Privacy Vulnerability of Published Anonymous Mobility Traces

Chris Y. T. Ma, David K. Y. Yau, *Member, IEEE*, Nung Kwan Yip, and Nageswara S. V. Rao, *Fellow, IEEE*

***Abstract*—Mobility traces of people and vehicles have been collected and published to assist the design and evaluation of mobile networks, such as large-scale urban sensing networks. Although the published traces are often made anonymous in that the true identities of nodes are replaced by random identifiers, the privacy concern remains. This is because in real life, nodes are open to observations in public spaces, or they may voluntarily or inadvertently disclose partial knowledge of their whereabouts. Thus, snapshots of nodes' location information can be learned by interested third parties, e.g., directly through chance/engineered meetings between the nodes and their observers, or indirectly through casual conversations or other information sources about people. In this paper, we investigate how an *adversary*, when equipped with a small amount of the snapshot information termed as *side information*, can infer an extended view of the whereabouts of a *victim* node appearing in an anonymous trace. Our results quantify the loss of victim nodes' privacy as a function of the nodal mobility, the inference strategies of adversaries, and any noise that may appear in the trace or the side information. Generally, our results indicate that the privacy concern is significant in that a relatively small amount of side information is sufficient for the adversary to infer the true identity (either uniquely or with high probability) of a victim in a set of anonymous traces. For instance, an adversary is able to identify the trace of 30%–50% of the victims when she has collected 10 pieces of side information about a victim.**

***Index Terms*—Mobility traces, privacy, security and protection.**

I. Introduction

MOBILITY traces of people and vehicles have been collected and published to assist the design and evaluation of mobile networks. One example application of such networks is urban sensing, where mobile nodes carried by ordinary city residents or their vehicles are used to monitor various events of interest in their city areas. Example activities include traffic monitoring [25], road surface condition sensing [10], chemical detection [28], and radiation detection [17]. This type of large-coverage, everyday sensing is made possible by advances in sensor technologies, which produce small form-factor, low-power, low-cost, and multimodal sensors that can be readily embedded into widely adopted personal handheld devices including smartphones. Clearly, mobility patterns of potential *real-world* participants in these networks, including their correlations and interactions with each other, will have profound effects on the network performance (e.g., coverage and connectivity of a collaborative sensing network). Indeed, researchers have found that existing synthetic movement models of mobile entities, such as pedestrians and different kinds of vehicles, though attractive for their low cost and high repeatability, generally fail to capture essential behaviors of real users. Therefore, the use of synthetic traces in network design can lead to wrong conclusions about network performance (e.g., routing efficiency) in reality [19]. Hence, there are increasing efforts to trace the locations of real users leading to the public availabilities of many such traces through either consolidated data portals such as Crawdad [7] or Web sites set up by individual research groups [34].

In order to protect the privacy of participants in real user traces, the true identity of each participant is often replaced by a consistent, unique, and random identifier (not correlated in any way with the true user identity). Moreover, the precision of the traces in the spatial and temporal domains can be often reduced by *cloaking* techniques such as reducing the resolution of the recorded data or introducing noise deliberately in the data. It is not clear, however, if these "anonymization" and cloaking techniques are sufficient to protect the privacy of the participants. This is because movements or whereabouts of participants in public spaces can be openly observed by others through chance/engineered meeting opportunities. Similar location/movement information can also be inferred indirectly from conversations, news articles, online social networks, or Web blogs, though the inference could be noisy. By gathering one or a few such (possibly rough) snapshots of a participant's location over time, which we term as *side information*, an adversary may be able to identify (either uniquely or with high probability) the participant's trace from a set of anonymous traces. Hence, the complete whereabouts of the participant (the *victim*) over an extended time duration will be revealed to the adversary.

In this paper, we formulate the above privacy problem. We analytically develop inference strategies that the adversary may use to maximize its effectiveness in identifying one or more victims under different system assumptions. We show

Manuscript received November 22, 2011; revised May 11, 2012; accepted July 02, 2012; approved by IEEE/ACM TRANSACTIONS ON NETWORKING Editor M. Allman. This work was supported in part by the Human Sixth Sense Programme at the Advanced Digital Sciences Center from Singapore's Agency for Science, Technology and Research (A*STAR) under a research grant; the US National Science Foundation under Grants No. CNS-0963715, CNS-0964086, and DMS-0707926; the Mathematics of Complex, Distributed, Interconnected Systems Program, Office of Advanced Computing Research, US Department of Energy; and the National Natural Science Foundation of China (NSFC) under Grant No. 61028007. This work was performed in part at the Advanced Digital Sciences Center, Purdue University, and Oak Ridge National Laboratory managed by UT-Battelle, LLC, for the US Department of Energy under Contract No. DE-AC05-00OR22725.

C. Y. T. Ma is with the Advanced Digital Sciences Center, Illinois at Singapore, Singapore 138632, Singapore (e-mail: chris.ytma@gmail.com).

D. K. Y. Yau is with the Advanced Digital Sciences Center, Illinois at Singapore, Singapore 138632, Singapore, and also with Purdue University, West Lafayette, IN 47907 USA.

N. K. Yip is with Purdue University, West Lafayette, IN 47907 USA.

N. S. V. Rao is with Oak Ridge National Laboratory, Oak Ridge, TN 37831 USA.

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

Digital Object Identifier 10.1109/TNET.2012.2208983

how the adversary can gainfully incorporate general world knowledge—in the form of a *movement model* accounting for global movement constraints and preferences—in its inference strategies. We also quantify experimentally the loss of victim nodes' privacy (possibly as a process over time) as a function of several important system parameters, including the nodal mobility, the inference strategies of the adversaries, and any noise that may appear in the traces or side information (due to either the application of cloaking techniques or inherently imprecise observations). Our contributions are twofold.

1) We provide extensive analysis both theoretically and experimentally to demonstrate that with the current practice of capturing and publishing anonymous location traces of real users, the concern exists that an adversary could identify the traces of one or more victims in the published data with high probability by invoking a small amount of side information about the participants. In particular, we present comprehensive attack strategies available to the adversary when it collects information about a victim's movement either through direct observations or indirect information sources and show that these attacks are effective in breaching privacy. We also provide a mathematical framework to show the optimality of specific attack strategies in that they utilize all the available information in the most effective way.

2) We give comprehensive experimental analysis to show the differences between different real traces from the perspective of the privacy problem. Their different characteristics will result in quite different performance under various privacy attacks.

II. Related Work

Privacy of published data sets has received much attention [2], [3], [23], [37], [41]. Sweeney [37] proposes a privacy measure of *k-anonymity*. When k-anonymity is satisfied, each individual is indistinguishable from at least $k-1$ other individuals. Bayardo and Agrawal [3] propose a practical method to identify a provably optimal k-anonymization of real census data, or a "good" anonymization for general data, since the general problem is NP-hard. The concept of k-anonymization does not capture the diversity of the anonymity set. To solve the issue, Machanavajjhala *et al.* [21] propose an l-diversity measure to ensure diversity in the published data. Li and Li [20] propose a t-closeness metric, which ensures that the distance of a sensitive attribute's distribution in one class is no more than a threshold t from that of the whole table. Xiao and Tao [41] propose *m-invariance* to limit the risk of privacy disclosure in data *republications* since potential correlations among snapshots of data in the different publication instances can be used to derive sensitive information.

Identification of users, or their attributes, who access location-based services has been studied [4], [11], [13]–[15], [18], [24], [35], [36]. Golle and Partridge [14] quantify the likelihood of identifying an individual using her home and working locations and show that revealing at census block level is able to identify most of the US working population. Freudiger *et al.* [11] quantify the probability of identifying the home or office location of a user based on the number of queries issued to a server.

One basic technique to improve location privacy is to reduce the spatial/temporal granularity of the location information given to the service provider while still supporting satisfactory service quality [13], [15]. Hoh *et al.* [18] devise a protection method that releases user data only when certain privacy constraints are met. Meyerowitz and Choudhury [24] propose to send fake requests with real ones in order to reduce one's ability to trace a mobile node over time. Shokri *et al.* [35], [36] propose an evaluation framework for location-privacy protection, assuming that the adversary knows the spatial distribution or transition probabilities of each user between locations. Chow *et al.* [4] use granularity reduction to provide privacy in peer-to-peer systems that support location-based services.

Approaches to improve the privacy of geo-located data sets include data perturbation, data swapping, data generalization or granularity reduction, and data withholding. Abul *et al.* [1] propose the use of space translations to achieve (k,δ)-anonymity for databases of moving objects, where δ is the radius of a cylindrical volume representing the allowed trajectory imprecision. Terrovitis and Mamoulis [39] use the suppression of location information to achieve an acceptable probability of privacy breach. Nergiz *et al.* [29] use a notion of k-anonymity that is specific to trajectories and propose a generalization method to enhance the privacy of published trajectories.

Martin *et al.* [23] quantify how background knowledge possessed by an attacker may impact privacy breach. They express the background knowledge in a language and provide an algorithm to determine the amount of disclosed sensitive information in the worst case as a function of the background knowledge. In a data mining context, Agrawak and Srikant [2] propose a reconstruction method to build a decision-tree classifier without accessing precise information in individual data records, so that the data value distributions can be reconstructed with sufficient accuracy. They also propose value-class memberships and value distortions as privacy preservation techniques.

The literature above assumes that an attacker has limited knowledge and power and analyzes privacy and its protection in application-specific situations. We take a similar approach. Our specific focus is on the privacy of anonymous mobility traces as they are published in various public data portals [7], [34]. Our analysis assumes basic spatial and temporal cloaking techniques since the basic protection can more easily ensure the applicability of the data sets for diverse application scenarios, which befits the intention of the data portals. Please see Section VII for a discussion.

Currently, differential privacy (DP) is an extremely active research area. It is important because it adopts a strong notion of privacy that does not limit the power of the attacker and measures privacy loss by basic information metrics. Dwork *et al.* [8] consider how much noise is needed to perturb true answers from a statistical database in order to preserve privacy. They show that the extent of noise needed is proportional to the sensitivity of the query function. Ho and Ruan [16] propose to provide DP by dynamic sizing of grid cells and addition of noise to data sets. Machanavajjhala *et al.* [22] use a modified form of DP to have the published statistics match more closely with the actual statistics, without breaching privacy. Rastogi and Nath [32] propose an algorithm to ensure DP using transformation and encryption, such that users can compute the amount of noise needed to perturb the published data in a distributed manner, while keeping the noise in the statistics obtained by the aggregator small. They

mitigate the problem that for time-series data, such as our mobility traces, the amount of noise needed for standard DP approaches will in the worst case grow linearly with the number of queries. Despite its importance, DP's assumptions about the use of the data sets are fundamentally incompatible with our problem context. Please see Section VII for a discussion.

