likely to point to frequently visited destinations. Finally, it filters out clusters located in non-residential zones[3] or clusters with many daytime arrivals.

3.2 Tracking

Target tracking techniques can be used to reconstruct paths from anonymous samples or segments [21]. For example, target tracking techniques can allow an adversary to follow the traces reported by a vehicle from an identifying home to other potentially sensitive locations, thereby learning who visited these sensitive places.

These algorithms generally predict the target position using the last known speed and heading information and then decide which next sample to link to the same vehicle through Maximum Likelihood Detection. If multiple candidate samples exist, the algorithm chooses the one with the highest a posteriori probability based on a probability model of distance and time deviations from the prediction (in our evaluation, we assume a strong adversary with a good model of these deviations). Tracking starts with an arbitrary sample m_{i,t_j} in the set M. Following are the key tracking steps.

Prediction. We predict the next measurement from the same vehicle,

$$\tilde{m}_{i,t_{j+1}} = \left(x_{i,t_j} + v^x_{i,t_j} * \Delta t, y_{i,t_j} + v^y_{i,t_j} * \Delta t\right),$$

using the known speed.

Selection. We compute how likely each of the measurements collected at time instant t_{j+1} is the next measurement of vehicle for which we predicted movements: $P(m_{p,t_q} | \tilde{m}_{i,t_{j+1}})$ for an arbitrary measurement from pth user at the time instant t_q $(t_j < t_q < t_k)$. The Bayes rule and the assumption on equal priori probabilities on all hypotheses simplify the a posteriori probability into a function dependent on the distance between the predicted position $\tilde{m}_{i,t_{j+1}}$ and the actual position m_{p,t_q}. Smaller distances imply greater likelihood. Out of all k_q samples, we select the sample with index p that maximizes this probability (i.e., that is closest to the predicted position). Let $N(\tilde{m}_{i,t_{j+1}}, \tilde{M}_j)$ denote a function that returns the sample closest to a predicted position and $\tilde{M}_j$ is a subset of M that includes location samples collected only after the time instant t_j.

Update and repeat. The process uses the selected sample m_{p,t_q} as new starting points and repeats with the prediction step.

3.3 Enhanced Tracking Models

Several tracking enhancements are possible. We also consider the following models in our experiments and discuss the potential effect of other models in Section 9:

- **Reacquisition.** Under the reacquisition model, an adversary skips samples with high tracking confusion under certain conditions, and thus, may be able to reacquire the correct trace even after a point of confusion. We predict several steps ahead and use the closest match among all steps, the step which minimizes the confusion metric. We will define the confusion metric in Section 4. This change simply increases the number of hypotheses in the selection step of our previous description. Thus, given the sample m_{i,t_j}, we predict its movement for w time intervals as follows: for $\forall i \in 1, \dots w$,

$$\tilde{m}_{i,t_{j+1}} = \left(x_{i,t_j} + v^x_{i,t_j} * \Delta t * i, y_{i,t_j} + v^y_{i,t_j} * \Delta t * i\right).$$

- **Heading information.** Using the knowledge of a driving direction can also enhance tracking performance. This is particularly the case when updates are frequent enough so that direction does not change significantly between updates. Consider the scenario in Fig. 4a, where a single vehicle is driving against rush hour traffic. Using the algorithm described so far, tracking might easily fail because a car on the opposite lane may actually be closer to the predicted position than the vehicle we intend to track. By using heading information in the tracking metric, however, tracking is straightforward in this case where only a single vehicle continues to drive in this direction.

[3] In our experiment, we manually eliminated centroids located outside residential areas by plotting and checking them on the satellite imagery of Google Earth. However, this process could be automated with GIS city zoning information.

At each step, the likelihood that a particular candidate sample belongs to the tracked vehicle is now dependent both on distance and heading information. We represent this through a joint probability density function as follows: For an arbitrary measurement from the pth user at the time instant t_q $(t_j < t_q < t_k)$,

$$P(m_{p,t_q}|\tilde{m}_{i,t_{j+1}}) = P_{emp,distance}(\delta) * P_{emp,heading}(\theta),$$

where δ and θ denote the distance gap and heading gap between $\tilde{m}_{i,t_{j+1}}$ and m_{p,t_q}, respectively.

This pdf models general knowledge about peoples' movement patterns. Tracking performance improves if the model closely fits movement patterns in the data set. In our study, we analyze a worst-case scenario, where the adversary has a perfect pdf model (which we empirically derived from the data itself). The empirical PDF of heading difference between two adjacent location samples is depicted in Fig. 10a.

In addition, we included a heuristic to reduce the computational complexity. It is also based on the observation that users driving at higher speeds more likely maintain their directions. Thus, we do not consider candidate location samples with a heading difference of more than 90 degrees, if the last known speed of the vehicle was high.

While we use a relatively straightforward tracking model in this study, the metrics and algorithms developed here can also be used with more sophisticated tracking models. We chose the simpler model because we did not observe significant improvements in tracking performance with other algorithms in our evaluations scenario, as we discuss in Section 9.

4 The Time-to-Confusion Privacy Metric

We first present a novel privacy metric for location traces called *time-to-confusion* and then discuss a related metric for evaluating home identification risks.

We observe that the degree of privacy risk strongly depends on how long an adversary can follow a vehicle. For a privacy breach, a trace must contain a privacy-sensitive event (e.g., visited a sensitive destination) and the adversary must be able to identify the driver generating this trace. Both the probability that sensitive information is included and the probability of identification increase with longer traces. Identification may be possible, for example, if the vehicle returns to a known home or work location of a specific individual.

We therefore characterize privacy in terms of the maximum possible tracking time. We refer to this time as *time-to-confusion* since it is usually limited by areas of confusion, where too many users are traveling, so that an adversary cannot track a user with a high degree of certainty.

Tracking uncertainty. To formally define this metric, let us start by considering the uncertainty inherent in a single tracking step. Inspired by the use of entropy in anonymous communication systems [37], we use information-theoretic metrics to measure uncertainty or confusion in tracking. We denote the uncertainty by $U(\cdot)$. For any point on the trace, *Tracking Uncertainty* is defined as $U = -\sum p_i \log p_i$, where p_i denotes the probability that location sample i belongs to the vehicle currently tracked. Lower values of U indicate more certainty or lower privacy. Given no other information than the set of location samples, intuitively, the probability for a sample reported at time t is high, if the sample lies close to the predicted position of the vehicle at time t and if no other samples at the same time are close to the vehicle. We can then also define the inverse, tracking confidence, as $C = (1 - U)$.

Empirically, we found that distances of the correct sample to the predicted position appear monotonically decreasing in Fig. 6b. Therefore, we compute the probability p_i for a given location sample by first evaluating the exponential function

$$\hat{p_i} = e^{-\frac{d_i}{\mu}}$$

for every candidate sample and then normalizing all $\hat{p_i}$ to obtain p_i. The parameter μ can be interpreted as a distance difference that can be considered very significant. We obtain the value of μ from empirical pdf of distance deviation in Fig. 6b which we fit with exponential function using unconstrained nonlinear minimization (μ is 2,094 meters).

The proposed algorithm is not dependent on the use of an exponential function for estimating the probability that a location sample belongs to the same trace. It does assume, however, that a publicly known "best" tracking model exists and that the adversary does not have any better tracking capabilities. In this paper, we have empirically derived this probability model by fitting an exponential function.

Given this uncertainty definition, we can now define linkability. We present definitions on *linkability* and *traceable path* that we use for defining a novel privacy metric for location traces, a **time-to-confusion**.

Definition 1. *An arbitrary sample m_{i,t_j} in the set M is said to be* linkable *to the sample m_{q,t_r} if the following conditions hold: 1) $t_1 < t_j < t_r$, 2) $m_{q,t_r} = N(m_{i,t_j}, \tilde{M}_j)$, where $N(\cdot)$ is a function that returns the sample closest to a predicted position and $\tilde{M}_j$ is a subset of M that includes location samples collected only after the time instant t_j, and 3) $U(m_{i,t_j}, \tilde{M}_j) \leq U_{th}$, where U_{th} is an uncertainty threshold and $U(\cdot)$ computes the entropy for this tracking step as described before.*

From the above definition, we build a function $Track(m_{i,t_j}, \tilde{M}_j)$, which returns m_{q,t_r} only if m_{i,t_j} is *linkable*.

Definition 2. *We call a* traceable path *a set $P = \{m_{i,t_j}, Track(m_{i,t_j}, \tilde{M}_j), \ldots, Track^n(m_{i,t_j}, \tilde{M}_j)\}$, applying $Track(\cdot)$ function repeatedly until its return value does not exist.*

Finally, we measure the degree of privacy as the *Mean Time to Confusion (MTTC)*, the time that an adversary could correctly follow a trace. Note that this includes time while a user remains stationary unless otherwise specified. More specifically, the time to confusion is the tracking time between two points, where the adversary reached confusion (i.e., could not determine the next sample with sufficient certainty).

Definition 3. *For a given set $P = \{m_{i,t_j}, Track(m_{i,t_j}, \tilde{M}_j), \ldots, Track^n(m_{i,t_j}, \tilde{M}_j)\}$, we compute total tracking time by time difference between the first location sample and the last location sample of P, and we call it Time-To-Confusion.*

As privacy metrics in this paper, we compute the median and maximum Time-To-Confusion over all *traceable paths*. The median over all TTCs is obtained by applying the function Track to every location sample in the set M.

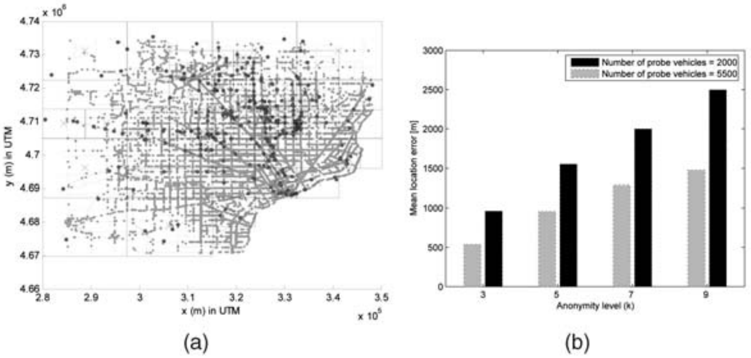


Fig. 5. Data accuracy of samples processed with spatial cloaking algorithm fails to meet the accuracy requirement in our scenario (b). (a) Snapshot of spatial cloaking applied to GPS traces. (b) Data accuracy.

