Privacy Vulnerability of Published Anonymous Mobility Traces*

Chris Y. T. Ma* David K. Y. Yau*♯ Nung Kwan Yip* Nageswara S. V. Rao♢
* Purdue University, West Lafayette, IN, USA
♯ Advanced Digital Sciences Center, Illinois at Singapore
♢ Oak Ridge National Laboratory, TN, USA
{ma18, yau}@cs.purdue.edu, yip@math.purdue.edu, raons@ornl.gov

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 (captured in both real and synthetic traces), 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.

Categories and Subject Descriptors

C.2.4 [**Computer-Communication Networks**]: Distributed Systems; K.6.5 [**Management of Computing and Information Systems**]: Security and Protection

General Terms

Algorithms, Security, Experimentation

Keywords

Location Privacy, Entropy

*Research was supported in part by U.S. Department of Energy under SensorNet grant number AC05-00OR22725 and Mathematics of Complex, Distributed, Interconnected Systems program, Office of Advanced Computing Research, in part by U.S. National Science Foundation under grant numbers DMS-0707926 and CNS-0964086, and in part by Bilsland Dissertation Fellowship awarded to C. Ma.

MobiCom'10, September 20–24, 2010, Chicago, Illinois, USA.

1. 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 [17], road surface condition sensing [5], chemical detection [19], and radiation detection [9]. 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 multi-modal sensors that can be readily embedded into widely adopted personal handheld devices including smart phones. 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 [12]. 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 [4] or websites set up by individual research groups [21].

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 develop analytically inference strategies that the adversary may use to maximize its effectiveness in identifying one or more victims under different system assumptions. We show 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 captured in both real and synthetic traces, 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 two-fold.

(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 synthetic and real traces from the perspective of the privacy problem. Despite generated from the same basic statistics (e.g., area of the network and average/maximum speeds of the mobile nodes) of the real traces, the synthetic traces may behave quite differently from the real traces. Their different characteristics will result in quite different performance under various privacy attacks. For instance, mobile nodes in the synthetic traces are more sparsely distributed in the network. This leads to easier de-anonymization of synthetic traces when the adversary attacks by collecting side information passively, but the de-anonymization may take a longer time if the adversary attacks by observing the participants directly.

2. RELATED WORK

Various privacy issues of published data sets have been studied in the literature [1, 22, 2, 24, 15]. Sweeney [22] proposes a protection model named *k-anonymity* and a set of accompanying policies for privacy protection. When k-anonymity is satisfied, each individual should be indistinguishable from $k-1$ other individuals. Bayardo and Agrawal [2] propose a practical method to identify a provably optimal k-anonymization of real census data, or a good anonymization when the optimal one cannot be found in reasonable time, because the general problem is NP-hard. Xiao and Tao [24] propose a generalization principle of *m-invariance* to effectively limit the risk of privacy disclosure in data republications, given the many potential correlations among various snapshots of each data entry in subsequent publications that can be used to derive sensitive information. Martin *et al.* [15] quantify the impact of background knowledge possessed by an attacker on privacy breach. They express the attacker's background knowledge in a language, and provide an algorithm to determine the amount of disclosed sensitive information in the worst case with respect to the amount of the background knowledge. In the data mining context, Agrawak and Srikant [1] propose a reconstruction procedure to build a decision-tree classifier without accessing the precise information in individual data records, so that the distributions of the data values can be reconstructed with sufficient accuracy. They also suggest the use of value-class membership and value distortion to preserve privacy.

Narayanan and Shmatikov [18] study the privacy implications of releasing anonymous and perturbed user ratings of movies in the Netflix database for a research competition to better predict customers' preferences for different new movies.[1] They propose an effective algorithm to de-anonymize the data set, and verify its performance with the published data. Their problem, in which an adversary is given perturbed side information about a victim to identify the victim's record among all the database records, is one of the attack scenarios we consider in this paper.

Privacy protection of mobile nodes in location-based services has also been studied [8, 7, 10, 16]. One proposed approach is to reduce the spatial/temporal granularity of the location information made available to the service provider while achieving satisfactory service effectiveness [8, 7]. Hoh *et al.* [10] devise a protection strategy to release user data only when certain privacy constraints are met. Meyerowitz and Choudhury [16] suggest sending fake requests with the real ones to reduce the ability to trace a mobile node over time.