III. Problem Definition

We assume that a set of traces, each of which is recording intermittently the time and corresponding location of a mobile node, are released to the public. We call a node that is included in a trace set a *participant* in the trace set. The samples can be collected using, say, a GPS-enabled device carried by the participant, which reports the participant's location and the corresponding time periodically to a data collector. The traces are anonymous in that the true identity of a participant has been replaced by a random and unique identifier. The true node identity is not correlated in any way to the random identifier, but the same true identity is always mapped to the same random identifier. The times at which locations of a participant are recorded in a trace are called the *sampled times*. We assume that the recorded participant location at a sampled time, say t, is imprecise for anonymization purpose as explained in Section I. Specifically, instead of recording the precise point in space p at which the node is located at time t, the trace records a larger *cell* enclosing p. For simplicity, we assume that the cell is a square of dimension x (in distance units). The imprecision is higher if x is higher, and vice versa.

There is an *adversary* who tries to identify the complete path histories of one or more participants (of known true identities) from the anonymous traces. We call a node whose whereabouts are being exposed a *victim* node. For the adversary to achieve its purpose, we assume that it can collect certain *side information* about one or more participants by chance or effort through noisy real-world channels. Each piece of side information gives the location of a participant at an associated time instant, although the information may not be exact. In practice, the side information may be obtained through a number of practical means. First, nodes are open to observations in public spaces. Hence, the adversary may obtain the side information *directly* through meeting the victim by chance or engineered encounters. Direct side information may be noisy due to imperfect vision or memory of the adversary about the meeting. Second, nodes may disclose information on their whereabouts either voluntarily or inadvertently. For example, a casual conversation between Alice and Bob may make references to where Alice was around 9 p.m. the night before, or it may make reference to the whereabouts of another person Charlie. Clearly, such location information might be released through many other means, including published media such as news articles or Web blogs. Hence, the adversary may also obtain the side information *indirectly*, i.e., through a channel other than direct encounter with the victim. Similarly, the indirect information may be noisy due to imprecise observations, memories, references, etc. In this paper, we will consider the following two attack scenarios.

A. Problem A: Passive Adversary

In this problem setting, the adversary is given the complete (anonymized) traces. The adversary's goal is, given some pieces of side information about a predetermined but unknown victim, to identify in some optimal fashion the complete path history of the chosen victim. The key assumptions are: 1) the adversary is *passive* in the sense that it does not actively go out to seek encounters with potential victims; 2) the side information given to the adversary contains noise. We will consider two cases. In the first case (*Problem A1*), the side information references time instants that coincide with sampled times in the trace only. That is, if a piece of side information refers to a participant's location at time t, then the set of traces must also contain a sampled location of some participant at t. In the second, more general case (*Problem A2*), the side information may also reference time instants between two consecutive sampled times in the set of traces. We study the worst-case scenario in which *all* pieces of the side information refer to times different from the sampled times in the set of traces. In either cases, we assume that the adversary is "sophisticated" and will attempt to incorporate all known information in its inference strategy by employing some form of Bayesian inferencing. We further assume that, in applying the Bayesian inferencing, the adversary can make use of some general knowledge it has about the world, including global constraints on nodal movements imposed by (publicly known) geography of the deployment area, and general movement preferences of all the nodes viewed as an aggregate (but not the individual preferences of specific nodes).

B. Problem B: Active Adversary

In this section, the adversary is *active* in the sense that it obtains side information about participants by physically encountering the participants. The complete trace history is still revealed to the adversary, but now in a real time and gradual fashion, i.e., as time progresses, the adversary is provided with the trace information together with the information acquired up to the real time instants. The goal here is to identify *as many identities of the traces as possible*. Specifically, we will consider the following three forms of the problem: B1) The adversary is itself one of the mobile nodes included in the set of traces (i.e., it is one of the participants in the trace set); B2) The adversary minimizes its efforts by simply staying at one fixed location; B3) The adversary predetermines a movement strategy to presumably maximize the amount of useful side information it can obtain, subject to the same physical movement constraints and speed limits as the participant mobile nodes. However, we will not consider the case in which the adversary may adapt its movement strategy to prior information it has learned about the potential victims. For example, after encountering a victim, the adversary will not attempt to henceforth follow the victim. This is reasonable if the objective of the adversary is to identify as many trace identities as possible. In fact without further given information, it is not clear if modifying the path can improve the performance.

The goal in all of the scenarios in the above two problems is to identify the victim's trace from the published set based on all the available (noisy) information. The results will be presented in the most quantitative manner possible.

C. Notations and Model Assumptions

We first define some notations and general assumptions about the *a priori* knowledge.

Θ : The collection of all cell location IDs.

$\{L_i\}_{i=1,2,...N}$: The collection of all the traces of the participants, each indexed by an anonymous index i. N is the

total number of traces. Precisely, for each i, L_i is a function of time $L_i : \mathbf{R}_+ \longrightarrow \Theta$ giving the ID of the cell visited by participant i.

$\{s_k\}_{k=1,2,\ldots}$: The sampled times at which the actual node locations are published, i.e., $L_i(s_k)$ is the published location ID of the cell visited by mobile node i at time s_k.

$\{t_k\}_{k=1,2,\ldots}$: The time instants at which some noisy side information about the victim's locations are revealed.

R : The noisy side information of the victim. Specifically, it is a map, $R : \{t_k\}_k \longrightarrow \Theta$ so that $R(t_k)$ is the (corrupted) location ID of the cell visited by the victim at time t_k as revealed to the adversary.

In order to concentrate on the key issue of privacy breach, we further make the following assumptions.

1) The sampled times s_k's are equally spaced. In addition, for *Problem A1*, we have $\{t_k : k = 1, 2, \ldots\} \subset \{s_k : k = 1, 2, \ldots\}$; For *Problem A2*, we have $\{t_k : k = 1, 2, \ldots\} \not\subset \{s_k : k = 1, 2, \ldots\}$; then we assume that for each t_k, there exists $\tilde{k}$ such that $s_{\tilde{k}} < t_k < s_{\tilde{k}+1}$ and $t_k = \frac{1}{2}(s_{\tilde{k}} + s_{\tilde{k}+1})$.

2) The noise in the side information in each revelation instant is assumed to be some i.i.d. random variable Z_k's of some given distribution $\Pr_Z$. Hence, we have

$$R(t_k) = L_{i^*}(t_k) + Z_k \tag{1}$$

where i^* is the victim's trace ID (which is of course not known to the adversary).

3) All the mobile nodes follow the same movement model that is assumed to be Markovian. Hence, the statistics of the whole collection of traces can be completely described by some one-step transition matrix $\{P_{ij}\}_{i,j\in\Theta}$. The time interval for the transition matrix is denoted by T. For the convenience of later presentation, we set T to be $s_2 - s_1$ for *Problem A1* and $\frac{1}{2}(s_2 - s_1)$ for *Problem A2*. This matrix is either given or estimated by some general world knowledge.

We take the time here to note that the last assumption is clearly for simplification purposes. There are many well-known prediction, interpolation, and filtering algorithms for (even non-Markovian) time series analysis (see, for example, [12, Ch. 3 and 8]). On the other hand, our simulation results already produce robust results even for the non-Markovian real traces. Hence, we will not be sidetracked by invoking the more refined models. Instead, we will emphasize the implications of general knowledge about nodal movements toward the privacy issues.

IV. Strategies of the Adversary

In this section, we give details of the possible strategies used by the adversary for each of the attack scenarios listed in Section III.

A. Strategies for A1 and A2

As noted before, the side information often contains noise. The adversary thus needs to perform Bayesian inference or use the maximum likelihood estimator (MLE) to make the best guess. The goal is that given R, find the L_i that gives the best match. The formulation of such a procedure is described in the following. Given $\{R(t_k)_{k=1,2,\ldots}\}$, compute

$$\begin{aligned}
&\Pr(L_i|\{R(t_k), k = 1, 2, \ldots\}) \\
&= \frac{\Pr(L_i, R(t_k), k = 1, 2, \ldots)}{\Pr(R(t_k), k = 1, 2, \ldots)} \\
&= \frac{\Pr(R(t_k), k = 1, 2, \ldots|L_i)\Pr(L_i)}{\sum_{j=1}^{N}\Pr(R(t_k), k = 1, 2, \ldots|L_j)\Pr(L_j)}.
\end{aligned} \tag{2}$$

The goal of the MLE is to find i which maximizes the expression (2). Note that the denominator does not depend on i. In addition, without any knowledge about how the victim is chosen, we set the *a priori* distribution of the victim to be uniform: $P(L_i) = \frac{1}{N}$ for $i = 1, 2, \ldots N$. Hence, the solution of the MLE is given by

$$\max_{i=1,2,\ldots N} \Pr(R(t_k), k = 1, 2, \ldots|L_i). \tag{3}$$

With the assumption of the noise model given in (1), the expression (3) can be given in the following form.

Case A1: Because the noise is i.i.d., we have

$$\Pr(R(t_k), k=1, 2, \ldots|L_i) = \prod_k \Pr_Z(R(t_k) - L_i(t_k)) \tag{4}$$

where the location difference is computed using the Cartesian distance between the two cells. Recall that $R(t_k) - L_i(t_k)$ equals the noise random variable in the perturbation process give by (1).

Case A2: By the Markovian assumption of the movement model, (3) can be given by

$$\begin{aligned}
&\Pr(R(t_k), k = 1, 2, \ldots|L_i) \\
&\quad = \Pr(R(t_k), k = 1, 2, \ldots|L_i(s_k), i = 1, 2, \ldots) \\
&\quad = \frac{\prod_k[\Pr(L_i(s_{\tilde{k}+1})|R(t_k)) \times \Pr(R(t_k)|L_i(s_{\tilde{k}}))]}{\prod_k[\Pr(L_i(s_{\tilde{k}+1})|L_i(s_{\tilde{k}}))]}.
\end{aligned} \tag{5}$$

Recall that there exists a $\tilde{k}$ such that $s_{\tilde{k}} < t_k < s_{\tilde{k}+1}$ and $t_k = \frac{1}{2}(s_{\tilde{k}} + s_{\tilde{k}+1})$. Hence, (5) can be easily expressed in terms of the transition matrix P_{ij}: The numerator involves transitions between time intervals of length T and hence the matrix P, while the denominator involves intervals of length $2T$ and hence the matrix P^2.

The expression (4) can be greatly simplified if the noise Z_k takes on specific forms. For example, we have the following.

1) Gaussian random variables $N(0, \sigma^2)$:

$$\begin{aligned}
&\Pr(R(t_k), k = 1, 2, \ldots|L_i) \\
&\qquad = C\exp\left\{-\frac{1}{2\sigma^2}\sum_k |R(t_k) - L_i(t_k)|^2\right\}
\end{aligned} \tag{6}$$

for some constant C. Hence, the MLE is essentially the same as the following *minimum square* approach:

$$\min_i \sum_k |R(t_k) - L_i(t_k)|^2. \tag{7}$$

2) Uniform Distribution with on the interval $(-\frac{l}{2}, \frac{l}{2})$:

$$\Pr(R(t_k), k = 1, 2, \ldots|L_i) = \prod_k \frac{1}{l}\chi_{(-\frac{l}{2},\frac{l}{2})}(R(t_k) - L_i(t_k)) \tag{8}$$

where $\chi_A(x, y) = 1$ or 0 depending on if $x - y \in A$ or not. Upon taking the log of the above equation, we have

$$\begin{aligned}
&\log\Pr(R(t_k), k = 1, 2, \ldots|L_i) \\
&\quad = \sum_k \log\chi_{(-\frac{l}{2},\frac{l}{2})}(R(t_k) - L_i(t_k)) + (\text{a constant}).
\end{aligned} \tag{9}$$

Optimizing the above expression is equivalent to identifying the trace that has the *largest number* of sampled times such that the trace location falls within a fixed range of the noisy side information.