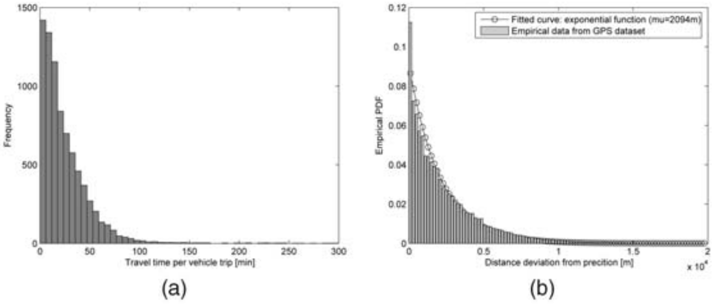


Fig. 6. Fitting distance errors in tracking using an exponential function. (a) Empirical distribution of travel times per vehicle trip. (b) Empirical probability distribution function of distance deviation from prediction to correct sample.

The MTTC can then be defined as the median tracking time during which uncertainty stays below a confusion threshold. If the uncertainty threshold is chosen high, tracking times increase but so also does the number of false positives (following incorrect traces). Since the adversary cannot easily distinguish correct tracks and false positives, we assume that high uncertainty thresholds will be used.

Uncertainty in home identification. We measure the home identification risk through a similar uncertainty metric. To calculate an uncertainty, we measure distance between a cluster centroid and each of the five nearest homes from it. For each distance, we assign a likelihood by computing a probability,

$$\hat{p}_i = e^{-\frac{d_i}{\mu}},$$

normalize all likelihoods for five corresponding candidates, and calculate the entropy. In the experiment conducted in this study, we chose to consider the five nearest homes to balance accuracy of uncertainty computation with the time needed to conduct the (partially manual) experiment.

We chose this metric because building density significantly affects the accuracy of the home identification technique. In dense areas, false positives can be caused by many vehicles waiting at traffic lights or stop signs that shift the cluster centroid to a neighbor's house, for example. In areas with few buildings, this is less likely.

5 Existing Privacy Algorithms

Several techniques have been proposed to protect against location privacy breaches through inference methods. However, we are aware of only one class of techniques, spatial cloaking algorithms for k-anonymity, which can guarantee a defined degree of anonymity for all users. Other algorithms can be categorized as best-effort algorithms that increase average privacy levels, but offer no specific guaranteed privacy level for an individual user. We briefly review these algorithms and evaluate their effect on data quality and their level of privacy protection.

5.1 Spatial Cloaking for Guaranteed Privacy

k**-anonymity** [36] formalizes the notion of strong anonymity and complementary algorithms exist to anonymize database tables. The key idea underlying these algorithms is to generalize a data record until it is indistinguishable from the records of at least $k-1$ other individuals. Specifically, for location information, spatial cloaking algorithms have been proposed [20], [19] that reduce the spatial accuracy of each location sample until it meets the k-anonymity constraint. To achieve this, the algorithms require knowledge of the nearby vehicles' positions, thus, they are usually implemented on a trusted server with access to all vehicles' current position.

k-anonymous data sets produced with known algorithms cannot meet traffic monitoring's accuracy requirements. Fig. 5b shows the spatial accuracy results obtained after applying a spatial cloaking algorithm to guarantee k-anonymity of each sample. We use the same data set in Section 7.1 so that we could directly compare k-anonymity with our proposed solution in terms of spatial accuracy. The results were obtained with the CliqueCloak algorithm [19], which, to our knowledge, achieves the best accuracy. *The results show that even for very low privacy settings, $k=3$, location error remains close to 1,000 m for an emulated deployment of 2,000 vehicles, far over the accuracy requirement of the traffic monitoring application.* While these results can be expected to improve with increased penetration rates as the deployment case of 5,500 vehicles shows 500 m for $k=3$

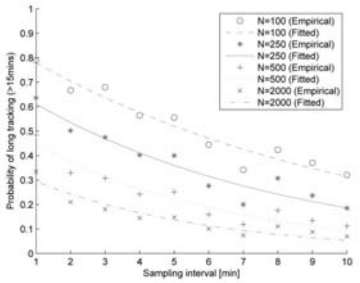


Fig. 7. Dependency of traceable path on sampling period and probe density.

(indeed, Gruteser and Grunwald [20] show that median accuracies of 125 meters and below can be obtained when *all* vehicles act as probes), other privacy approaches are necessary to enable probe systems operating with lower penetration rates.

5.2 Best-Effort Algorithms for Probabilistic Privacy

Given that in dense environments paths from many drivers cross, drivers intuitively enjoy a degree of anonymity, similar to that of a person walking through an inner city crowd. Thus, Tang et al. [39] lay out a set of privacy guidelines and suggest that the sampling frequency, with which probes send position updates, should be limited to larger intervals. The authors mention that a sample interval of 10 min appears suitable to maintain privacy, although the choice appears somewhat arbitrary (for reference, a typical consumer GPS chipset implementation offers a maximum sampling frequency of 1 Hz). We refer to data collection with reduced sampling frequency as subsampling.

Other best-effort algorithms suppress information only in certain high-density areas rather than uniformly over the traces as the subsampling approach. The motivation for these algorithms is that path suppression in high-density areas increases the chance for confusing or mixing several different traces. This approach was first proposed by Beresford and Stajano [11]. The path confusion [22] algorithm also concentrates on such high-density areas although it perturbs location samples rather than suppressing them. These techniques increase the chance of confusion in high-density areas, but they also cannot guarantee strong privacy in low-density areas where paths only infrequently meet. Thus, in terms of worst-case privacy guarantees, their advantage over subsampling remains unclear.

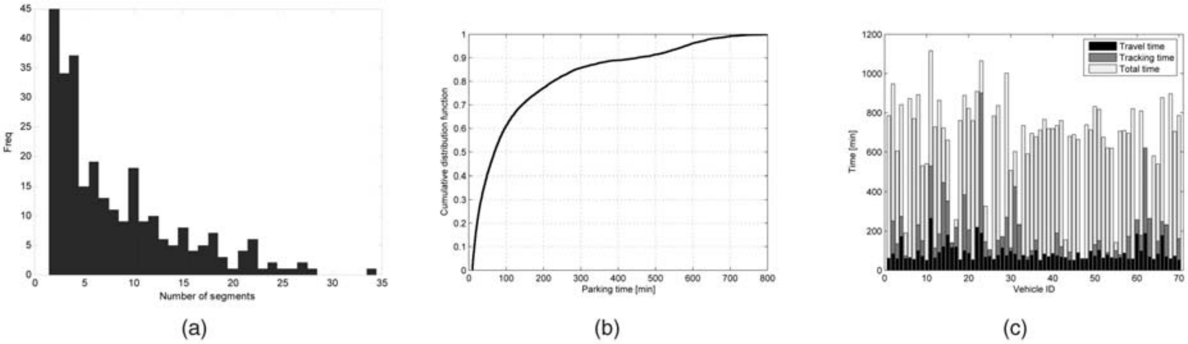
5.3 Privacy of Best-Effort Subsampling

We conjecture that the performance of inference attacks depends on the sampling interval and the user density. Here, we investigate this parameter space through tracking and home identification experiments on real week-long GPS traces from 312 probe vehicles.

Fig. 7 illustrates the percentage of paths that can be tracked longer than 15 mins, the median trip time in the US. This means that for about half these tracked trips, the adversary can observe both origin and destination of the trip. The graph shows the percentage of tracked paths dependent on user densities and sampling intervals. This datum is empirically derived using the tracking model described in Section 3 from 24 hours of traces in the suburban area (see Fig. 2a). As evident, the tracking time appears to follow an exponential function as either the sampling interval or the probe density increases. We observe that even LBSs with a seemingly large 10-minute sampling interval allow long tracking for 7 percent of users in our high user density scenario (2,000 probe vehicles in our road network).

In this subsequent case study, we characterize how easily an adversary can link multiple trips to the same individual, instead of just tracking a single trip. Linking multiple trips is challenging because vehicles will not update their location while turned off, thus, the tracking model has to evaluate at every point in time whether nearby location updates are generated by other vehicles passing by, or whether the tracked vehicle has started a new trip. We account for this by using a different likelihood model in our tracking algorithm and by looking ahead several hours in time. We empirically obtain the distribution of time deviation between two successive trips as shown in Fig. 8b. With this CDF of time deviation and an empirically fitted PDF of distance deviation (exhibiting quite similar pattern to Fig. 6b), we can characterize the likelihood that a given location sample stems from a new trip of the same tracked vehicle.

We apply this tracking model against paths of 315 different users with a duration of 2.5 days to measure the total tracking duration. Each probe vehicle's trace consists of multiple trips (2-13 trips per day) as shown in Fig. 8a. Fig. 8c depicts the path tracking performance over 2.5 days on 70 longest tracked traces out of 315 users. We plot each traced trace in terms of total time, tracking time, and travel time. Total time denotes the whole time duration between the origin of the first trip to the destination of the last trip of a single probe vehicle, tracking time describes how long an adversary follows the vehicle including parking time, and



(a) (b) (c)

Fig. 8. Some statistics on collected real GPS traces. The rightmost figure shows some of tracking outliers stretches over multiple trips. (a) Number of trips in each vehicle's trace. (b) Parking time cumulative distribution. (c) Traceable paths.