Our problem differs from the previous work by its specific focus on privacy leaks of user location information considering the characteristics of anonymous mobility traces of users, under the assumption that an attacker, assisted by different amounts of side information that can be realistically obtained, employs various well grounded strategies to infer the private information. As discussed in Section 1, we are motivated by the emerging practice of collecting traces of real users in a mobile network and publishing anonymized and possibly cloaked versions of these traces through various data portals, to assist in the design and evaluation of these networks.

3. PROBLEM DEFINITION

We assume that a set of traces, each of which 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

[1]Achievement of this goal is expected to have significant impact on the profitability of Netflix.

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 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 1. 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. 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 pm 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.

3.1 Problem A: Passive adversary

In this problem setting, the adversary is given the complete (annonymized) traces. The adversary's goal is, given some pieces of side information about a pre-determined but unknown victim, to identify in some optimal fashion the complete path history of the chosen victim. The key assumptions are: (i) the adversary is *passive* in the sense that it does not actively go out to seek encounters with potential victims; (ii) 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).

3.2 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 identity 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 pre-determines 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.

3.3 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.

$\{L_i\}_{i=1,2,\ldots N}$: 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 iid random variable Z_k's of some given distribution $\Pr_Z$. Hence we have

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

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 which 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 [6, Chapters 3, 8]). On the other hand, our simulation results already produce robust results even for the non-Markovian real traces. Hence we will not be side tracked by invoking the more refined models. Instead, we will emphasize the implications of general knowledge about nodal movements towards the privacy issues.

4. 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 3.

4.1 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 below. Given $\{R(t_k)_{k=1,2,\ldots}\}$, compute

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

The goal of the MLE is to find i which maximizes the expression (2). Note that the denominator is a constant. 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:

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

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 iid, we have

$$\Pr(R(t_k), k=1,2,\ldots|L_i) = \Pi_k \Pr_Z(R(t_k) - L_i(t_k)), \quad (4)$$

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:

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

which can be easily expressed in terms of the transition matrix P_{ij}. (Note that 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's takes on specific forms. For example,

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

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

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

$$\min_i \sum_k \Big|R(t_k) - L_i(t_k)\Big|^2. \quad (7)$$

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

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

where $\chi_A(x, y) = 1$ or 0 depending on if $x - y \in A$ or not.

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 $\Pi_k \Pr_Z(R(t_k) - L_i(t_k))$. The trace with the *maximum similarity value* is declared to be the victim's.

(2) Minimum Square Approach (MSQ). This is the same as formulation (7), i.e., the *similarity* value of trace i is given by $-\sum_k \Big|R(t_k) - L_i(t_k)\Big|^2$. 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) 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:

$$\sum_{k=1}^{M} I_{2\sigma}\left(L_i(t_k), R(t_k)\right), \quad (9)$$

where $I_{2\sigma}(x, y) = 1$ if $|x - y| \le 2\sigma$ 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 [18], 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,

$$\sum_{k=1}^{M} \frac{1}{\text{Weight}(R(t_k))} \exp\left\{-\frac{1}{C}\left|L_i(t_k) - R(t_k)\right|\right\}, \qquad (10)$$

where $\text{Weight}(R(t_k))$ is some weight assigned to the revealed cell $R(t_k)$ and C is a constant. In the simulations we assign equal weights to all of the cells 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,

MLE$_2$:

$$\Pi_k\Big(\sum_{l\in\Theta} P_i(l, t_k)\Pr_Z(R(t_k) - l)\Big) \qquad (4_2)$$

MSQ$_2$:

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

BAS$_2$:

$$\sum_{k=1}^{M} \left(\sum_{l\in\Theta} P_i(l, t_k) \times I_{2\sigma}(l, R(t_k))\right) \qquad (9_2)$$

EXP$_2$:

$$\sum_{k=1}^{M} \left(\sum_{l\in\Theta} \frac{P_i(l, t_k)}{\text{Weight}(R(t_k))} \exp\left\{-\frac{1}{C}\left|l - R(t_k)\right|\right\}\right) \qquad (10_2)$$

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 particpants. 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.