The above provides a rigorous mathematical formulation for the Bayesian inferencing equipped with the side information. On the other hand, the above also leads to some simplified heuristic approaches for tackling the victim identification problem. Qualitatively, they are all similar to the minimum square approach, but we find it a useful contribution to record and compare their performances. In the following, we consider four strategies used by the adversary to identify the victim's trace from the published trace set. We first describe them for case A1.

1) *MLE Approach (MLE):* This is the same as formulation (4), i.e., the *similarity* value of trace i is given by $\prod_k \Pr_Z(R(t_k) - L_i(t_k))$. The trace with the *maximum similarity value* is declared to be the victim's.

2) *Minimum Square Approach (MSQ):* This is essentially formulation (7), i.e., the *similarity* value of trace i is given by $-\sum_k |R(t_k) - L_i(t_k)|^2$. The trace with the *least negative similarity value* is declared to be the victim's.

3) *Basic Approach (BAS):* In this approach, motivated by the uniform noise distribution (8) and (9), but to allow more flexibility, the adversary assumes that the noise is zero-mean and has a specific standard deviation (σ), but makes no assumption about its exact distribution. The adversary then computes the *similarity* value of trace i with the collected side information using the following equation:

$$\sum_{k=1}^{M} I_{2\sigma}(R(t_k), L_i(t_k)) \tag{10}$$

where $I_{2\sigma}(x, y) = 1$ if $|x - y| \leq 2\sigma$, and 0 otherwise. Hence, the adversary accepts a trace as a potential candidate if it is possible for the trace owner to appear in a radius of $2 \times \sigma$ of the revealed location, which encloses all possible noise if it is uniformly distributed, or 95.6% of noise if it is Gaussian. The trace with the *maximum similarity value* is declared to be the victim's.

4) *Weighted Exponential Approach (EXP):* In this approach, which is proposed and analyzed in [26], we assume that the adversary does not know the type of noise or its magnitude. Similar to BAS, the adversary computes and maximizes the *similarity* value of trace i using the following equation:

$$\sum_{k=1}^{M} \frac{1}{\text{Weight}(R(t_k))} \exp\left\{-\frac{1}{C}|R(t_k) - L_i(t_k)|\right\} \tag{11}$$

where $\text{Weight}(R(t_k))$ is some weight assigned to the revealed cell $R(t_k)$ and C is a constant. This formulation describes a similar concept as in the BAS approach, but one that is not as drastic. The exponential function assigns a higher weight when the trace location is closer to the side information, but the weight decays to zero more slowly than the abrupt vanishing property of the characteristic function in the BAS formula. In the simulations, we let the weights in the denominator be equal because, with possible errors in the revealed location, it is unclear how different weights could be assigned.

The above formula can be easily modified for case A2. For convenience, we first define for each trace i, the function $P_i : \Theta \times \{t_k : k = 1, 2, \ldots M\} \longrightarrow \mathbf{R}_+$

$$P_i(l, t_k) = \frac{P_{x,l} P_{l,y}}{P_{x,y}}$$

where $x = L_i(s_{\tilde{k}})$, $y = L_i(s_{\tilde{k}+1})$, and $s_{\tilde{k}} < t_k < s_{\tilde{k}+1}$. Then, we have the following:

MLE$_2$:

$$\prod_k \left(\sum_{l \in \Theta} P_i(l, t_k) \Pr_Z(R(t_k) - l) \right). \tag{4$_2$}$$

MSQ$_2$:

$$-\sum_k \left(\sum_{l \in \Theta} P_i(l, t_k) |R(t_k) - l|^2 \right). \tag{7$_2$}$$

BAS$_2$:

$$\sum_{k=1}^{M} \left(\sum_{l \in \Theta} P_i(l, t_k) \times I_{2\sigma}(R(t_k), l) \right). \tag{10$_2$}$$

EXP$_2$:

$$\sum_{k=1}^{M} \left(\sum_{l \in \Theta} \frac{P_i(l, t_k)}{\text{Weight}(R(t_k))} \exp\left\{-\frac{1}{C}|R(t_k) - l|\right\} \right). \tag{11$_2$}$$

Notice that the four approaches have the same computational complexity, which is linear in the number of pieces of revealed side information and the number of nodes.

A remark in place is that our exposition assumes attack strategies where the victim is assumed to be one of the participants. However, the strategies apply or can be easily extended to the case in which it is uncertain if the side information collected for a mobile node actually corresponds to any participant. In particular, the MLE approach can be used directly without modification, while a properly picked *threshold* can be used for the other attack strategies to remove traces from consideration if their similarity to the victim's trace is lower than the threshold. This can certainly be formulated rigorously in terms of statistical hypothesis testing.

B. Strategy for Problems B1–B3

In this scenario the adversary observes the participants directly. Note that the information about the traces is only revealed progressively in time, in a synchronized way with respect to the information collected by the adversary. The overall algorithm is specified in Fig. 1. As there is no noise when additional information is acquired, the adversary does not need to use any inference strategy. The Attack program takes as input the traces that are published progressively. The algorithm first assumes that all the traces are candidate traces for each participant. A trace is said to be a candidate trace of a participant if it appears at the same set of times and locations as when/where the adversary meets the participant, and the trace has not yet been identified. As time evolves, the adversary removes candidate traces that do not agree with the observed information about each victim from

```
Cascade(candidate_set, i)
  let j = trace id where candidate_set_i = {j}
  /* remove the identified trace from candidate set
     of other victims */
  For(m = 0; m < number_of_trace; m++)
    If trace j in candidate_set_m and m ≠ i
      remove trace j from candidate_set_m
      If candidate_set_size_m = 1
        Cascade(candidate_set, m)
      Endif
    Endif
  Endfor

Attack({L_i}_{i=1,2,...N})
  /* initially all traces are possible candidates
     to each victim */
  For (m = 0; m < number_of_trace; m++)
    add all traces to candidate_set_m
  Endfor

  While (sampling_time not ended)
    For each node i met at sampling_time and
        each trace j in candidate_set_i
      /* check if a candidate trace appear at the
         observed location */
      If (met node i at location r at sampling_time and
          L_j(sampling_time) != r)
        remove trace j from candidate_set_i
        If candidate_set_size_i = 1
          Cascade(candidate_set, i)
        Endif
      Endif
    Endfor

    report average k-anonymity
    evolve sampling_time
  Endwhile

  report all identified victims
```

Fig. 1. Specification of Attack algorithm.

the set for that victim. The function Cascade takes two input parameters, where `candidate_set` is the candidate set of all victim nodes and i is the victim ID identified. The function is called when a victim's trace is identified, so as to remove that trace from the candidate set of other victims. The candidate set size is the k-anonymity of the victim, as every trace in the candidate set is possibly the victim's.

Notice that the adversary may not identify a participant at times they meet each other, but the identification can occur at a later time when all but one candidate traces are identified and removed, as indicated by the recursive Cascade function call in Fig. 1. Hence, the adversary identifies a participant more efficiently when it tries to identify as many participants as possible.

V. Trace Characteristics

In this section, we begin by analyzing the differences in behaviors between the real traces. Their differences will be illustrated by three types of real mobility traces: 1) cabs in San Francisco (cab) [31]; 2) buses in a Shanghai grid system (bus) [34]; and 3) cabs in the Shanghai area (shcab) [34]. Basic statistics of these three sets of traces are listed in Table I. We assume that the published traces are snapshots taken every minute with spatial granularity of 0.01° in latitude and longitude for anonymization purpose as explained in Section I unless stated otherwise. Characteristics of traces are studied using the four metrics as described in Sections V-A–V-D. Observations that can be explained using differences between movement preferences of the mobile nodes are summarized at the end of this section.

A. Distribution of Correlation Between Traces

Here, we study the correlation between different traces. We use the Pearson product-moment correlation coefficient to

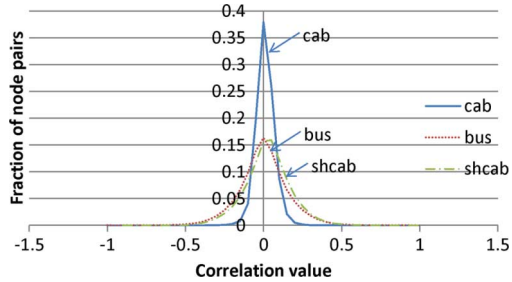


Fig. 2. Distribution of correlations between traces of the same set.

TABLE I
Basic Statistics of the Real Traces

	San Francisco cabs	Shanghai Grid buses	Shanghai cabs
Min. latitude	37.05	30.7217	30.00
Max. latitude	38.00	31.5899	32.00
Min. longitude	-122.86	121.0001	120.00
Max. longitude	-122.00	121.9117	122.00
# cells[a]	8170	8004	40000
# active cells[b]	3997	2108	19746
# nodes	536	2348	4438
Min. timestamp (local time)	Sat May 17 03:00:04 2008	Mon Feb 19 08:00:01 2007	Wed Jan 31 13:00:01 2007
Max. timestamp (local time)	Tue June 10 02:25:34 2008	Sat Feb 24 08:00:00 2007	Sat Feb 24 13:00:00 2007

[a] when spatial granularity is 0.01°.
[b] cells ever visited by any node.

quantify the correlations between node pairs, which is used in the study in the relationship between taxonomy of texts [38, Appendix]. It is also related to the cross-correlation function between stochastic processes [30, Ch. 10]. For any mobile node pair i and j, the quantity is defined as follows:

$$C(i,j) = \lim_{M\to\infty} \frac{1}{M}\sum_{k=1}^{M}\left(\frac{L_i(s_k) - EL_i}{\sigma_{L_i}}\right)\left(\frac{L_j(s_k) - EL_j}{\sigma_{L_j}}\right)$$

where EL_i and σ_{L_i} are respectively the average and standard deviation of node i's locations

$$EL_i = \lim_{M\to\infty} \frac{1}{N}\sum_{k=1}^{M} L_i(s_k) \tag{12}$$

$$\sigma_{L_i} = \lim_{M\to\infty} \sqrt{\frac{1}{M}\sum_{k=1}^{M}(L_i(s_k) - EL_i)^2}. \tag{13}$$

The distribution of the correlations between different node pairs is depicted in Fig. 2.

The figure shows that movements of different San Francisco cabs have little or no correlation. It is because cabs are unlikely to follow each other for a long time. Moreover, the Shanghai cabs have higher correlation than the San Francisco cabs. Investigation reveals that some of the Shanghai cabs did not move at all over the trace collection period, and their positions are indifferentiable from each other under the spatial granularity of the cloaking. This is possibly because they are parked close to each other, and their identical cloaked locations lead to the high correlation.