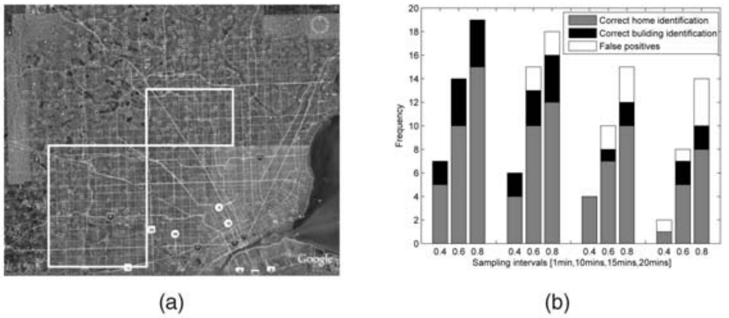


Fig. 9. Plausible home locations in two target regions (in white rectangles) according to manual inspection. The study considered a total of 65 homes in a $25\ \mathrm{km} \times 25\ \mathrm{km}$ area. (a) Four different sampling intervals are depicted by four circles and the specific parameter is next to each mark. (b) Original location traces have one location report per minute, which corresponds to 1 minute interval. (a) Plausible home locations. (b) Home identification rate.

finally, travel time only measures the driving time. We observe that many tracking outliers go beyond a single trip, even stretching up to a few trips.

Home identification risks versus sampling intervals. Let us now examine the level of protection offered by subsampling against the home identification technique. We again use the same location traces and consider subsampling intervals of 1, 10, 15, and 20 min. We measure the home identification rate, meaning how many homes out of the total (65 home locations shown in Fig. 9a) are correctly detected, and the false positives rate, meaning how many are incorrect among the estimated home locations. The home locations were obtained by manually inspecting all traces in the depicted target region.

Again, Fig. 9b demonstrates that longer sampling intervals do not necessarily address the privacy problem. The figure shows the home identification uncertainty for each centroid returned from the clustering procedure described earlier, on data sets with sampling intervals of 1, 10, 15, and 20 min. The different bars show the number of correct home identifications for different uncertainty thresholds. Presumably, an adversary will only select locations with high certainty to reduce false positives. Note that even with a sampling interval of 20 minutes, the adversary can still correctly identify a home with high certainty. Reductions in sampling frequency can reduce the probability that samples are taken nearby a driver's home, but this probability is also a function of the length of the trace. If location traces are never discarded, sufficient samples around a user's home will eventually be available. When taking 0.6 as a threshold, an adversary correctly locates 10, 10, 7, and 5 homes under 1, 10, 15, and 20 minutes, respectively. The number of correct centroids by an adversary increases up to 15, 12, 10, and 8 homes with a 0.8 threshold.

5.4 Summary

In summary, we observe the following about existing algorithms:

- Spatial cloaking algorithms that can achieve a guaranteed privacy level for all drivers fail to provide sufficient spatial accuracy for the range of user densities studied in our deployment. For $k = 3$, spatial accuracy remains over 1,000 m, for probe deployments of 2,000 and 5,500 vehicles, one order of magnitude over the applications accuracy requirement. Thus, they are not suitable for probe vehicle systems that operate with low probe densities, or are incrementally deployed over a longer time period.
- Best-effort privacy techniques such as subsampling improve privacy but fail to provide a defined level of privacy for all users. The tracking algorithm described in Section 3 will be able to track some subscribers, particularly those in lower density regions.
- Similarly, best-effort privacy techniques do not fully protect against home identification. While the evaluated home identification intrusion technique suffered from many false positives, this mechanism is at least effective as an automated prefiltering step that can be followed by manual inspection.

To provide a high degree of privacy protection, more sophisticated data suppression mechanisms are needed that can guarantee a level of privacy for all users while maintaining high data quality.

6 Path Privacy Preserving Mechanism

Throughout this section, we develop a disclosure control algorithm that provides privacy guarantee even for users driving in low-density areas. Given a maximum allowable time-to-confusion and a tracking uncertainty threshold, the algorithm can control the release of a stream of received position samples to maintain the tracking time bounds.

Since the algorithm must be aware of the positions of other vehicles, we develop a centralized solution that relies on a trustworthy privacy server. This server filters traces and identities, and thus, prevents exposing sensitive location and identity data to untrusted external service providers. Specifically, it limits the chance that several successive location samples can be linked to the same user, since such partially reconstructed trajectory could act as a quasi-identifier and allow reidentification of users. The use of a trusted location server is widely accepted because it allows users to select one service provider they trust rather than sharing sensitive information with many service providers and it is a natural model for intermediaries that already have access to location information (e.g., cellular service providers) but may need to share it with others.

We first consider the stepwise tracking model without the possibility of path reacquisition. We observe that a specified maximum time to confusion (for a given uncertainty level) can be guaranteed if the algorithm only reveals location samples when 1) the time since the last point of confusion is less than the maximum specified time to confusion or 2) the tracking uncertainty is currently above the specified threshold.

Algorithm 1 shows how this idea can be implemented. Note that it describes processing of data from a single time interval, it would be repeated for each subsequent time slot with the state in the vehicle objects maintained. It takes as input the set of GPS samples reported at time t (v.currentGPSSample updated for each vehicle), the maximum time to confusion (confusionTimeout), and the associated uncertainty threshold (confusionLevel). Its output is a set of GPS samples that can be published while maintaining the specified privacy guarantees.

Algorithm 1. Uncertainty-aware privacy algorithm

```
 1: // Determines which location samples can be release
    while maintaining privacy guarantee.
 2: releaseSet = releaseCandidates = {}
 3: for all vehicles v do
 4:   if start of trip then
 5:       v.lastConfusionTime = t
 6:     else
 7:       v.predictedPos = v.lastVisible.position +
          ($t$-v.lastVisible.time) $*$ v.LastVisible.speed
 8:   end if
 9:
10:   // release all vehicles below timeout
11:   if t - v.lastConfusionTime $<$ confusionTimeout then
12:      add v to releaseSet
13:   else
14:     // consider release of others dependent on
        uncertainty
15:     v.dependencies = k vehicles closest to the
        predictedPos
16:     if uncertainty(v.predictedPos, v.dependencies) $>$
        confusionLevel then
17:       add v to releaseCandidates
18:     end if
19:   end if
20: end for
21:
22: // prune releaseCandidates
23: for all v $\in$ releaseCandidates do
24:     if $\exists$ w $\in$ v.dependencies, w $\ni$ releaseCandidates
        $\cup$ releaseSet then
25:       if uncertainty(v.predictedPos, v.dependencies $\cap$
          (releaseCandidates $\cup$ releaseSet)) $<$
          confusionLevel then
26:         delete v from releaseCandidates
27:       end if
28:     end if
29: end for
30: repeat pruning until no more candidates to remove
31: releaseSet = releaseSet $\cup$ releaseCandidates
32:
33: // release GPS samples and update time of confusion
34: for all v $\in$ releaseSet do
35:     publish v.currentGPSSample
36:     v.lastVisible = v.currentGPSSample
37:     neighbors = k closest vehicles to v.predictedPos
        in releaseSet
38:     if uncertainty(v.predictedPos, neighbors) >=
        confusionLevel then
39:         v.lastConfusionTime=t
40:   end if
41: end for
```

The algorithm proceeds as follows: It first predicts the current position of each vehicle based on prior observations (line 7f.); note that speed is a vector that includes heading information. Second, it identifies the vehicles that can be safely revealed because less time than confusionTimeout has passed since the last point of confusion (line 11f.) Third, it identifies a set of vehicles that can be revealed because current tracking uncertainty is higher than specified in confusionLevel (lines 14-30). Finally, it updates the time of the last confusion point and the last visible GPS sample for each vehicle (line 32ff., the latter is needed for path prediction in the uncertainty calculation). This step can only be performed when the set of revealed GPS samples had been decided, since confusion should only be calculated over the revealed samples.

The third step relies on several approximations. To reduce the computational complexity, it calculates tracking uncertainty only with the k closest samples to the prediction point, rather than with all samples reported at time t. This is a conservative approximation, since uncertainty would increase if additional samples are taken into account (see the proof in the Appendix, which can be found on the Computer Society Digital Library at http://doi.ieeecomputersociety.org/10.1109/TMC.2010.62). We illustrate the effect of k on the computation complexity and the number of released samples in Section 7. Second, it builds a set of releaseCandidates since uncertainty should only be calculated with released samples, but the set of released samples is not determined yet. The algorithm subsequently prunes the candidate set until only vehicles remain who meet the uncertainty threshold. The key property to achieve after the pruning step is that $\forall$ v $\in$ releaseCandidates, uncertainty(v.predictedPos, k closest neighbors in releaseSet $\cup$ releaseCandidates) $\geq$ confusionLevel. The algorithm uses the approximation of calculating the k closest neighbors before the pruning phase, and ensuring during pruning that only vehicles remain if all k neighbors are in the set. While this approximation could be improved in order to release more samples, the current version is sufficient to maintain the privacy guarantee.

6.1 Algorithm Extensions for the Reacquisition Tracking Model

The algorithm described so far does not provide adequate privacy guarantees under the reacquisition tracking model because it only ensures a single point of confusion after the maximum time to confusion has expired.

We observe that such reacquisitions are only possible over short timescales, since movements after more than several minutes become too unpredictable. To verify this assumption, Fig. 10b shows the longest reacquisition and distribution of reacquisition length in minutes, empirically obtained from our data set. As expected, no reacquisitions occur over gaps longer than 10 minutes. Thus, the following extensions can prevent reacquisitions within a time window w. For the experiments reported in the following section, we set $w = 10$:

- **After the *confusionTimeout* expires:** In addition to maintaining confusion from the last released position, it is calculated from every prior released

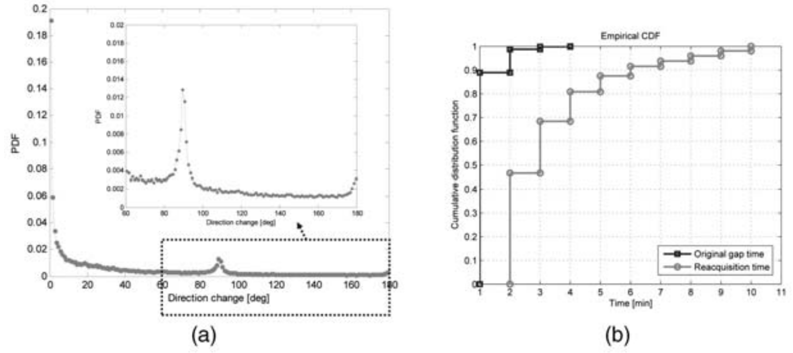


Fig. 10. (a) PDF of heading difference between successive locations. Joint PDF is computed by the multiplication of PDFs of distance gap and direction change. (b) Cumulative distribution function of reacquisitions. No reacquisition occurs over gaps longer than 10 minutes.

location sample (of the same vehicle) within the last w minutes. Samples can only be released if all these confusion values are above the confusion threshold.