4.2 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 Figure 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 which do not agree with the observed information about each victim from 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.

```
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
```

Figure 1: Specification of Attack algorithm.

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 Figure 1. Hence, the adversary identifies a participant more efficiently when it tries to identify as many participants as possible.

5. TRACE CHARACTERISTICS

In this section we begin by analyzing the differences in behaviors between the real traces and simple synthetic traces. Their fundamental differences will be illustrated by six types of mobility traces. They include two sets of real traces: (1) cabs in the San Francisco area [20], and (2) buses in the ShangHai Grid system [21]. Basic statistics about these two sets of traces are listed in Table 1. We then make use of statistics from the San Francisco cab trace to generate four other synthetic traces using the random waypoint mobility model ($rway$) and two of its variants which impose a maximum trip length of x (in km) ($rway - x$ with $x = 10,\ 20$),

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

Table 1: Basic statistics of the real traces.

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

and the random walk model (*rwalk*). In particular, we use the number of cells visited by the cabs as the size of the map of the synthetic traces (approximated by a square of size $0.63^\circ \times 0.63^\circ$), the average speed of the cabs (about 13.8 mph) as that of the synthetic mobile nodes, and the average time the cabs are active (about 15 days) as the simulation run time of the synthetic traces. We notice that other mobility models [11, 14, 13] have been proposed to better approximate the movement of real-world entities. However, the main goal of the current paper is not to decide what type of synthetic traces is best in realism in what situations, but to provide a quantitative measurement of the mobility characteristics and their impact on the privacy issue. Hence we will only use the synthetic traces to establish the necessary intuition in understanding the impact of information collection with respect to privacy. To continue, 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 1 unless stated otherwise. Characteristics of traces are studied using the four metrics as described in Sections 5.1–5.4. Observations that can be explained using differences between movement preferences of the mobile nodes are summarized at the end of this section.

5.1 Correlation between traces

We use the Pearson product-moment correlation coefficient [23] to quantify the correlations between node pairs. 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:

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

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

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

The figure shows that movements of different cabs have little or no correlation, while those of the random walk nodes have higher correlations. It is because cabs are unlikely to follow each other for a long time. Random walk nodes show the highest correlation since their movements are synchronized and their choices of next movement are limited to the

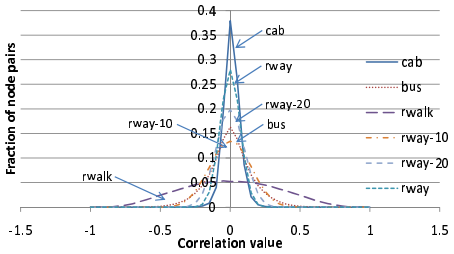


Figure 2: Distribution of correlations between traces of the same set.

immediate neighboring cells. Although a low correlation between the traces indicates that they do not share common paths over a long time, having a high correlation between the traces does not imply otherwise. It is because the computed correlation is the distance relative to the mean position of each trace but not to a common location, and it does not take into account the orientation of the nodes.

5.2 Autocorrelation of the same trace

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

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

Figure 3 depicts the autocorrelation as a function of the time shift s.

The figure shows that for 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. The more localized movement of the random walk nodes makes their autocorrelation value only drops slowly as a function of shifting.