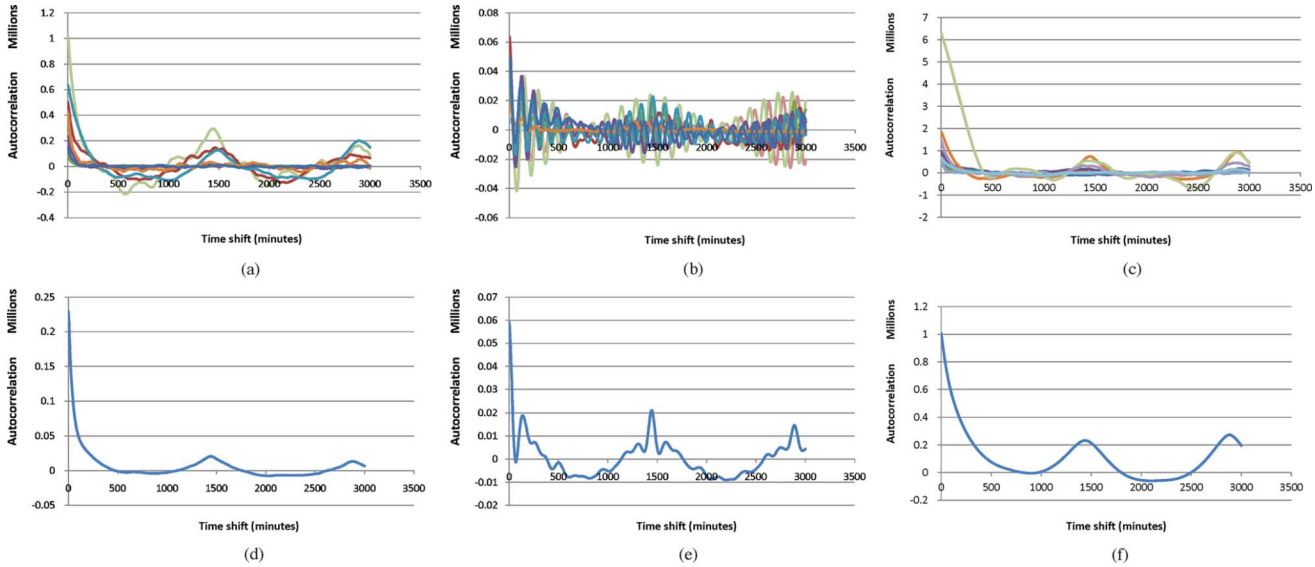


Fig. 3. Autocorrelation of each trace for different sets of traces as a function of the time shift s. (a) Samples for cab traces. (b) Samples for bus traces. (c) Samples for SH cab traces. (d) Average for cab traces. (e) Average for bus traces. (f) Average for SH cab traces.

B. Autocorrelation of the Same Trace

The autocorrelation $C(i,s)$ of trace i with time shifting of s is defined as

$$\lim_{M\to\infty}\frac{1}{M}\sum_{k=1}^{M}(L_i(s_k+s)-EL_i)(L_i(s_k)-EL_i).$$

In the case of a (stationary) Markov chain $\{L(n)\}_{n=0,1,2,\ldots}$ with transition matrix p_{ij}, the above value can be explicitly computed as

$$E(L(n)-\bar{L})(L(n+s)-\bar{L}) = E(L(n)L(n+s)) - (\bar{L})^2$$
$$= \sum_{k,l\in\Theta} klp_{kl}^{(s)}\mu_k - \left(\sum_k k\pi_k\right)^2$$

where $p^{(s)}$ is the sth-step transition matrix and μ is the stationary distribution of the Markov chain.

The (individual and average) results for the various traces are shown in Fig. 3 as a function of the time shift s. The figure shows that for the real traces, there are sharp rises in autocorrelation individually and on average when the time shift is one day. The bus traces also show repeatedly oscillating autocorrelation values throughout a day because each bus runs on a periodic schedule. Such oscillations are much less obvious for the cabs as they move more randomly.

C. Complexity of Movement

In this section, we demonstrate the complexity of nodal movements as quantified through the order-n model complexity given by

$$H_n(L) := -H(L(n)|L(1),L(2),\ldots L(n-1))$$
$$= -\sum_{\Theta^{n-1}} p(l_1,\ldots,l_{n-1})$$
$$\times\left\{\sum_{\Theta} p(l_n|l_1,\ldots,l_{n-1})\log p(l_n|l_1,\ldots,l_{n-1})\right\}. \quad (14)$$

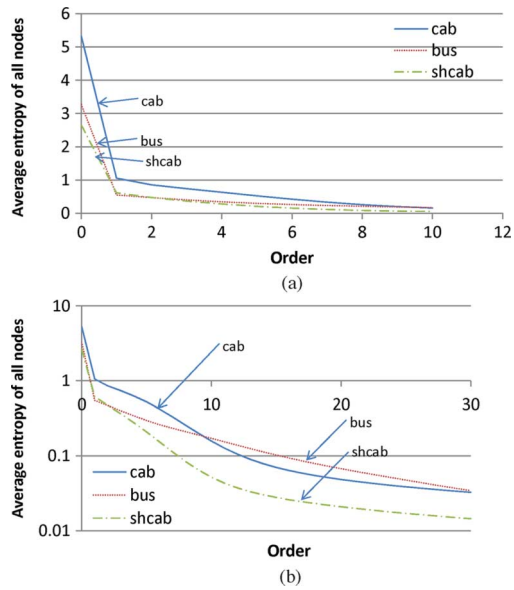


Fig. 4. Order-n complexity of different sets of traces as a function of order. (a) Zoomed. (b) Log scale.

In (14), L denotes a general outcome of L_i's and the functions $p(\cdots):\Theta^{n-1}\longrightarrow\mathbf{R}_+$ and $p(\cdot|\cdots):\Theta\times\Theta^{n-1}\longrightarrow\mathbf{R}_+$ are the joint probability and conditional probability densities, respectively, of the locations in the collection of traces $\{L_i\}_{i=1,2,\ldots N}$.

The above function is defined for general stochastic processes (see for example [6, Ch. 3]). The value of H_n represents the uncertainty of the order-n model. The smaller the value, the less uncertainty there is in the model. Notice that H_0 is essentially the entropy of the stationary distribution.

The behavior of (14) as a function of n is shown in Fig. 4. The result conforms to the theoretical result that for any stationary process X, $H_n(X)$ is a decreasing function of n, and the limit $\lim_{n\to\infty}H_n(X)$ thus exists. The limiting value is called the *entropy rate* of the process X, usually denoted by $H(L)$.

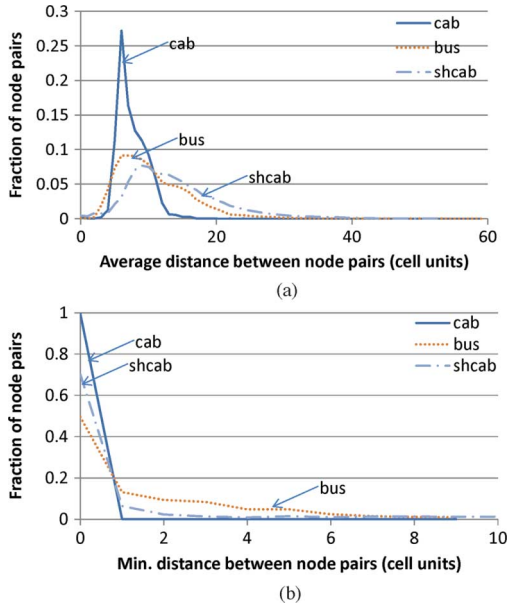


Fig. 5. Distribution of (a) average and (b) minimum distances between pairs of traces.

Again, in the case of a Markov chain $\{L(n)\}_{n=0,1,2,\ldots}$, $H(L)$ can be explicitly given by

$$H(L) = -\sum_{kl} \pi_k p_{kl} \log p_{kl}.$$

D. Distribution of Distances Between Traces

Fig. 5(a) shows the distribution of average distance between trace pairs, which is defined as

$$\mathrm{Dist}_{\mathrm{Ave}}(i,j) = \lim_{N\to\infty} \frac{1}{N} \sum_{k=1}^{N} |L_i(k) - L_j(k)|$$

for trace pair i and j, where $i \neq j$, and Fig. 5(b) depicts the distribution of minimum distance between trace pairs, which is defined as

$$\mathrm{Dist}_{\mathrm{Min}}(i,j) = \min_k (|L_i(s_k) - L_j(s_k)|).$$

E. Implications of the Trace Characteristics

Many of the observed different characteristics of the mobility traces can be summarized and explained using the natural preferences for certain places visited by the mobile nodes as shown in the real traces, such as the busy downtown area for the cabs and the assigned routes for the buses.

For instance, the real traces have their preferred visiting places, resulting in a small H_0, and the entropy drops slowly when the order increases. This is also reflected in the small average distances between cabs or between buses. Moreover, the cab and bus traces collected from Shanghai exhibit a broader spatial range than the cab traces from San Francisco, which is likely because the size of Shanghai is larger than that of San Francisco. At the same time, the popular places in Shanghai are distributed more sparsely than those in San Francisco, such that the Shanghai cabs have a larger average distance from each other than the San Francisco cabs. Because some of the bus routes are closer together while some are farther apart, the result is that only 50% of the buses have met each other as shown in Fig. 5(b), while almost 100% of the San Francisco cabs and 70% of the Shanghai cabs have met each other.

When nodes are more sparsely distributed in the network area, more efficient victim identification results when the adversary collects the side information passively (*Problems A1* and *A2*). On the other hand, sparsity of nodes can both be beneficial and detrimental to the performance of an adversary who observes the participants directly (*Problems B1–B3*). It is because when the mobile nodes are sparsely distributed, it could take much longer time for the adversary to meet them, thus harming the attack efficiency. Meanwhile, once the adversary meets a mobile node, it could identify the trace of the node almost instantaneously as no other mobile nodes (and hence traces) are around at the same time, thus helping the attack performance. We will verify these expectations experimentally in Section VI.

VI. Simulation Results

A. Results for Passive Adversary

In this section, we study the attack scenario where the adversary tries to identify the trace of one participant (the victim) by gathering side information passively. In each simulation, the victim is randomly picked from all the participants. Pairs of $\langle \mathrm{time}, \mathrm{location} \rangle$ of the victim are then randomly sampled from the trace, and noise is introduced in the spatial domain. The noisy data are revealed to the adversary as side information, which the adversary utilizes to identify the complete movement history of the victim from the published traces. We assume that the published traces are snapshots taken every minute with spatial granularity of 0.01° in latitude and longitude for anonymization purpose unless stated otherwise. Results reported are for simulation experiments each repeated 100 000 times.

We quantify the performance of the strategies with the following metrics.