- **Before the *confusionTimeout* expires:** Every released sample must maintain confusion to any samples that are released during the last w minutes *and* before the *confusionTimeout* was last reset.

6.2 Algorithm Extensions for the Place Identification Attack

The proposed algorithm, by virtue of its design, automatically identifies the low-density areas and removes location samples in those areas. This property of the algorithm helps prevent home identification since home identification is the easiest in lower density residential areas, and the algorithm removes location samples in these areas.

To further strengthen protection for sensitive origins or destinations such as hospitals, intuitively, one needs to remove all location updates until between the sensitive place and high-confusion (i.e., crowded) area. To incorporate this into the algorithm, we introduce two extensions as follows:

- First, we modified the algorithm not to apply time-out windowing until the confusion for at least one location update exceeds *confusionLevel*. This means that for traces originating in a lower density area, location updates will not be filtered out until the user has arrived in a higher density area where confusion exceeds the specified threshold.
- Second, we modified the algorithm to disable time-out windowing during last T_{guard} minutes before a location trace stops. The T_{guard} interval should be the same length as the maximum TTC. This means that the last part of a trace leading to a destination in a low-confusion area will not be released. Since the algorithm does not know a priori where the destination is or when a trace will stop, it has to delay the release of all location samples by T_{guard} minutes. This delay allows discarding the last updates if a trace stops.

7 Experimental Evaluation

In this section, we present the experimental evaluation of the proposed privacy preserving techniques. Specifically, we demonstrate the effectiveness of our proposed techniques for privacy protection in the analysis of GPS traces. The analysis of the evaluation includes privacy preservation against home identification and target tracking attacks. Also, we evaluate how our proposed privacy preserving techniques can maintain the quality of service for the traffic monitoring application.

7.1 Experimental Setup

Experimental data sets. Throughout the experiments, we used (offline collected) real GPS traces from 312 probe vehicles in our trace-driven simulations. In the experiments, we first applied privacy preserving techniques (i.e., the proposed one and the baseline) on the GPS traces and then measured the performance of privacy protection using target tracking and home identification techniques on these privacy-preserved GPS traces.

Since target tracking typically is only effective for a short time period, we only use 24-hour GPS traces out of a set of week-long GPS traces. This approach helps create a high-density scenario (500 and 2,000 probe vehicles on a $70\ \mathrm{km}^2$ region) with a limited number of probe vehicles. We overlay GPS traces of different volunteer drivers at the same time frame (24 hours) of different dates. This overlay method has a limitation in that it generates similar routes by aggregating GPS traces from the same set of drivers. However, we still believe that it provides insights into higher density scenarios. We will revisit this limitation in Section 9.

Evaluation metrics. In our experiments, we applied the following metrics to evaluate our privacy preserving algorithms for GPS traces.

Tracking Time. Minimizing tracking time reduces the risk that an adversary can correlate an identity with sensitive locations. We use *time to confusion (TTC)*, which we defined in Section 3 as a privacy metric, to measure the tracking duration. To better demonstrate the bounded privacy protection of our proposed algorithm, we report two statistics: the maximum value of TTC and the median value of TTC.

(Relative) Weighted Road Coverage. Through this metric, we measure the data quality that the privacy-preserved traces provide for the traffic monitoring applications. Also, this metric captures the value of each sample based on whether sampled on busier roads or not. Since privacy protection techniques, in general, incur a trade-off between privacy protection and quality of service, our proposed solution aims to provide reasonable privacy protection while delivering the same road coverage for satisfying the need of the traffic monitoring applications. In this paper, we use *relative* road coverage as we defined in Section 2.1. In addition to this metric, we also provide the percentage of

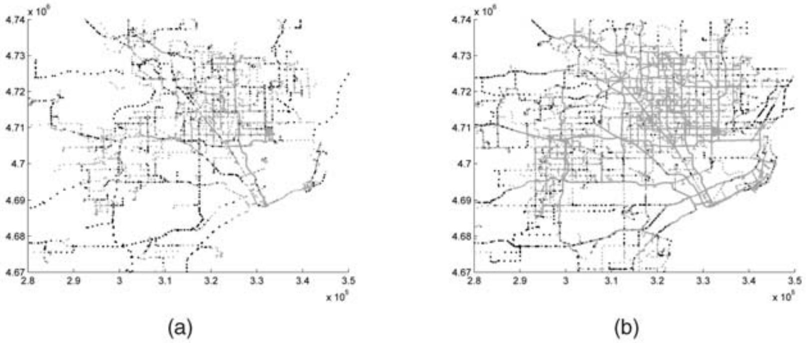


Fig. 11. Uncertainty-aware privacy algorithm removes more samples in low-density areas in which vehicles could be easily tracked. Gray dots indicate released location samples and black ones denote removed samples. (a) Snapshot of privacy preserving GPS traces generated by uncertainty-aware path cloaking at off-peak time (over 1.5 hours) in a high-density scenario. (b) Snapshot of privacy preserving GPS traces generated by uncertainty-aware path cloaking algorithm at peak time (over 1.5 hours) in a high-density scenario.

released location sample compared to the original traces which we consider 100 percent. Note that both metrics are normalized by values of the original GPS traces.

Home Identification Rate. This metric measures the percentage of plausible home position identifications. This measure acts as a proxy for the chance of reidentifying a user. Since no ground truth is available, we have manually inspected the *unmodified* traces and chosen selected 65 traces, where the vehicle visited one residential building significantly more frequently than others. We marked the position of this building as a likely real home position and measure which percentage of these positions is also selected by the automated home identification algorithm based on the *privacy-enhanced* traces. We also measure *false positives*, positions selected by the algorithm that do not match the manually chosen ones, to provide an indication of the accuracy of the algorithm.

7.2 Snapshots of Privacy Preserving GPS Traces

Before evaluating the performance of our proposed technique, let us compare the privacy-preserved GPS traces generated by the proposed path cloaking algorithm with the original GPS traces to highlight major changes in the modified traces. Figs. 11a and 11b show both in a high user density scenario for off-peak (over 1.5 hours at 10 am) and peak time (over 1.5 hours at 5 pm), respectively. Gray dots indicate released location samples while black dots illustrate samples removed by path cloaking. We observe two characteristics from these traces. First, uncertainty-aware path cloaking removes fewer location samples at peak time, and second, it retains more location samples within the presumably busier downtown area. This illustrates how the algorithm, by virtue of its design, retains information on busier roads, where traffic information is most valuable.

7.3 Balancing Computation Load against Data Quality Using k

To illustrate the effect of the parameter k (chosen by the system designer) in the uncertainty-aware path cloaking algorithm on the computational load and the data quality, we conduct experiments where we vary k (i.e., $k = 2, 3, 5, 10, 20$). We measure the computational load in terms of simulation time (seconds) with our Java implementation on an Intel Core2 Duo Processor (2.0 GHz) with 2 GB RAM and the data quality in terms of the number of retained samples. As shown in Fig. 12, there exists a trade-off between data quality and computational complexity. A larger k results in increased computational complexity of the cloaking algorithm. However, the more accurate confusion computation with larger k values helps retain more GPS samples in the filtered trace. A smaller k leads to removing more GPS samples, which, in turn, can reduce the quality of the data set. Note, however, that a smaller k always provides a lower bound on real uncertainty, thus, it is always a conservative choice from a privacy perspective. In Section 8, we use the two most relevant candidates in tracking uncertainty computation ($k = 2$).

8 Results

The following target tracking experiment illustrates how the path cloaking algorithm prevents an adversary from reconstructing an individual's path *and* locating an individual's home using the cleansed GPS traces. Specifically, we compare our uncertainty-aware privacy algorithm and its *with-reacquisition* version with random subsampling in terms of maximum and median TTC for configurations that produce the same number of released location samples (as a metric of data quality). Also, we compare the same algorithms in terms of home identification rate against the number of released location samples. We evaluate the effectiveness of our proposed privacy preserving algorithms by answering the following questions:

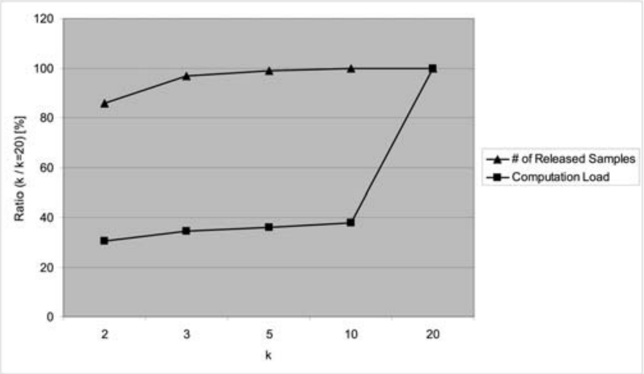


Fig. 12. Choosing k creates a trade-off between data quality and computation complexity. While reducing the computation complexity, the case of $k = 2$ provides a lower bound on uncertainty. This leads to retaining smaller numbers of GPS samples.