5.3 Complexity of movement

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

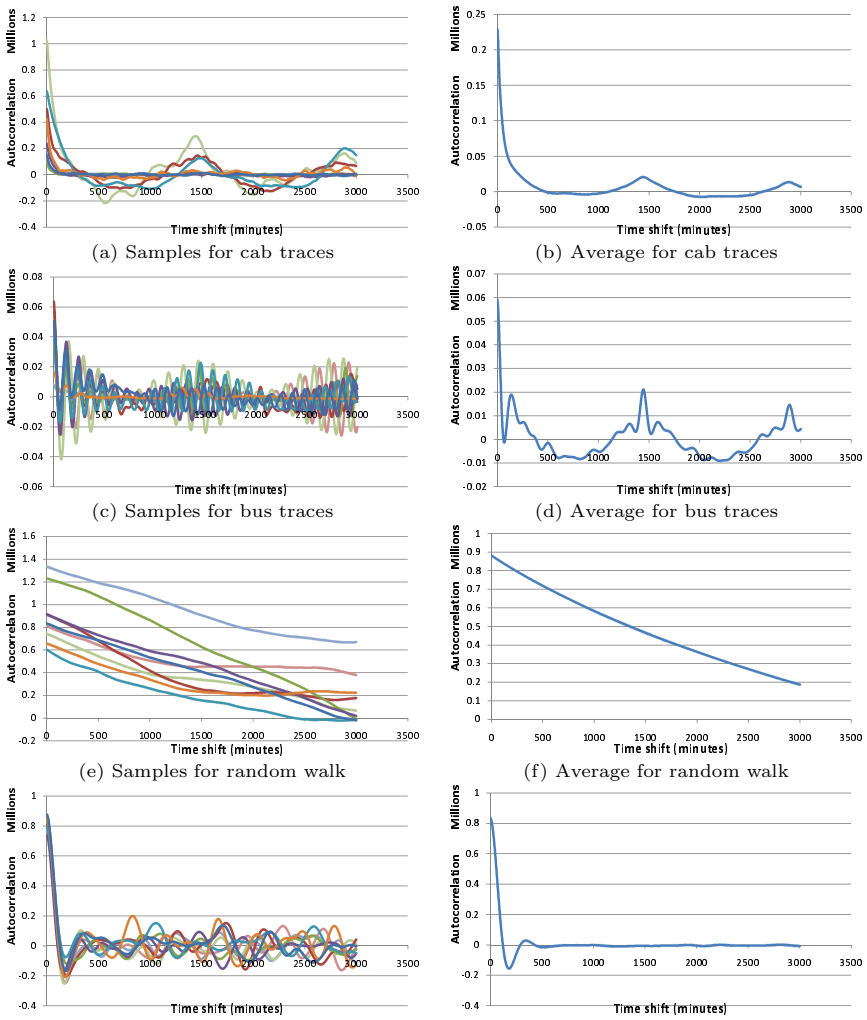
$$H_n\left(\{L_i\}_{i=1,2,\ldots N}\right) = \sum_{\Theta^{n-1}} p(l_1,\ldots,l_{n-1}) \times \left\{ \sum_{\Theta} p(l_n|l_1,\ldots,l_{n-1}) \log p(l_n|l_1,\ldots,l_{n-1}) \right\}. \quad (11)$$

In the above, 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 of the locations in the collection of traces $\{L_i\}_{i=1,2,\ldots N}$.

The above function, defined for general stochastic processes, is well-known in the information theory community (see for example [3, Chapter 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 (11) as a function of n is depicted in Figure 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.

Notice that because of the relatively slow movement of the mobile nodes, the synthetic traces do not have enough time to reach steady state if we limit the synthetic traces to the same length as the real traces in quantifying their



(a) Samples for cab traces

(b) Average for cab traces

(c) Samples for bus traces

(d) Average for bus traces

(e) Samples for random walk

(f) Average for random walk

(g) Samples for random waypoint

(h) Average for random waypoint

Figure 3: Autocorrelation of each trace for different sets of traces as a function of the time shift s.

characteristics. As a result, although one may expect the random walk traces to have a constant entropy as the order increases, we observe otherwise in the figure, and the entropy of the synthetic traces also drops more significantly than the real traces as the order increases.

5.4 Distance between traces

Figure 5(a) depicts the distribution of average distance between trace pairs, which is defined as

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

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

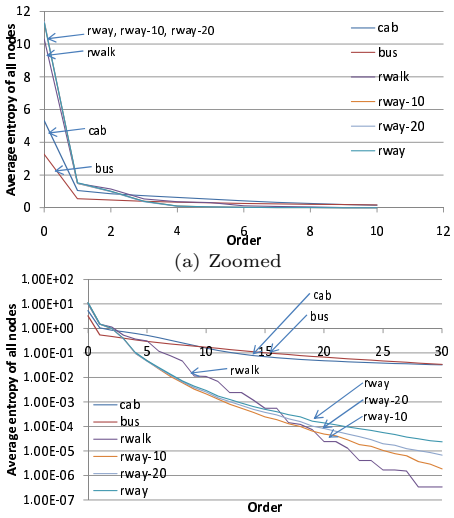
$$\text{Dist}_{\text{Min}}(i,j) = \min_{k} \left(|L_i(s_k) - L_j(s_k)|\right).$$