1) *Fraction of correct conclusions:* A conclusion is correct if the victim is uniquely identified according to the criterion of highest similarity metric, or the victim is among the set of candidates with the highest similarity metric and all the candidates are indifferentiable from each other.
2) *Fraction of incorrect conclusions:* A conclusion is incorrect when the victim is not among the set of candidates having the highest similarity metric.
3) *Fraction of undecided conclusions:* A conclusion is undecided when the victim is among the set of candidates having the highest similarity metric and the candidates are not indifferentiable from each other.

1) Problem A1: We present the results based on the perception of the passive adversary on the noise when the side information references time instants that coincides with sampled times in the traces.

1) *Correct assumption about the noise distribution:* We first consider the case when the revealed location of the victim is perturbed with zero-mean Gaussian noise with standard deviation σ, which matches the assumption made by the adversary in MLE. Fig. 6 shows the performance of the attack strategies using the cab, bus, and shcab traces.

When we compare the two attack strategies that assume knowledge of the noise, namely MLE and BAS, MLE is more aggressive as it excludes a trace from further consideration as

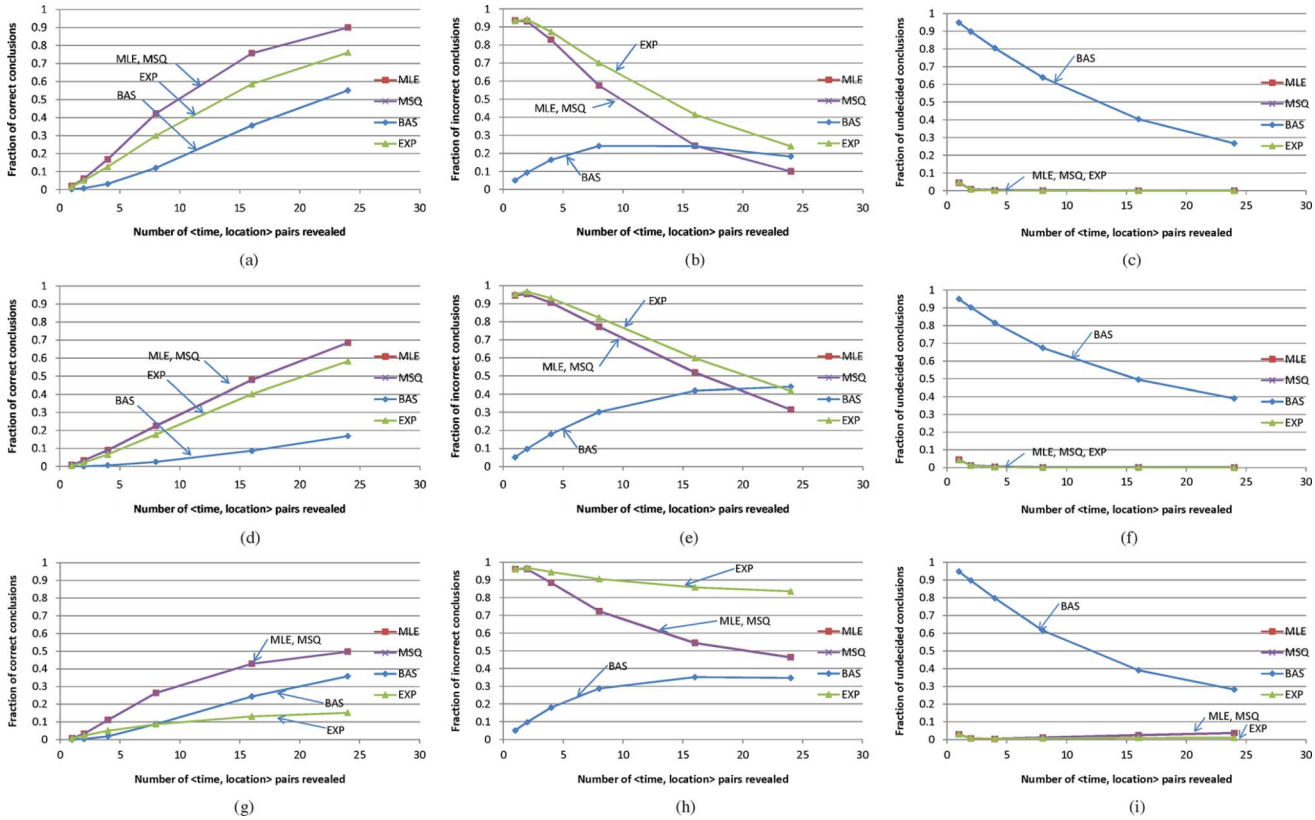


Fig. 6. (Problem A1) Performance of various metrics as a function of the number of $\langle \text{location}, \text{time} \rangle$ pairs revealed. (a)–(c) San Francisco cab traces, (d)–(f) Shanghai Grid bus traces, and (g)–(i) Shanghai cab traces. Zero-mean Gaussian noise with $\sigma = 5$. (a) Correct conclusions, cabs. (b) Incorrect conclusions, cabs. (c) Undecided conclusions, cabs. (d) Correct conclusions, buses. (e) Incorrect conclusions, buses. (f) Undecided conclusions, buses. (g) Correct conclusions, Shanghai cabs. (h) Incorrect conclusions, Shanghai cabs. (i) Undecided conclusions, Shanghai cabs.

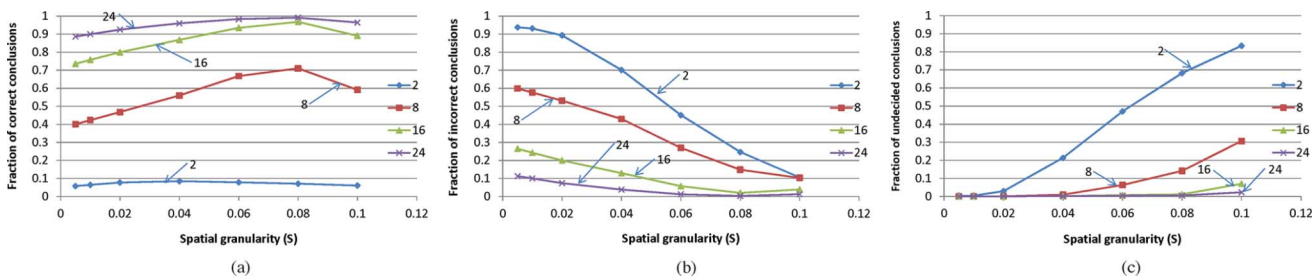


Fig. 7. (Problem A1) Performance of various metrics as a function of the spatial granularity of the trace set, for different numbers of $\langle \text{location}, \text{time} \rangle$ pairs of side information. San Francisco cab traces. MLE attack strategy. Zero-mean Gaussian noise with $\sigma = 5$. (a) Correct conclusions. (b) Incorrect conclusions. (c) Undecided conclusions.

soon as it determines that the trace cannot be perturbed to the revealed locations of the victim given the type and magnitude of the noise assumed. Hence, when the adversary's assumption is correct, this approach can give very good results in the fraction of correct conclusions, although it can also give a large fraction of incorrect conclusions initially, when the adversary has only a few pairs of the side information because the traces with the highest similarity for only a few pieces of noisy side information may not be truly the victim's. In comparison, BAS generally returns lower fractions of both correct and incorrect conclusions as it gives equal weights to traces that agree with the side information within the error bounds. This results in more undecided conclusions, i.e., there is more than one trace, including the correct one, which shares the same highest similarity value, and the victim's trace is undecided among the set. Notice that because the error bounds are not large enough to enclose all possible noise, the fraction of incorrect conclusions increases initially for BAS when more pieces of side information are available to the adversary.

We now look at the other two approaches that do not use knowledge of the noise, namely MSQ and EXP. We can see that although MSQ does not require the knowledge, its performance is similar to the best-case performance of MLE in terms of the fraction of correct conclusions. Meanwhile, EXP performs the worst as it puts too much weight on traces that give little deviations from some of the pieces of side information.

Next, we evaluate the impact of the granularity of spatial cloaking. Results of the MLE attack on San Francisco cab traces are shown in Fig. 7. In the experiments, the adversary's side information is inaccurate or noisy. We fix the amount of noise in the side information and vary the spatial granularity of the published trace set.

If the adversary's side information is accurate, we expect that a finer spatial granularity of the traces will increase the

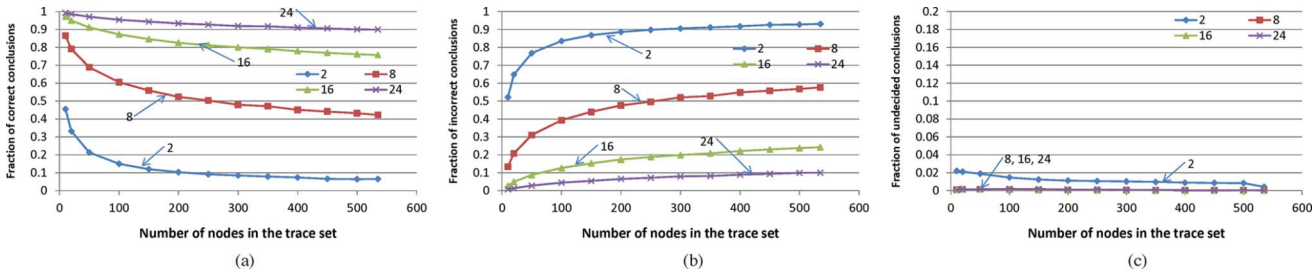


Fig. 8. (Problem A1) Performance of various metrics as a function of the number of nodes in the published trace set, for different numbers of $\langle \text{location}, \text{time} \rangle$ pairs of side information. San Francisco cab traces. MLE attack strategy. Zero-mean Gaussian noise with $\sigma = 5$. (a) Correct conclusions. (b) Incorrect conclusions. (c) Undecided conclusions.