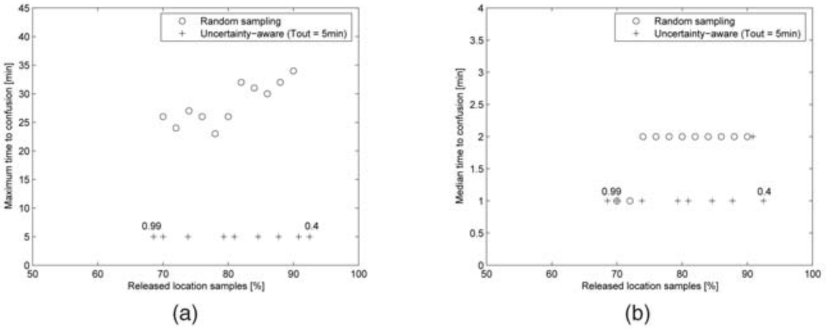


Fig. 13. Maximum/median tracking duration for different privacy algorithms in high-density scenarios (2,000 vehicles/1,600 sqm). The uncertainty-aware privacy algorithm outperforms random sampling for a given number of released location samples. (a) The maximum value of TTC using uncertainty-aware privacy algorithm without reacquisition. (b) The median value of TTC using uncertainty-aware privacy algorithm without reacquisition.

- Do uncertainty-aware privacy algorithms effectively limit tracking time (i.e., guarantee time-to-confusion)? Are these limits maintained even in low user density scenarios?
- How does the average tracking time allowed by path cloaking compare to the subsampling baseline, at the same data quality level?
- How are the results affected by the choice of data quality metric (percentage of released location samples versus relative weighted road coverage)?
- Do the proposed algorithms effectively suppress home identification risks? How efficient is it compared to the subsampling baseline for the same data quality level?
- How much quality should the proposed solutions sacrifice for protecting driver's privacy against target tracking model that utilizes driving direction information?

8.1 Protection against Target Tracking

Throughout the results presented in the following sections, one graph depicts many experiment trials, where one trial comprises the following steps. We first apply a privacy algorithm to the low-density (500 vehicle) or high-density (2,000 vehicle) data set generated from the 312 original vehicle traces. We then remove vehicle identifiers and execute the target tracking algorithm (see Section 3) to measure tracking time for the first 312 vehicles. For each vehicle, we compute the tracking time starting from each sample of the trace and report the maximum. One data point shown in the graph then corresponds to the median or maximum over the 312 vehicle tracking times computed for one trial. For each graph, these trials are then repeated with different uncertainty thresholds for the path cloaking algorithms and different probabilities of removal in the subsampling algorithm.

Bounded tracking time without reacquisition. First, we ascertain whether the uncertainty-aware privacy algorithm guarantees bounded tracking under the no reacquisition tracking assumption. Figs. 13a and 13b show the maximum and median tracking time plotted against the relative amount of released location samples, respectively, for a high-density scenario with 2,000 vehicles in the $70 \text{ km} \times 70 \text{ km}$ area. Fig. 13a shows results for the uncertainty-aware privacy algorithm (marked with $+$) for varying uncertainty levels with time-out fixed at 5 minutes and for the random subsampling algorithm for varying probabilities of removal. Since the configuration parameters from these algorithms are not directly comparable, the graph shows the percentage of released location samples on the x-axis, allowing comparison of TTC at the same data quality level. Also note that graph compares the algorithms in terms of maximum tracking time to illustrate differences in tracking time variance and outliers. During tracking, we set the adversary's uncertainty threshold to 0.4. This means that the adversary will give up tracking if at any point the uncertainty level rises above this threshold, because the correct trace cannot be determined. A 0.4 uncertainty level corresponds to a minimum probability of 0.92 for the most probable next location sample.

As evident from the data, the uncertainty-aware privacy algorithm effectively limits time to confusion to 5 min, except for very low privacy settings (i.e., low uncertainty threshold less than 0.4), while the random sampling algorithm allows some vehicles to be tracked up to about 35 min. Our proposed algorithm can release up to 92.5 percent of original location samples while achieving the bounded tracking property.

In Fig. 13b, we see that naturally occurring crossings and merges in the paths of nearby vehicles lowers median TTC to 1 or 2 minutes (with reacquisition it would be higher, though). However, with random subsampling (20 percent removal), about 15 percent of vehicles (34 out of 233) can still be tracked longer than 10 minutes. The uncertainty-aware path cloaking can guarantee the specified maximum tracking time of 5 min even for these vehicles with higher data quality, removing only 17.5 percent of samples.

Dependence on reacquisition and density. We now repeat the same experiment under the reacquisition tracking model, where an adversary may skip ahead over a point of confusion. Fig. 14a (note scaled x-axis) shows that the uncertainty-aware privacy algorithm with reacquisition extensions can also effectively limit tracking time under this model, while subsampling allows a worst-case tracking time of 42 min. The maximum allowable amount of released location samples is decreased compared to that of Fig. 13.

Let us now investigate whether the privacy guarantee is also maintained in a very low user density scenario with only 500 probe vehicles. Fig. 14b shows that this is indeed the case with the reacquisition model. Note that the privacy is also guaranteed in the case without the reacquisition model (refer to conference version [25]). While subsampling allows a longer maximum TTC due to the low user density, our proposed scheme still preserves the maximum TTC guarantee of 5 minutes for uncertainty thresholds between

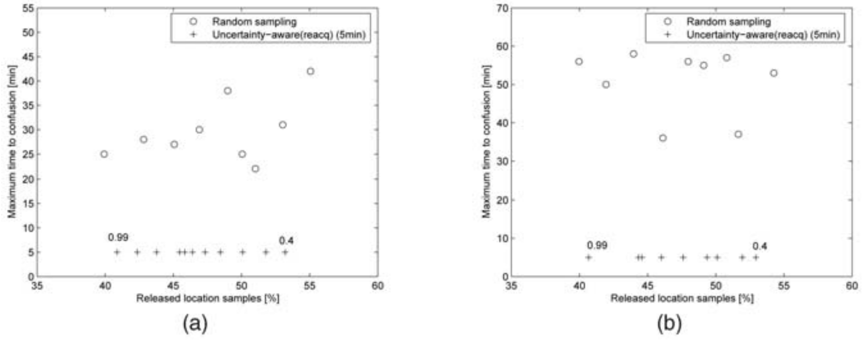


Fig. 14. Maximum value of TTC under the reacquisition tracking model. The uncertainty-aware path cloaking (with reacquisition) version outperforms a random subsampling at a given range of sample removal regardless of density. (a) High-density scenario (2,000 vehicles/1,600 sqm). (b) Low-density scenario (5,000 vehicles/1,600 sqm).

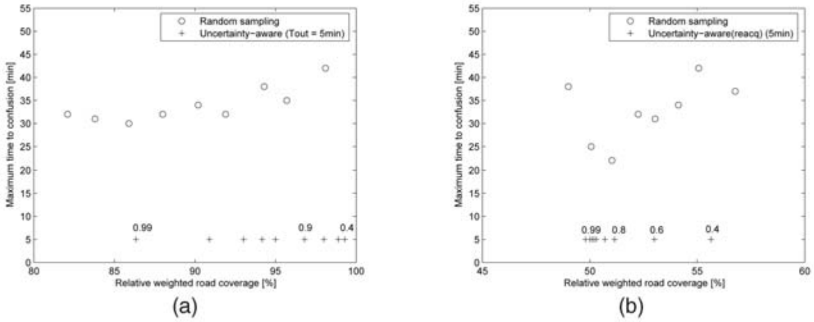


Fig. 15. Time-to-confusion advantages of uncertainty-aware path cloaking become even more pronounced when comparing algorithms with the traffic-monitoring-specific (relative) weighted road coverage data quality metric. (a) Comparison of maximum TTC against weighted road coverage in high-density scenario (uncertainty-aware privacy algorithm). (b) Comparison of maximum TTC against weighted road coverage in high-density scenario ((with reacquisition) uncertainty-aware privacy algorithm).

TABLE 3
Quality-of-Service Enhancement in Each of Uncertainty-Aware Privacy Algorithm (with Reacquisition) Uncertainty-Aware Privacy Algorithm, and Random Sampling Compared to the QoS Level Which Original Traces Can Achieve

	QoS metrics	
	Released location samples	Weighted road coverage
Original traces	100%	100%
Uncertainty-aware privacy (5min,0.95)	81%	95.0%
Random sampling (0.8)	80%	79.3%
(with reacq) Uncertainty-aware (5min,0.4)	53.2%	55.6%
Random sampling (0.53)	53%	52.9%

0.4 and 0.99. Compared to the high-density scenario, our proposed algorithm requires removing more samples to achieve the bounded tracking property in the lower user density scenario.

Quality-of-service analysis. So far, we have measured quality of service in terms of the percentage of samples removed by the algorithm. Since samples in higher density areas are more important for the traffic monitoring application, the benefits of our proposed privacy algorithm are even more significant if we consider *relative weighted road coverage* as shown in Figs. 15a and 15b. We select a few points from Figs. 15a and 15b for random sampling and uncertainty-aware path cloaking to have similar numbers of released location samples for a fair comparison, and we observed how each approach has improved weighted road coverage. More details are shown in Table 3. It shows that the uncertainty-aware privacy algorithm achieves a relative weighted road coverage similar to that of original location traces even though the actual number of released location samples is lower than that of the original location traces. Fig. 11 explains this result; the algorithm retains most samples in high-density areas and removes most from lower densities. However, the uncertainty-aware privacy algorithm with reacquisition extensions provides a slight improvement of relative QoS for weighted road coverage.

8.2 Protection against Home Identification

In this section, we demonstrate the effectiveness of our proposed algorithm against the home identification techniques of a map-aware adversary.

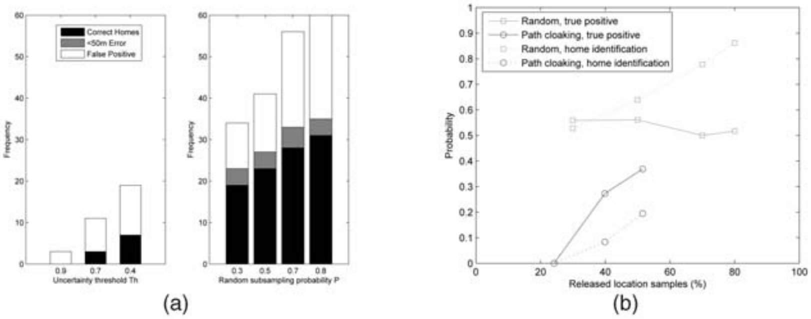


Fig. 16. Note that removing 70 percent of location traces still allows 19 out of 37 homes identified correctly and 4 homes narrowed down within 50 m. (a) Home identification details. (b) Random sampling versus path cloaking.