5.5 Implications of the trace characteristics

Many of the observed different characteristics of the mobility traces can be summarized and explained using the lack of *preferred locations* and random initial positions of the synthetic traces. Specifically, real traces show natural preferences for certain places visited by the mobile nodes, such as the busy downtown area for the cabs and the assigned routes for the buses, whereas the synthetic traces do not.

Since synthetic traces do not have a set of preferred locations to visit, and they are placed randomly in the network initially, they exhibit much sparser spatial distributions in the network than the real traces. Moreover, the shorter the maximum trip length in the synthetic traces, the sparser they are in the network. These observations are shown through the higher order-0 entropy of the synthetic traces than the real traces (Figure 4), the larger autocorrelation values of the synthetic traces (Figure 3), the longer average distances between synthetic node pairs (Figure 5(a)) (with random walk having the largest average distance), and the fact that not all the random walk nodes have met each other during the simulation (Figure 5(b)). On the other hand, the real traces have their preferred visiting places, resulting in a smaller H_0 than the synthetic traces, and the entropy drops much more slowly when the order increases. This is also reflected in the smaller average distances between cabs or between buses, although the bus distances exhibit a broader range 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 Figure 5(b), while almost 100% of the cabs have met each other.

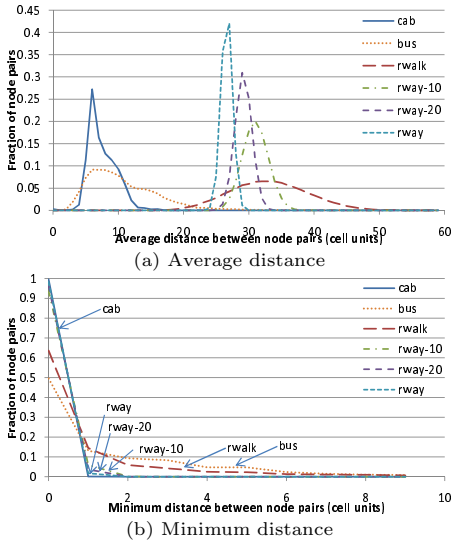
When nodes are more sparsely distributed in the network area, more efficient victim identification results when the ad-



(a) Zoomed

(b) Log scale

Figure 4: Order-n complexity of different sets of traces as a function of order.



(a) Average distance

(b) Minimum distance

Figure 5: Distribution of average and minimum distances between pairs of traces.

versary collects the side information passively (**Problems A1** and **A2**). Hence, we would expect victims in the synthetic traces to be more easily identified than in the real traces. 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. On the other hand, 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 the following section.

6. SIMULATION RESULTS

6.1 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 <time, location> 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. Results reported are for simulation experiments each repeated 100,000 times.

We quantify the performance of the strategies with the following metrics, (i) *Fraction of correct conclusions.* A conclusion is correct if the victim is uniquely identified according to the criterion of highest similarity metric; (ii) *Fraction of incorrect conclusions.* A conclusion is incorrect when the victim is not among the set of candidates having the highest similarity metric.