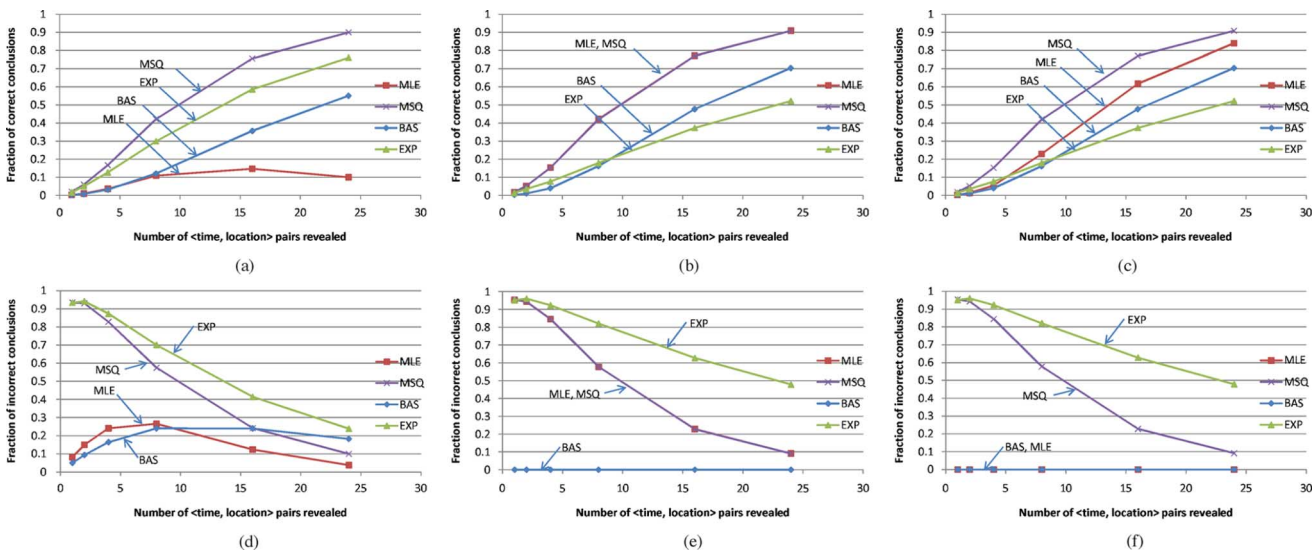


Fig. 9. (Problem A1) Performance of various metrics as a function of number of $\langle \text{location}, \text{time} \rangle$ pairs revealed. (a), (d) Uniform noise assumed, Gaussian actual. (b), (e) Gaussian noise assumed, Uniform actual. (c), (f) Uniform noise both assumed and actual. San Francisco cab traces. Noise with $\sigma = 5$. (a)–(c) Correct conclusions. (d)–(f) Incorrect conclusions.

effectiveness of the attack. This is because, when the grid cell is larger, more traces are likely to share a common cell, which makes it harder for the adversary to differentiate between the traces.

When the side information is inaccurate, however, the results in Fig. 7 show that interestingly, a coarser spatial granularity is not always bad for the adversary. This is because when the side information has mistakes, these mistakes may in fact be mitigated, or sometimes even masked, by a coarser cell structure. There is hence a competitive effect between the higher discriminative power of finer cells on the one hand, and possible error mitigating effects of coarser cells on the other. The result is that, as shown in Fig. 7, the adversary generally has the best performance at an *intermediate* cell size.

We now evaluate the impact of the number of nodes in the trace set. In this set of experiments, we always include the victim node in the trace set, but vary the number of other nodes that coexist with the victim. Fig. 8 shows the performance of the adversary using MLE as a function of the total number of nodes in the trace set. The results show that, as expected, the adversary has a higher ability of identifying the victim when the size of the trace set gets smaller. Furthermore, the adversary derives more benefit from a smaller number of nodes when she possesses fewer pieces of the side information.

2) *Incorrect assumptions about the noise distribution:* We now consider the case when the assumption of noise distribution made by the adversary in MLE is incorrect. Fig. 9(a) and (d) shows the performance of the strategy when the actual and assumed noise is Gaussian and Uniform, respectively. Fig. 9(b) and (e) shows the results when the actual and assumed noise is Uniform and Gaussian, respectively. Fig. 9(c) and (f) shows the results when the noise distribution is Uniform, and the adversary assumes the same.

Notice that among the approaches that assume about the noise, MLE is affected the most by the wrong assumptions. In particular, the performance of MLE varies depending on the types of actual and assumed noise. When the adversary assumes the noise to be Uniform but it is Gaussian, the performance is much worsened since the victim's trace can be mistakenly and permanently removed from consideration due to occasional Gaussian noise that exceeds the range of the assumed Uniform noise. On the other hand, when Gaussian noise is assumed but it is actually Uniform, MLE surprisingly gives a greater fraction of correct conclusions than when the correct noise distribution is assumed, albeit at the price of getting a greater fraction of incorrect conclusions also. In contrast to MLE, the performance of BAS is less sensitive to the type of noise.

2) Problem A2: Fig. 10 shows the performance of the attack approaches for different sampling time intervals for the cab traces when the side information references time instants that does not coincide with sampled times in the traces. Zero-mean Gaussian noise with $\sigma = 5$ is introduced into the spatial domain

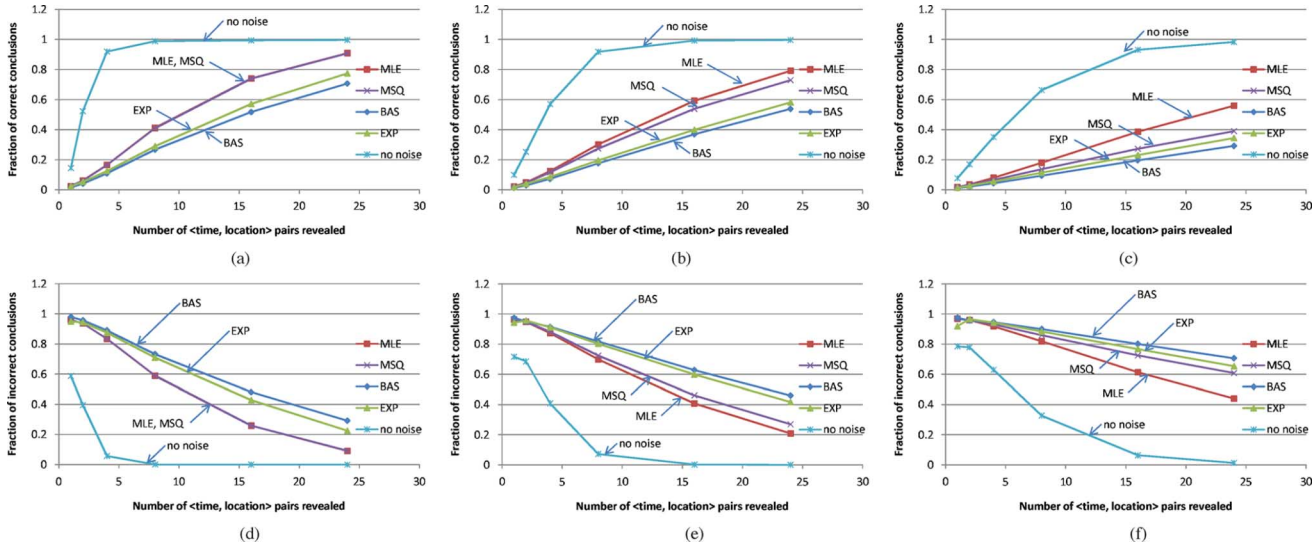


Fig. 10. (Problem A2) Performance of various metrics for attacks requiring different degrees of movement inference for each trace as a function of number of $\langle \text{time}, \text{location} \rangle$ pairs revealed. San Francisco cab traces, zero-mean Gaussian noise with $\sigma = 5$. (a), (d) $S = 10$ min; $T = 5$ min; (b), (e) $S = 30$ min; $T = 15$ min; (c), (f) $S = 1$ h; $T =$ half an hour. (S is the trace sampling time, and T is the interval for computing the transition matrix.) (a) Correct conclusions, 10-min sampling, (b) Correct conclusions, 30-min sampling, (c) Correct conclusions, 60-min sampling, (d) Incorrect conclusions, 10-min sampling, (e) Incorrect conclusions, 30-min sampling, (f) Incorrect conclusions, 60-min sampling.

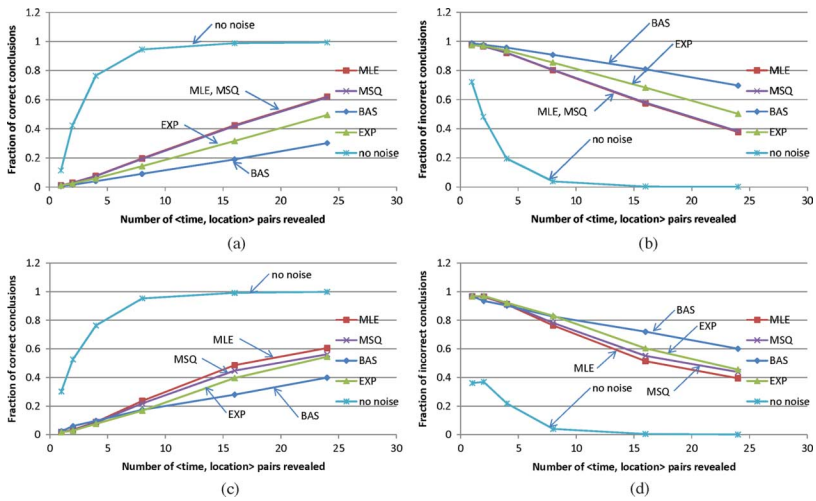


Fig. 11. (Problem A2) Performance of various metrics for attacks requiring different degrees of movement inference for each trace as a function of number of $\langle \text{time}, \text{location} \rangle$ pairs revealed. Traces are sampled every half an hour, and the transition matrix is generated using sampling information every 15 min. (a) Correct conclusions, buses. (b) Incorrect conclusions, buses. (c) Correct conclusions, Shanghai cabs. (d) Incorrect conclusions, Shanghai cabs.

of the side information except for the line labeled "no noise." The figure shows that the sparser the samples in the traces, the less effective the attacks are in general. This is expected since when samples are sparser, inference of nodal movements between the sampling points becomes less reliable. Fig. 11 shows the results for the bus traces and the Shanghai cab traces. The figure shows that without noise in the side information, even with a sampling temporal granularity of an hour and spatial granularity of $0.01°$, the adversary is able to identify the victim's trace by fewer than 25 pairs of side information with high probability. When noise is introduced, however, the results depend heavily on the traces. For instance, the effect of noisy side information on the attack strategies is more noticeable for both the Shanghai bus and cab traces.

When we compare the performance of the attack approaches in this case with the special case in Section VI-A.1, in which no inference using a general movement model is necessary, the performance here does not degrade significantly for MLE_2 and MSQ_2. Interestingly, BAS_2 gives a much larger fraction of incorrect conclusions and slightly larger fraction of correct conclusions initially when movement has to be inferred, while EXP_2 performs about the same in both cases.