The following experiment illustrates how an uncertainty-aware path cloaking algorithm suppresses the home identification risk. We compare our proposed algorithm with the random subsampling baseline that we also used in the target tracking analysis, but now focus on the trade-off between home identification risks and data quality.

We select the subset of 37 homes marked by (white) house symbol in Fig. 9a for this evaluation. These lie in dense residential areas and are home locations, where complete one-week set of arrivals and departures has been recorded. Such longer traces will make it easier for the adversary to determine the home location, thus they represent the worst-case privacy within our data set. In this experiment, we did not overlay GPS traces from different dates, since this would lead to unrealistic distributions of user near home locations. Thus, we can only consider a relatively low-density scenario with 312 week-long traces.

Fig. 16a shows the home identification depending on algorithm parameters. For random subsampling, we varied a probability of anonymous location sample selection in the range 0.3-0.8. To achieve a similar percentage of release location samples (i.e., similar data quality), we varied the uncertainty threshold in the path cloaking algorithm from 0.9 to 0.4. After executing the home identification algorithm, we evaluate the following metrics: 1) number of cluster centroids that exactly point to correct homes (buildings), 2) number of clusters centroids that are located within 50 m from the correct home, and 3) number of clusters that point to other buildings or homes that are not found in a set of manually identified homes (or the so-called false positive). Each bar in Fig. 16a represents a tuple of three numbers for each method. Note that with the baseline subsampling algorithm, even after removing 70 percent of location data, 19 out of 37 homes can still be correctly identified and 4 additional homes can be narrowed down to within 50 m. The path cloaking algorithm, however, significantly reduces the number of homes that can be identified, and in the 0.9 uncertainty threshold case, does not allow any home to be identified.

To enable a more precise comparison in terms of quality of service, we also provide a quality versus privacy graph in Fig. 16b. Recall that we only use 312 traces so that more location samples must be withheld than what we observed in target tracking analysis with identical values of uncertainty threshold. For our proposed technique, we plot two different metrics, *true positive* meaning how many homes are correct among the estimated home locations, and *home identification rate*, meaning how many homes out of 37 manually identified homes are correctly detected. The former metric is more meaningful to evaluate the confidence of an adversary who does not know a priori users' home locations. The latter metric describes the absolute number of correctly identified homes. Note that random subsampling returns a constant level of true positive even though we decrease a selection probability. Compared to random subsampling techniques, the proposed path cloaking techniques better preserve user privacy against home identification attack.

Protection against place identification. The uncertainty-aware path cloaking algorithm will also offer protection for other places travelers visit, particularly those not that popular. By removing location updates from low-density areas, it pushes out centroids of an adversary's clustering toward high-density roads, which frequently protects against building identification. It does allow some location samples close to buildings where many different users visit at the same time such as shopping malls or large work sites as shown in Fig. 17b, however. One can argue that places visited by large numbers of different people tend to be less private than places visited only by few people. It also does not compromise user privacy because the popular destinations such as shopping malls generally do not allow inferences about a single user's identity. The two snapshots in Fig. 17 qualitatively illustrate our observations.

8.3 Outlier Removal against Adversary Model with Heading Information

Heading information does not help track an anonymous target a lot in relatively large sample interval such as 1 minute in our data set. The benefit becomes larger when either time interval decreases or vehicles do not change their driving directions much (e.g., when running on a straight highway). For instance, heading information can be used in pruning unlikely hypotheses by assigning higher likelihoods to candidate samples with similar heading information because vehicles do not change directions abruptly while driving on highways. We observed several cases in the analysis of real GPS traces, where the enhanced algorithm in Section 6 prevents tracking outliers, utilizing heading information in tracking uncertainty computation.

Fig. 18 illustrates one example observed in our collected location traces. Fig. 18a shows real GPS traces from five different probe vehicles, one of which (colored in green) runs in an opposite direction from others. Fig. 18b displays two different sets of the modified location traces: 1) the modified location traces that our uncertainty-aware path cloaking computes without heading information and 2) with heading information. Note that our modified version of the algorithm effectively senses the dissimilarity between the heading information of the target and those of candidate samples,

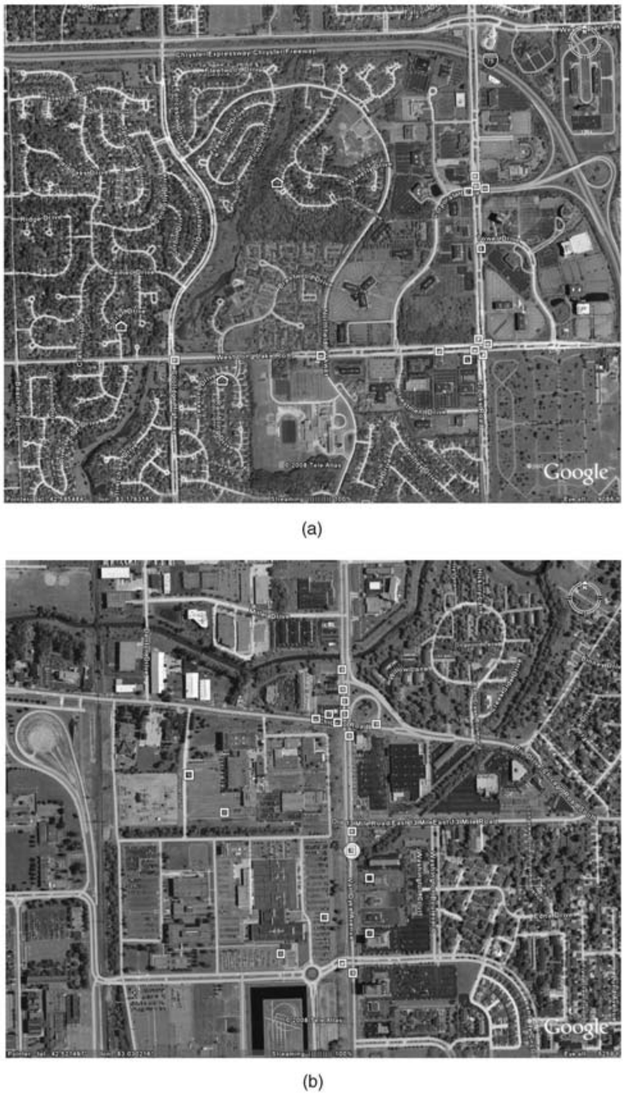


Fig. 17. The uncertainty-aware path cloaking pushes clusters toward roads from residential areas (see (a)). However, it still leaves clusters near destinations such as workplaces, where multiple users visit at the same time (as shown in (b)). The symbol of white house and the rectangle symbol depict manually identified homes and the estimated homes, respectively. (a) Clusters around homes after uncertainty-aware patch cloaking. (b) Clusters around workplaces after uncertainty-aware patch cloaking.

and finally, removes the trajectory of the target until a few samples with a similar direction exist.

To show that our uncertainty-aware path cloaking algorithm detects outliers based on driving behaviors (i.e., speed and direction) as well as density, we measure how many samples are additionally removed out from original traces if heading information is considered. Table 4 summarizes the percentage of released samples (relative to original traces) and the relative weighted road coverage of two modified traces, depending on whether heading information is considered or not.

9 Discussion

In this section, we discuss the limitation of the data set and the experiment, the possible extensions of our proposed algorithm, and some future directions.

Ground truth. Since the real home addresses are unavailable (driver identities were omitted in the data set for privacy reasons), we manually inspected the unmodified week-long traces to identify likely home locations for use as ground truth. We overlayed the traces on satellite images for this purpose. This inspection led to 65 reference locations, shown in Fig. 9a. These were chosen because they contained a single home that stood out as a likely home location and the drivers visited this home much more frequently at night than other locations. Therefore, we believe the manual inspection provides a reasonable approximation of real home locations.

Prior knowledge on subjects. In this work, we assume that an adversary does not have any a priori knowledge about the subjects being tracked when we develop inference attack models. We cannot rule out that inferences are possible with additional information about subjects. For example, if the destination and a likely path of a subject's trip are a priori known, tracking probabilities may change and it may make it easier to link anonymous location updates to this subject. Of course, in this model, there is less sensitive information to protect because the destination is already known, but it still leaks some information about the exact timing and speed of the trip, making it possible to determine whether a subject was present at an accident site, for example. Still, we believe that the cloaking algorithm significantly raises the bar for extracting information from traces.

Relaxing trust in location server. The algorithm described so far relies on a trustworthy location server, since the algorithm needs the full GPS traces of all vehicles. A fully distributed algorithm poses a research challenge by itself, since clients would need to monitor the positions of neighboring cars, which again raises privacy and trust issues. It also appears possible, though, to relax the trust assumptions in the location server through a hybrid approach, with additional in-vehicle disclosure control based on coarser information about neighbors. Since data quality would only be marginally affected by missing updates in low-density areas, one could devise schemes to inform vehicles of the approximate probe density in their area. Then vehicles could reduce location updates to the server in the most sensitive low-density areas. To prevent spoofing of such density information, further research could investigate data cross-validation schemes or secure multiparty computation schemes to compute density.