6.1.1 Problem A1

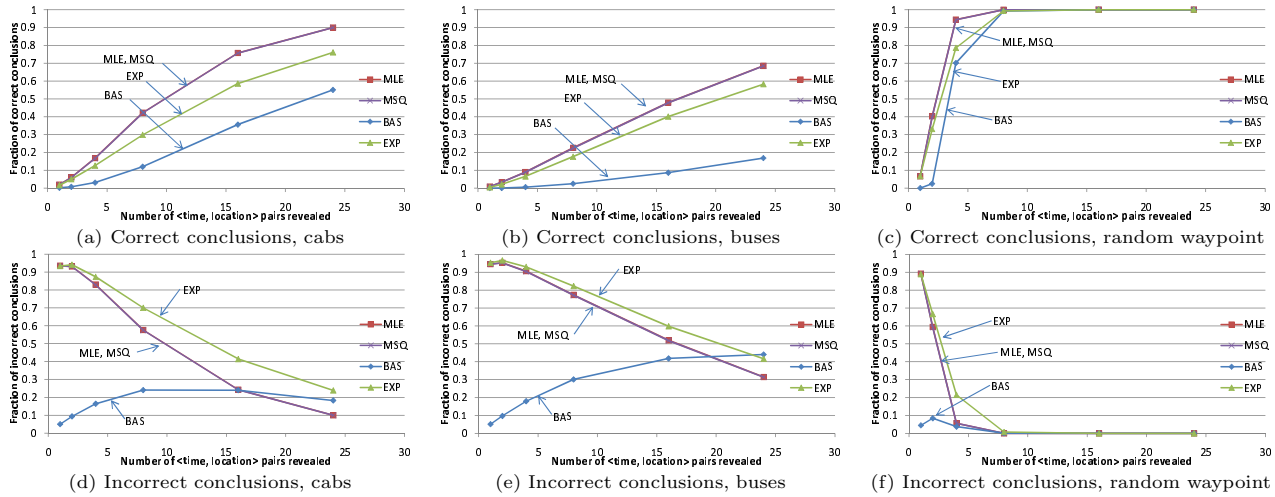
We present the results based on the perception of the adversary on the noise.

(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**. Figure 6 shows the performance of the attack strategies using the cab, bus, and random waypoint traces. Results of random walk traces are not shown because they give similar trends as random waypoint.

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 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.

(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. Figures 7(a) and (d) show the performance of the strategy when the actual and assumed noise is Gaussian and Uniform, respectively. Figures 7(b) and (e) show the results



(a) Correct conclusions, cabs (b) Correct conclusions, buses (c) Correct conclusions, random waypoint

(d) Incorrect conclusions, cabs (e) Incorrect conclusions, buses (f) Incorrect conclusions, random waypoint

Figure 6: (Problem A1) Performance of various metrics as a function of the number of <location, time> pairs revealed. (a), (d) San Francisco cab traces, (b), (e) ShangHai Grid bus traces, and (c), (f) random waypoint traces. Zero-mean Gaussian noise with $\sigma = 5$.

when the actual and assumed noise is Uniform and Gaussian, respectively. Figures 7(c) and (f) show 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.

6.1.2 Problem A2

Figure 8 depicts the performance of the attack approaches for different sampling time intervals for the cab traces. Zero-mean Gaussian noise with $\sigma = 5$ is introduced into the spatial domain 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. Figure 9 depicts the results for the bus traces and the synthetic random waypoint 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 insignificant for the synthetic traces, but it is more noticeable for the bus traces.

When we compare the performance of the attack approaches in this case with the special case in the previous subsection, in which no inference using a general movement model is necessary, the performance here does not degrade significantly for $\mathbf{MLE}_2$ and $\mathbf{MSQ}_2$. Interestingly, $\mathbf{BAS}_2$ gives a much larger fraction of incorrect conclusions and slightly larger fraction of correct conclusions initially when movement has to be inferred, while $\mathbf{EXP}_2$ performs about the same in both cases.

6.1.3 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.