3) Summary on Passive Adversary Strategies: The results show that approaches relying on the assumption of noise could have very poor performance when the assumption is wrong, as illustrated by the MLE results. On the other hand, an approach not having knowledge of the noise may still perform well. In particular, MSQ performs equally well as MLE even when the latter has the correct noise assumption. *Since MSQ also performs better than the heuristic approaches of BAS and EXP, it appears to be the preferred adversary strategy overall.*

B. Results for Active Adversary

In this section, we examine the performance of the active adversary who gains side information by direct meetings with the participants. Recall that this adversary can identify a victim by

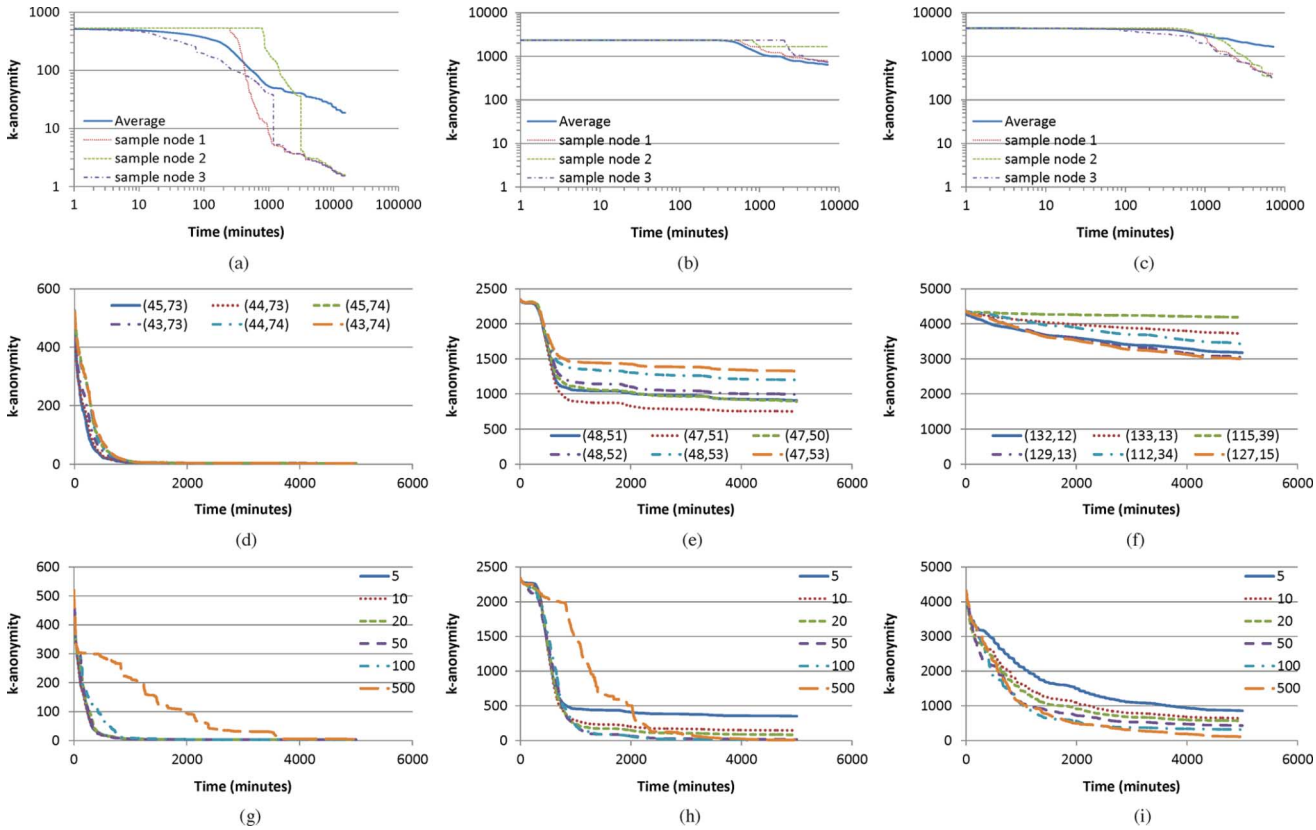


Fig. 12. k-anonymity of the victim as observed by the adversary as a function of attack time, when the adversary is (a)–(c) one of the mobile nodes (Problem B1), (d)–(f) static (Problem B2), and (g)–(i) mobile within a predetermined path (Problem B3). (a) One of mobile nodes (cab). (b) One of mobile nodes (bus). (c) One of mobile nodes (SH cab). (d) Static attacker (cab). (e) Static attacker (bus). (f) Static attacker (SH cab). (g) Mobile attacker (cab). (h) Mobile attacker (bus). (i) Mobile attacker (SH cab).

elimination, and the process is most efficient if the adversary meets the participants as quickly as possible. We assume that the adversary operates to achieve this goal. We further assume that the adversary's side information is gained only at times coinciding with sampled times of the traces.

1) Problem B1: Fig. 12(a)–(c) shows the average k-anonymity of the victims as observed by the adversary as a function of the attack time for different sets of the traces, when the adversary is one of the mobile nodes. The figures show that the most reduction in k-anonymity for each participant results from observations made in the first day in the real traces. Notice also from the figures that there are flat regions in the bus trace results corresponding to night times of the days. The cab traces exhibit a similar behavior, but it is much less obvious due to the cabs' own mobility characteristics.

2) Problem B2: Fig. 12(d)–(f) shows the k-anonymity of the victims as observed by the adversary as a function of attack time, when the adversary stays at one of the cells. Each line in the figure represents the results for a particular staying cell, and the line label shows the relative coordinates of that cell in the network area. We plot the results of the six most popular cells in each figure, and the popularity of a cell is ranked according to the total number of visits made by the mobile nodes over the entire trace.

The figures indicate that for the real traces, staying at a cell for a day is sufficient to reduce the k-anonymity for each participant significantly. The improvement by staying longer at each cell is minimal.

3) Problem B3: Fig. 12(g)–(i) shows the k-anonymity of the victim as observed by the adversary as a function of attack time, when the adversary moves actively inside the network area. The label of each line in the figure indicates the number of popular cells visited by the adversary. The adversary uses a greedy algorithm to compute the shortest route that connects all the popular cells to be visited and follows this heuristic route throughout the simulation period. Notice that as the adversary travels between the popular cells, it may visit other cells during the journeys.

The figures show that travels made by the adversary generally improve the attack efficiency in identifying the traces. For instance, for the bus traces, traveling helps the adversary reduce the size of the candidate set for each participant from more than 2000 to only a few in about one day, while staying at a cell can only reduce the size by half. It is because by traveling, the adversary is able to meet more participants, especially when their spatial distribution is sparser, such as the bus and cab traces from Shanghai. However, traveling to too many places may hurt the performance because the adversary may spend too much time traveling over unpopular places.

4) Summary on Active Adversary Strategies: In this section, we studied different strategies for an active adversary to collect snapshots of the victims. The results show that for the real traces, *the ability of the active adversary to travel helps it identify many of the victim traces in a realistic amount of time*. When the adversary prefers to stay at a cell, the attack efficiency depends on the type of traces and the staying location of the adversary. In general, staying at a more popular location helps by allowing the adversary to identify more victims more quickly.

VII. Discussion

Our analysis in this paper is motivated by the existing practice of releasing mobility traces in various public data portals [7], [34]. This practice is well intended. For example, the traces can be used to provide realistic input for trace-driven simulations, which can better ensure the relevance of the simulation results than synthetic traces. For these intended purposes, preservation of information at the granularity of individual traces is crucial. Furthermore, in order not to impose unnecessary constraints on using the traces for diverse types of investigations, these portals release entire trace sets of data to the user and leave it up to the user to exploit the available data. It is natural for us to inquire the privacy implications of such comprehensive release of information, and our analysis is a contribution to this investigation, beyond a general realization of the potential problem. Specifically, we provide a systematic study of the privacy problem in order to quantify its severity when exploited by an intelligent adversary, whose power is however limited by the amount of side information available to her.

Our analysis assumes techniques of spatial and temporal data cloaking that are admittedly basic. A main advantage of more basic cloaking techniques is, however, their ability to better guarantee the relevance of the cloaked data to diverse applications, including those that cannot be characterized *a priori*. Our chosen approach should not be taken as ruling out the use of more sophisticated cloaking techniques, however. These techniques are certainly possible, and many examples are known [1], [16], [29], [32], [39]. However, they all come at the price of requiring more severe transformations of the original data that will render the data applicable for specific applications only, i.e., types of applications for which the transformations are carefully designed.

In particular, differential privacy (DP) [8], [9] is a widely studied approach for ensuring the privacy of a data set in the face of a powerful adversary. In spite of its importance, however, DP makes assumptions about the use of a data set that are fundamentally incompatible with our problem context. For example, in order to provide strong privacy, DP does not give users unlimited access to the data set. Rather, the user must issue queries to learn about the data, and the types, as well as numbers, of allowable queries are carefully restricted, e.g., only *aggregate-sum* queries may be allowed [32]. Based on these restrictions, DP may calculate the amount of noise needed to ensure that no private information can be learned from adjacent data sets [9]. In particular, for time-series data such as our mobility traces, the amount of noise needed may in the worst case grow linearly with the number of queries [32]. Moreover, by nature of its design, DP is able to provide summary or statistical answers about a data set only, but it does not allow to preserve information at the granularity of the individual traces. As we remarked, the loss of per-trace information makes the data unsuitable for certain purposes including trace-driven simulations.

VIII. Conclusion

In this paper, we studied the privacy vulnerability of published mobility traces even when the true node identities are made anonymous and the recorded node positions may be imprecise. We presented comprehensive strategies for an adversary to well utilize side information about node movements, collected either passively or actively, to achieve different privacy attacks. We proved mathematically an optimal approach for the adversary to identify a victim's trace from the published data exploiting all the available information.

Our analysis is verified and complemented by simulation results under comprehensive system parameters, such as the nodal mobility, adversary strategy, noise in the trace or the side information, and different extents of movement inference needed for the attack. In general, our results showed that the adversary is able to identify victims with high probability even when the available side information is limited. Furthermore, for the passive adversary, attacks that make detailed noise assumptions, such as MLE, could have poor performance when the assumptions are wrong. On the other hand, MSQ does not rely on these assumptions, and its performance is robust. It performs as well as MLE even when the latter has the correct noise assumption. It also performs better than the heuristic approaches of BAS and EXP. Overall, MSQ appears to be the preferred passive adversary strategy. For the active adversary, we show that its ability to travel can help it to identify many of the victim traces in a realistic amount of time.

References

[1] O. Abul, F. Bonchi, and M. Nanni, "Never walk alone: Uncertainty for anonymity in moving objects database," in *Proc. IEEE ICDE*, Cancun, Mexico, Apr. 2008, pp. 376–385.

[2] R. Agrawal and R. Srikant, "Privacy-preserving data mining," *SIGMOD Rec.*, vol. 29, no. 2, pp. 439–450, 2000.

[3] R. J. Bayardo and R. Agrawal, "Data privacy through optimal k-anonymization," in *Proc. IEEE ICDE*, Tokyo, Japan, Apr. 2005, pp. 217–228.

[4] C.-Y. Chow, M. F. Mokbel, and X. Liu, "Spatial cloaking for anonymous location-based services in mobile peer-to-peer environments," *GeoInformatica*, vol. 15, no. 2, pp. 351–380, 2011.

[5] S. E. Coull, C. V. Wright, F. Monrose, M. P. Collins, and M. K. Reiter, "Playing devil's advocate: Inferring sensitive information from anonymized network traces," in *Proc. NDSS*, San Diego, CA, Feb. 2007, pp. 16–28.

[6] T. M. Cover and J. A. Thomas, *Elements of Information Theory*. New York: Wiley, 1991.

[7] Dartmouth College, Hanover, NH, "CRAWDAD," 2012 [Online]. Available: http://crawdad.cs.dartmouth.edu/

[8] C. Dwork, F. McSherry, K. Nissim, and A. Smith, "Calibrating noise to sensitivity in private data analysis," in *Proc. TCC*, Mar. 2006, pp. 265–284.