Data set limitations. We need to point out that the tracking results can be affected by the choice of probe vehicles. In our data set, most drivers shared the same workplace. Thus, the workplace acted as a place of confusion, where the tracking algorithms failed. A random sample of the population would probably improve tracking performance. This would cause both our proposed algorithms and the random sampling method to remove more samples to meet the maximum TTC. The performance gap between them might also change from what we have observed in our study. In addition, our method of overlaying multiple data sets to create one high-density scenario may not be entirely faithful in representing true traffic conditions. Due to this overlay, some of the vehicles may also be driven by the same driver on similar routes, creating a further bias toward reduced tracking performance. Nonetheless, we observed that naive anonymization is problematic and our proposed algorithm filters out much less data

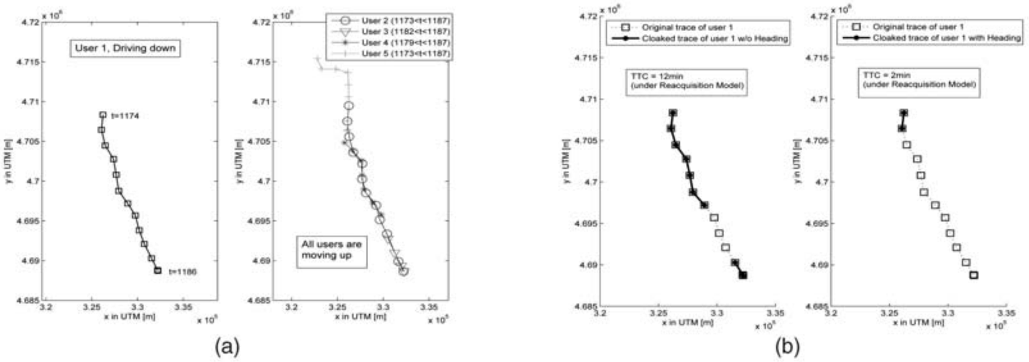


Fig. 18. The enhanced uncertainty-aware path cloaking algorithm removes not only location traces in low-density areas but also a location trace driving in a reverse direction from majority of surrounding probe vehicles. Each point is depicted by UTM coordinates (x, y) in meters. Modified trace of user 1 without heading information allows a longer tracking ($= 12$ mins) against the enhanced adversary model. The reconstructed path of User 1 is same as the modified paths as shown in (b). (a) Original GPS traces of five anonymous Vehicles (users 1-5). (b) Modified GPS traces of user 1 without heading information (left) and with heading information (right).

TABLE 4
Quality-of-Service Degradation due to Enhanced Tracking

	Probe Vehicles (500)		Probe Vehicles (2000)	
	w/o Heading (%)	w/ Heading (%)	w/o Heading (%)	w/ Heading (%)
Uncertainty threshold=0.9	74.2 (90.2)	64.1 (79.4)	84.4 (96.8)	73.6 (87.8)
Uncertainty threshold=0.95	69.2 (86.7)	53.8 (67.0)	80.9 (95.2)	64.1 (78.5)
Uncertainty threshold=0.98	62.0 (80.0)	40.6 (50.9)	75.2 (91.9)	48.9 (60.6)
Uncertainty threshold=0.99	55.7 (74.2)	34.9 (44.9)	70.2 (88.2)	40.2 (49.0)

Each entry denotes the percentage of released location samples, and each value in () denotes relative weighted road coverage.

than baseline algorithms. We still believe that our current results provide a valuable step toward understanding the tracking performance in probe vehicle scenarios.

Variants of target tracking algorithms. We also considered other possible tracking algorithms than the ones described so far. We cannot rule out that more sophisticated tracking techniques can achieve longer tracking times, but the algorithms we experimented with only showed very incremental gains compared to our tracking model.

- Linear Kalman Model: We observed that linear Kalman filtering does not enhance the tracking capability. The linear Kalman model is an effective tool to estimate the state (e.g., position, speed, and acceleration) of a system (e.g., vehicles) given a time series of noisy observations. Accurately estimated state then enables more accurate prediction of the next position of the moving target. In our data set, the noise in GPS samples on the order of a few meters was not a dominant factor compared to the relatively low sampling rate of one per minute.
- Tracking with road map information: If we use road network information, we can achieve better pruning over a set of hypotheses. For example, even though two observed samples are near in euclidean distance, it may be obvious that they do not belong to a same user if they are on very different roadways. While experiments we conducted on our data set did not show significant improvements in tracking performance, we generally expect that the use of map information helps tracking. In our study, we have focused on computationally simpler algorithms that could be applied to massive number of targets traces without sophisticated knowledge such as map information or a prior knowledge on subjects to be tracked. However, it is straightforward to extend the presented privacy model and algorithm to take into account road network information. Instead of using euclidean distance in the tracking model, the algorithm could match locations to road segments and calculate the road distance between two locations.

10 Related Work

Much work has focused on exploiting GPS or cellular phone data for traffic monitoring. MIT's CarTel [26] proposed to use the unused bandwidth of open wireless hotspots to deliver the GPS-based location and speed measurements of probe vehicles to the central server for traffic data mining and locating potholes. Previous study using cell-phone-based traffic monitoring [13] investigates the use of triangulation-based positioning technology to locate phones. Because of poor quality position estimates (100 m accuracy), vehicle speeds could not be consistently determined. Recently, Yoon et al. [42] proposed to use cellular network as a delivery method of GPS-based sensing information from probe vehicles. Several systems have also been deployed [5], [6]. These projects, however, have not focused on privacy of location information, and use only basic anonymization techniques, if using privacy enhancing techniques at all.

Several recent studies [32], [24], [21] analyzed the privacy risk of GPS traces and found that naive anonymization (i.e., omitting identifiers from a data set) does not guarantee anonymity due to a spatiotemporal correlation between periodic location updates. This raises an urgent need for stronger protection mechanisms.

Much research exists on guaranteeing anonymity in database records and k-anonymity [36] solutions are available, but their application for *time-series* location data set is questionable. The k-anonymity concept has been applied in location-based services [20], [34], [19] for single independent location updates. As shown in Section 5, these solutions can provide sufficient accuracy for applications such as point-of-interest queries in high-density scenarios, but they do not achieve the high-accuracy requirements of traffic monitoring applications with low penetration rates. In addition, a series of cloaking boxes applied to periodic location updates still allows an adversary to follow a target [41]. Many studies have subsequently extended the k-anonymity concept to allow cloaking through the use of hilbert curves [28], efficient cloaking of paths [41], and cloaking algorithms for l-diversity as well as k-anonymity [40]. Bettini et al. [12] recently provided a formal framework to define attack scenarios, defense techniques, and assumptions on the amount of knowledge that is accessible by an adversary.

Privacy preserving data mining [8] and anonymous communication are also not directly applicable to time-series location data. Random perturbation approaches cannot provide sufficient data accuracy and noise with small variance may be sometimes filtered by advanced signal processing techniques [30]. Anonymous communication techniques can relay messages between communication partners without revealing the source and/or destination identity (e.g., [14]), but does not protect potentially revealing information in the message payload. The work on measuring communication anonymity, however, [37] inspired us to use entropy in defining time-to-confusion.

Closest to our work are the best-effort location data protection algorithms [11], [33], [22], [27], which have in common that they create areas of confusion where the traces from several users converge. While these algorithms successfully achieve better accuracy and a defined level of privacy in such an area of confusion, they cannot provide overall privacy guarantees because these areas of confusion might not occur in lower density areas with few users. Recently, two research groups, Apu et al. [29] and Andreas et al. [31], proposed a privacy preserving data collection architecture for collaborative sensing applications. However, both works do not consider inference attacks that utilize existing correlation between location-based updates.

Another proposed approach builds on privacy policy languages [16] and their location-oriented extensions [38] to allow users (or their automated agents) to make more informed decisions about data sharing. Such policies may be enforced through access control mechanisms, such as [18], [43] for spatiotemporal data. Using these approaches, data can only be shared with trusted data consumers, while strong anonymization also allows more public distribution of data.

11 Conclusions

In this paper, we have proposed the time-to-confusion metric and cloaking algorithms to address privacy in an anonymous set of time-series location traces. We considered two specific privacy risks in anonymous location traces—target tracking and place identification—and found that these allow tracking and reidentifying data subjects in anonymous traces, particularly in areas with low user density. We quantify the tracking risk through the time-to-confusion metric and develop the uncertainty-aware path cloaking algorithm, which can filter a set of anonymous GPS traces to guarantee a maximum privacy-risk level (specified as time-to-confusion).

Using a real-world GPS data set, we measure the privacy gain and the achieved data quality for the proposed solutions compared to a baseline random sampling technique. We show that our uncertainty-aware path cloaking effectively guarantees worst-case tracking bounds (i.e., outliers), while achieving significant data accuracy improvements. Since the algorithm considers both density and driving behaviors (i.e., speed and direction), it effectively detects and removes traces that are sampled in low-density areas or could be easily tracked due to differences in driving direction from surrounding vehicles. It achieves better privacy than a random sampling technique at the same level of data quality. We also show that our solution is effective against clustering-based place identification techniques.

Acknowledgments

Most of this work completed while Baik Hoh was a PhD student at Rutgers University.

References

[1] TIER, http://tier.cs.berkeley.edu/wiki/home, 2010.
[2] Chronology of Data Breaches, http://www.privacyrights.org/ar/chrondatabreaches.htm, 2010.
[3] Urban Atmospheres, http://www.urban-atmospheres.net, 2010.
[4] Path Intelligence, http://www.pathintelligence.com, 2010.
[5] INRIX, http://www.inrix.com, 2006.
[6] Intellione, http://www.intellione.com, 2006.
[7] Participatory Urbanism, http://www.urban-atmospheres.net/ParticipatoryUrbanism/index.html, 2008.
[8] R. Agrawal and R. Srikant, "Privacy-Preserving Data Mining," *Proc. ACM SIGMOD*, pp. 439-450, May 2000.
[9] M. Allen, L. Girod, R. Newton, S. Madden, D.T. Blumstein, and D. Estrin, "VoxNet: An Interactive, Rapidly-Deployable Acoustic Monitoring Platform," *Proc. Int'l Conf. Information Processing in Sensor Networks (IPSN '08)*, pp. 371-382, 2008.
[10] M. Barbaro and T. Zeller Jr., "A Face Is Exposed for AOL Searcher No. 4417749," http://www.nytimes.com/2006/08/09/technology/09aol.html, 2010.
[11] A. Beresford and F. Stajano, "Mix Zones: User Privacy in Location-Aware Services," *Proc. IEEE Int'l Workshop Pervasive Computing and Comm. Security (PerSec '04)*, 2004.
[12] C. Bettini, S. Mascetti, X.S. Wang, and S. Jajodia, "Anonymity in Location-Based Services: Towards a General Framework," *Proc. Int'l Conf. Mobile Data Management (MDM '08)*, pp. 69-76, 2007.
[13] R. Cayford and T. Johnson, "Operational Parameters Affecting Use of Anonymous Cell Phone Tracking for Generating Traffic Information," *Proc. Inst. Transportation Studies for the 82nd TRB Ann. Meeting*, vol. 1, no. 3, pp. 03-3865, Jan. 2003.
[14] D. Chaum, "Untraceable Electronic, Mail Return Addresses, and Digital Pseudonyms," *Comm. ACM*, vol. 24, no. 2, pp. 84-90, 1981.
[15] T.M. Cover and J.A. Thomas, *Elements of Information Theory*. Wiley Interscience, 1991.
[16] L. Cranor, M. Langheinrich, M. Marchiori, and J. Reagle, "The Platform for Privacy Preferences 1.0 (P3P1.0) Specification," *W3C Recommendation*, Apr. 2002.
[17] X. Dai, M. Ferman, and R. Roesser, "A Simulation Evaluation of a Real-Time Traffic Information System Using Probe Vehicles," *Proc. IEEE Int'l Conf. Intelligent Transportation Systems*, pp. 475-480, 2003.