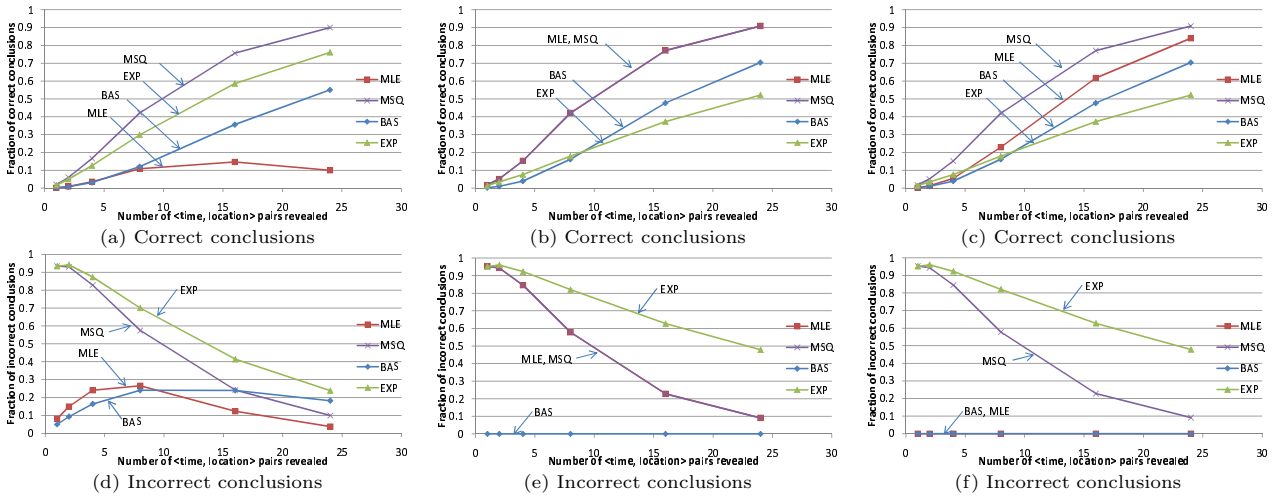
The results also verify our claim in Section 5 that victim identification is much easier for the synthetic than real traces, due to higher nodal sparsity in the former.

6.2 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 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. As discussed in Section 5, we expect the active adversary needs a longer time to identify all the synthetic traces than the cab traces, because the former have sparser node distributions. Further, we expect random walk to require the longest time among the synthetic traces, and the bus trace to require the longer time between the real traces.

6.2.1 Problem B1

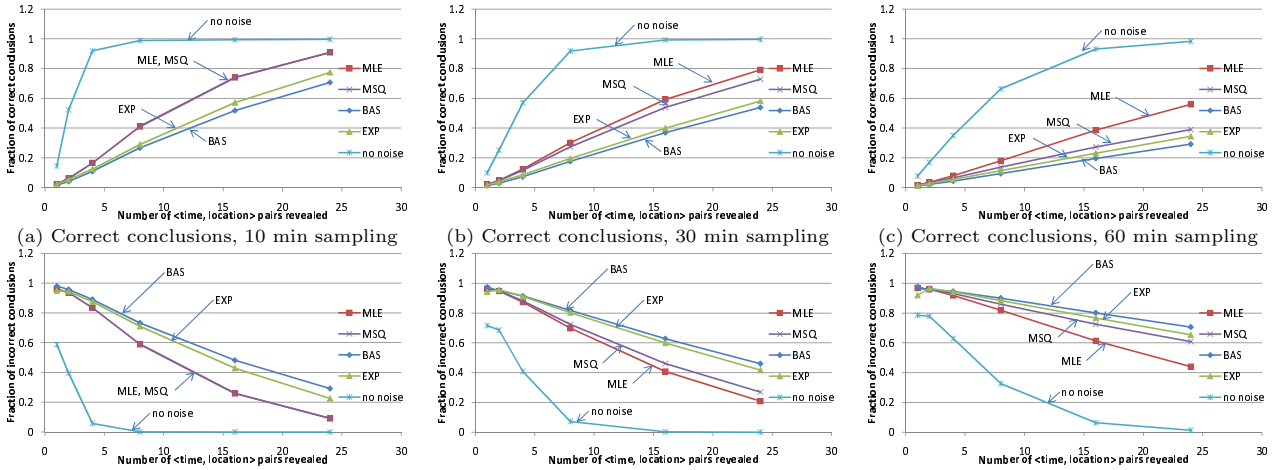
Figure 10 depicts the average k-anonymity of the victims as observed by the adversary as a function of the attack time for different sets of the traces, when the adversary is one of the mobile nodes. The figure shows that the most reduction in k-anonymity for each participant results from observations made in the first day in the real traces. The



(a) Correct conclusions (b) Correct conclusions (c) Correct conclusions

(d) Incorrect conclusions (e) Incorrect conclusions (f) Incorrect conclusions

Figure 7: (Problem A1) Performance of various metrics as a function of number of <location, time> 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) 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

Figure 8: (Problem A2) Performance of various metrics for attacks requiring different degrees of movement inference for each trace as a function of number of <time, location> pairs revealed. San Francisco cab traces, zero-mean Gaussian noise with $\sigma = 5$. (a), (d) S = ten minutes; T = five minutes; (b), (e) S = thirty minutes; T = fifteen minutes; (c), (f) S = one hour; T = half an hour. (S is the trace sampling time and T is the interval for computing the transition matrix.)