[9] C. Dwork, "Differential privacy, a survey of results," in *Proc. TAMC*, Xi'an, China, Apr. 2008, pp. 1–19.

[10] J. Eriksson, L. Girod, B. Hull, R. Newton, S. Madden, and H. Balakrishnan, "The pothole patrol: Using a mobile sensor network for road surface monitoring," in *Proc. ACM MobiSys*, Breckenridge, CO, Jun. 2008, pp. 29–39.

[11] J. Freudiger, R. Shokri, and J.-P. Hubaux, "Evaluating the privacy risk of location-based services," in *Proc. Finan. Cryptogr. Data Security (FC)*, Feb. 2011, pp. 31–46.

[12] B. Fristedt, N. Jain, and N. Krylov, *Filtering and Prediction: A Primer*. Providence, R.I.: Amer. Math. Soc., 2007.

[13] B. Gedik and L. Liu, "Location privacy in mobile systems: A personalized anonymization model," in *Proc. IEEE ICDCS*, Columbus, OH, Jun. 2005, pp. 620–629.

[14] P. Golle and K. Partridge, "On the anonymity of home/work location pairs," in *Proc. Pervasive*, Nara, Japan, May 2009, pp. 390–397.

[15] M. Gruteser and D. Grunwald, "Anonymous usage of location-based services through spatial and temporal cloaking," in *Proc. ACM MobiSys*, San Francisco, CA, May 2003, pp. 31–42.

[16] S.-S. Ho and S. Ruan, "Differential privacy for location pattern mining," in *Proc. ACM SIGSPATIAL—Int. Workshop Security Privacy GIS LBS*, Chicago, IL, Nov. 2011, pp. 17–24.

[17] D. S. Hochbaum and B. Fishbain, "Nuclear threat detection with mobile distributed sensor networks," *Ann. Oper. Res.*, vol. 187, no. 1, pp. 45–63, 2009.

[18] B. Hoh, M. Gruteser, H. Xiong, and A. Alrabady, "Preserving privacy in GPS traces via uncertainty-aware path cloaking," in *Proc. ACM CCS*, Alexandria, VA, Oct. 2007, pp. 161–171.

[19] T. Karagiannis, J.-Y. L. Boudec, and M. Vojnovic, "Power law and exponential decay of inter contact times between mobile devices," in *Proc. ACM MobiCom*, Montreal, QC, Canada, Sep. 2007, pp. 183–194.

[20] N. Li and T. Li, "t-closeness: Privacy beyond k-anonymity and l-diversity," in *Proc. IEEE ICDE*, Istanbul, Turkey, Apr. 2007, pp. 106–115.

[21] A. Machanavajjhala, J. Gehrke, D. Kifer, and M. Venkitasubramaniam, "l-diversity: Privacy beyond k-anonymity," in *Proc. IEEE ICDE*, Atlanta, GA, Apr. 2006, p. 24.

[22] A. Machanavajjhala, D. Kifer, J. M. Abowd, J. Gehrke, and L. Vilhuber, "Privacy: Theory meets practice on the map," in *IEEE ICDE*, Cancun, Mexico, Apr. 2008, pp. 277–286.

[23] D. J. Martin, D. Kifer, A. Machanavajjhala, J. Gehrke, and J. Y. Halpern, "Worst-case background knowledge for privacy-preserving data publishing," in *Proc. IEEE ICDE*, Istanbul, Turkey, Apr. 2007, pp. 126–135.

[24] J. Meyerowitz and R. R. Choudhury, "Hiding stars with fireworks: Location privacy through camouflage," in *Proc. ACM MobiCom*, Beijing, China, Sep. 2009, pp. 345–356.

[25] P. Mohan, V. N. Padmanabhan, and R. Ramjee, "Nericell: Rich monitoring of road and traffic conditions using mobile smartphones," in *Proc. ACM SenSys*, Raleigh, NC, Nov. 2008, pp. 357–358.

[26] A. Narayanan and V. Shmatikov, "Robust de-anonymization of large sparse datasets," in *Proc. 29th IEEE Symp. Security Privacy*, Oakland, CA, May 2008, pp. 111–125.

[27] A. Narayanan and V. Shmatikov, "De-anonymizing social networks," in *Proc. IEEE Symp. Security Privacy*, Oakland, CA, May 2009, pp. 173–187.

[28] NASA Ames Research Center, Moffett Field, CA, "NASA Ames scientist develops cell phone chemical sensor," 2009 [Online]. Available: http://www.nasa.gov/centers/ames/news/features/2009/cell_phone_sensors.html

[29] M. E. Nergiz, M. Atzori, Y. Saygin, and B. Guc, "Towards trajectory anonymization: A generalization-based approach," *Trans. Data Privacy*, vol. 2, no. 1, pp. 47–75, 2009.

[30] A. Papoulis, *The Fourier Integral and Its Applications*. New York: McGraw-Hill, 1962.

[31] M. Piorkowski, N. Sarafijanovic-Djukic, and M. Grossglauser, "CRAWDAD data set epfl/mobility," v. 2009-02-24, Feb. 2009 [Online]. Available: http://crawdad.cs.dartmouth.edu/epfl/mobility

[32] V. Rastogi and S. Nath, "Differentially private aggregation of distributed time-series with transformation and encryption," in *Proc. SIGMOD*, Indianapolis, IN, Jun. 2010, pp. 735–746.

[33] B. Ribeiro, W. Chen, G. Miklau, and D. Towsley, "Analyzing privacy in enterprise packet trace anonymization," in *Proc. NDSS*, San Diego, CA, Feb. 2008, pp. 87–100.

[34] HKUST, Hong Kong, "ShangHai grid," [Online]. Available: http://www.cse.ust.hk/scrg

[35] R. Shokri, G. Theodorakopoulos, G. Danezis, J.-P. Hubaux, and J.-Y. Le Boudec, "Quantifying location privacy: The case of sporadic location exposure," in *Proc. PETS*, Vigo, Spain, Jul. 2011, pp. 57–76.

[36] R. Shokri, G. Theodorakopoulos, J.-Y. Le Boudec, and J.-P. Hubaux, "Quantifying location privacy," in *Proc. IEEE Symp. Security Privacy*, Oakland, CA, May 2011, pp. 247–262.

[37] L. Sweeney, "K-anonymity: A model for protecting privacy," *Int. J. Uncertainty, Fuzziness Knowl.-Based Syst.*, vol. 10, no. 5, pp. 557–570, 2002.

[38] R. R. Sokal and P. H. A. Sneath, *Principles of Numerical Taxonomy*. San Francisco, CA: Freeman, 1963.

[39] M. Terrovitis and N. Mamoulis, "Privacy preservation in the publication of trajectories," in *Proc. MDM*, Beijing, China, Apr. 2008, pp. 65–72.

[40] G. Wondracek, T. Holz, E. Kirda, and C. Kruegel, "A practical attack to de-anonymize social network users," in *Proc. IEEE Symp. Security Privacy*, Oakland, CA, May 2010, pp. 223–238.

[41] X. Xiao and Y. Tao, "M-invariance: Towards privacy preserving re-publication of dynamic datasets," in *Proc. ACM SIGMOD*, Beijing, China, Jun. 2007, pp. 689–700.

Chris Y. T. Ma received the B.Eng. degree in computer engineering and M.Phil. degree in computer science and engineering from the Chinese University of Hong Kong, Hong Kong, in 2004 and 2006, respectively, and the Ph.D. degree in computer science from Purdue University, West Lafayette, in 2010.

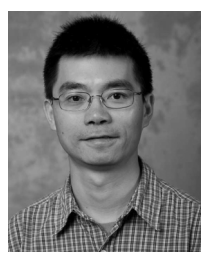
He is currently a Postdoctoral Researcher with the Advanced Digital Sciences Center, Illinois at Singapore, Singapore. His research interests include performance and security study of wireless networks, mobile sensor networks, and smart grids.

Dr. Ma was a recipient of the Bilsland Dissertation Fellowship and a Purdue Summer Research Grant.

David K. Y. Yau (M'10) received the B.Sc. degree (first class honors) from the Chinese University of Hong Kong, Hong Kong, in 1989, and the M.S. and Ph.D. degrees from the University of Texas at Austin in 1992 and 1997, respectively, all in computer science.

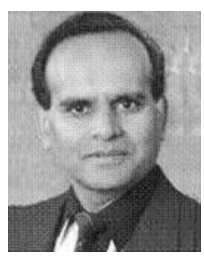
He is currently a Distinguished Scientist with the Advanced Digital Sciences Center, Illinois at Singapore, Singapore, and an Associate Professor of computer science with Purdue University, West Lafayette, IN. His other areas of research interest are protocol design and implementation, wireless and sensor networks, network security, network incentives, and smart grids.

Dr. Yau served as Associate Editor of the IEEE/ACM TRANSACTIONS ON NETWORKING from 2004 to 2009; Vice General Chair (2006), TPC Co-Chair (2007), and TPC Area Chair (2011) of the IEEE International Conference on Network Protocols (ICNP); TPC Co-Chair (2006) and Steering Committee member (2007–2009) of the IEEE International Workshop Quality of Service (IWQoS); and TPC Track Co-Chair (Network/Web/P2P Protocols and Applications) of the IEEE International Conference on Distributed Computing Systems (ICDCS) 2012. He was the recipient of an NSF CAREER Award for research in quality of service provisioning.



Nung Kwan Yip received the B.Sc. degree in electrical engineering from Purdue University, West Lafayette, IN, in 1991, and the Ph.D. degree in mathematics from Princeton University, Princeton, NJ, in 1996.

He was a Courant Instructor with New York University, New York, and a Van Vleck Assistant Professor with the University of Wisconsin–Madison before joining the Department of Mathematics, Purdue University, in 1999. He has also made long-term visits to the Institute of Pure and Applied Mathematics, Los Angeles, CA; Institute for Mathematics and Its Applications, Minneapolis, MN; and the Max Planck Institute for the Mathematics in the Sciences, Leipzig, Germany. His main research interests are partial differential equations, calculus of variations, and probability theory. His recent works include mathematical modeling in materials science and stochastic optimization in computer sensing networks.



Nageswara S. V. Rao (F'08) received the B.Tech. degree in electronics and communications engineering from the National Institute of Technology, Warangal, India, in 1982, the M.E. degree in computer science and automation from the Indian Institute of Science, Bangalore, in 1984, and the Ph.D. degree in computer science from Louisiana State University, Baton Rouge, in 1988.

He is currently a Corporate Fellow with the Computer Science and Mathematics Division, Oak Ridge National Laboratory, Oak Ridge, TN, which he joined in 1993. He has been on assignment at the US Missile Defense Agency as the Technical Director of the C2BMC Knowledge Center, Fort Belvoir, VA, since 2008. He was an Assistant Professor with the Department of Computer Science, Old Dominion University, Norfolk, VA, from 1988 to 1993. He has published more than 250 technical conference and journal papers in the areas of sensor networks, information fusion, and high-performance networking.

Dr. Rao received the 2005 IEEE Technical Achievement Award for his contributions in the information fusion area.