[18] A. Gal and V. Atluri, "An Authorization Model for Temporal Data," *Proc. Seventh ACM Conf. Computer and Comm. Security (CCS)*, pp. 144-153, 2000.

[19] B. Gedik and L. Liu, "Location Privacy in Mobile Systems: A Personalized Anonymization Model," *Proc. 25th IEEE Int'l Conf. Distributed Computing Systems (ICDCS '05)*, pp. 620-629, 2005.

[20] M. Gruteser and D. Grunwald, "Anonymous Usage of Location-Based Services through Spatial and Temporal Cloaking," *Proc. ACM Int'l Conf. Mobile Systems, Applications and Services (MobiSys '03)*, 2003.

[21] M. Gruteser and B. Hoh, "On the Anonymity of Periodic Location Samples," *Proc. Second Int'l Conf. Security in Pervasive Computing*, 2005.

[22] B. Hoh and M. Gruteser, "Protecting Location Privacy through Path Confusion," *Proc. IEEE/Create-Net Int'l Conf. Security and Privacy for Emerging Areas in Comm. Networks (SecureComm)*, Sept. 2005.

[23] B. Hoh, M. Gruteser, R. Herring, J. Ban, D. Work, J.-C. Herrera, A. Bayen, M. Annavaram, and Q. Jacobson, "Virtual Trip Lines for Distributed Privacy-Preserving Traffic Monitoring," *Proc. ACM Int'l Conf. Mobile Systems, Applications and Services (MobiSys '08)*, 2008.

[24] B. Hoh, M. Gruteser, H. Xiong, and A. Alrabady, "Enhancing Security and Privacy in Traffic-Monitoring Systems," *IEEE Pervasive Computing*, vol. 5, no. 4, pp. 38-46, Oct. 2006.

[25] B. Hoh, M. Gruteser, H. Xiong, and A. Alrabady, "Preserving Privacy in GPS Traces via Uncertainty-Aware Path Cloaking," *Proc. ACM Conf. Computer and Comm. Security (CCS '07)*, Oct. 2007.

[26] B. Hull, V. Bychkovsky, Y. Zhang, K. Chen, M. Goraczko, A.K. Miu, E. Shih, H. Balakrishnan, and S. Madden, "CarTel: A Distributed Mobile Sensor Computing System," *Proc. Fourth ACM Conf. Embedded Networked Sensor Systems (SenSys '06)*, Nov. 2006.

[27] T. Jiang, H. Wang, and Y.-C. Hu, "Preserving Location Privacy in Wireless LANs," *Proc. Fifth ACM Int'l Conf. Mobile Systems, Applications and Services (MobiSys '07)*, 2007.

[28] P. Kalnis, G. Ghinita, K. Mouratidis, and D. Papadias, "Preventing Location-Based Identity Inference in Anonymous Spatial Queries," *IEEE Trans. Knowledge and Data Eng.*, vol. 19, no. 12, pp. 1719-1733, Dec. 2007.

[29] A. Kapadia, N. Triandopoulos, C. Cornelius, D. Peebles, and D. Kotz, "AnonySense: Opportunistic and Privacy-Preserving Context Collection," *Proc. Sixth Int'l Conf. Pervasive Computing (Pervasive '08)*, May 2008.

[30] H. Kargupta, S. Datta, Q. Wang, and K. Sivakumar, "Random Data Perturbation Techniques and Privacy Preserving Data Mining," *Proc. IEEE Int'l Conf. Data Mining (ICDM '03)*, 2003.

[31] A. Krause, E. Horvitz, A. Kansal, and F. Zhao, "Toward Community Sensing," *Proc. ACM/IEEE Int'l Conf. Information Processing in Sensor Networks (IPSN '08)*, Apr. 2008.

[32] J. Krumm, "Inference Attacks on Location Tracks," *Proc. Fifth Int'l Conf. Pervasive Computing (Pervasive '07)*, May 2007.

[33] M. Li, K. Sampigethaya, L. Huang, and R. Poovendran, "Swing & Swap: User-Centric Approaches Towards Maximizing Location Privacy," *Proc. Fifth ACM Workshop Privacy in the Electronic Soc. (WPES '06)*, pp. 19-28, 2006.

[34] M.F. Mokbel, C.-Y. Chow, and W.G. Aref, "The New Casper: Query Processing for Location Services without Compromising Privacy," *Proc. 32nd Int'l Conf. Very Large Data Bases (VLDB '06)*, VLDB Endowment, pp. 763-774, 2006.

[35] A. Narayanan and V. Shmatikov, "Robust De-Anonymization of Large Datasets," *Proc. IEEE Symp. Security and Privacy*, pp. 111-125, 2008, doi:10.1109/SP.2008.33.

[36] P. Samarati and L. Sweeney, "Protecting Privacy When Disclosing Information: K-Anonymity and Its Enforcement through Generalization and Suppression," *Proc. IEEE Symp. Research in Security and Privacy*, 1998.

[37] A. Serjantov and G. Danezis, "Towards an Information Theoretic Metric for Anonymity," *Proc. Second Workshop Privacy Enhancing Technologies*, 2002.

[38] E. Snekkenes, "Concepts for Personal Location Privacy Policies," *Proc. Third ACM Conf. Electronic Commerce (EC '01)*, pp. 48-57, 2001.

[39] K.P. Tang, P. Keyani, J. Fogarty, and J.I. Hong, "Putting People in Their Place: An Anonymous and Privacy-Sensitive Approach to Collecting Sensed Data in Location-Based Applications," *Proc. Conf. Human Factors in Computing Systems*, pp. 93-102, 2006.

[40] M. Terrovitis and N. Mamoulis, "Privacy Preservation in the Publication of Trajectories," *Proc. Ninth Int'l Conf. Mobile Data Management (MDM '08)*, pp. 65-72, 2008.

[41] T. Xu and Y. Cai, "Exploring Historical Location Data for Anonymity Preservation in Location-Based Services," *Proc. IEEE INFOCOM*, pp. 547-555, 2008.

[42] J. Yoon, B. Noble, and M. Liu, "Surface Street Traffic Estimation," *Proc. Fifth Int'l Conf. Mobile Systems, Applications and Services (MobiSys '07)*, pp. 220-232, 2007.

[43] M. Youssef, V. Atluri, and N.R. Adam, "Preserving Mobile Customer Privacy: An Access Control System for Moving Objects and Customer Profiles," *Proc. Sixth Int'l Conf. Mobile Data Management (MDM '05)*, pp. 67-76, 2005.

Baik Hoh received the BS and MS degrees in electrical and computer engineering from the Korea Advanced Institute of Science and Technology and the PhD degree from the Department of Electrical and Computer Engineering at WINLAB, Rutgers University. He is currently a senior scientist at Nokia Research Center, Palo Alto, California. He was a recipient of the Academic Excellence Award in 2009 from Rutgers University. His research interests include privacy-enhancing technologies in pervasive computing and crowd sourcing applications.

Marco Gruteser received the vordiplom degree from the Darmstadt University of Technology, Germany, in 1998, and the MS and PhD degrees from the University of Colorado at Boulder in 2000 and 2004, all in computer science. He is currently an assistant professor of electrical and computer engineering at WINLAB, Rutgers University, where he is conducting research on location-aware networking, the design of location privacy techniques, and applications in vehicular networks. Previously, he was also a research associate at the IBM T.J. Watson Research Center from 2000 to 2001. He is a recipient of the US National Science Foundation CAREER Award. He also received the Schwartzkopf Prize for Technological Innovation as a member of the ORBIT Wireless Testbed Team and he has served on the program committee of numerous conferences, including MobiSys and INFOCOM, and on the editorial board of the journal *Computer Networks*.

Hui Xiong received the BE degree from the University of Science and Technology of China, the MS degree from the National University of Singapore, and the PhD degree from the University of Minnesota. He is currently an associate professor in the Management Science and Information Systems Department at Rutgers University. His research interests include data and knowledge engineering with a focus on developing effective and efficient data analysis techniques for emerging data-intensive applications. He has published more than 70 technical papers in peer-reviewed journals and conference proceedings. He is the coeditor of *Clustering and Information Retrieval* (Kluwer Academic, 2003) and the coeditor-in-chief of *Encyclopedia of GIS* (Springer, 2008). He is an associate editor of the *Knowledge and Information Systems* journal and has served regularly on the organization committees and program committees of a number of international conferences and workshops. He is a senior member of the IEEE and a member of the ACM.

Ansaf Alrabady received the PhD degree in computer engineering from Wayne State University. He is a senior research engineer at General Motors. His research interests include secure embedded system development for automotive applications. He received the Automotive Hall of Fame Young Leadership and Excellence Award in recognition of his contributions to the automotive industry.