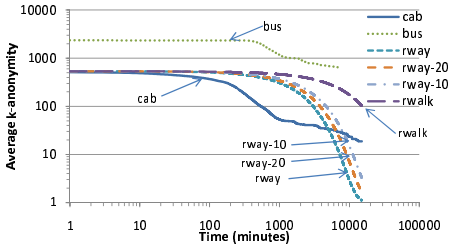


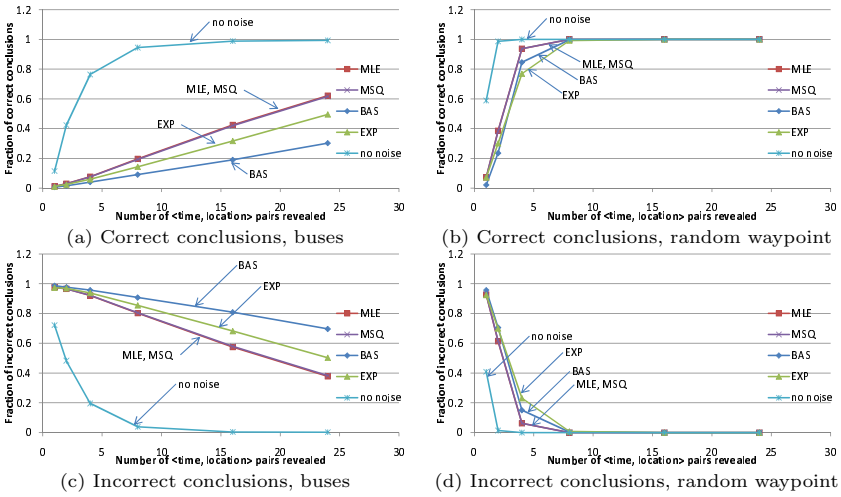
Figure 10: (Problem B1) Average k-anonymity as a function of attack time when the adversary is one of the mobile nodes.

figure also shows that as time increases, the k-anonymity always drops to close to one on average, except for random walk traces and the bus traces as expected. Notice also from the figure that there are flat regions in the bus trace results corresponding to night times of the days. The cab traces exhibit a similar behaviour, but it is much less obvious due to the cabs' own mobility characteristics. Synthetic traces have no concept of day and night, and they do not show such behaviour.

6.2.2 *Problem B2*

Figures 11(a)-(d) depict 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. The k-anonymity of the random walk and bus traces drops more slowly than the other traces as expected.



(a) Correct conclusions, buses

(b) Correct conclusions, random waypoint

(c) Incorrect conclusions, buses

(d) Incorrect conclusions, random waypoint

Figure 9: (Problem A2) Performance of various metrics for attacks requiring different degrees of movement inference for each trace as a function of number of <time, location> pairs revealed. Traces are sampled every half an hour and the transition matrix is generated using sampling information every fifteen minutes.

6.2.3 Problem B3

Figures 11(e)-(h) depict 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. 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 random walk and bus traces. However, traveling to too many places may hurt the performance because the adversary may spend too much time traveling over unpopular places.

6.2.4 Summary on active adversary strategies

The results show that for the real traces, the ability of the active adversary to travel helps it identify many of the victim traces within one day. For synthetic traces, however, the attack efficiency is lower because their spatial distribution is sparser, verifying our observation in Section 5. 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.

7. CONCLUSION

In this paper, we studied the privacy vulnerability of publishing traces of mobile nodes 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 degrees of movement inference needed for the attack. In particular, we pointed out some main differences between the synthetic and real traces with respect to the privacy problem. In general, our results showed that the adversary is able to identify victims with high probability even when the available side information is limited.

8. REFERENCES

[1] R. Agrawal and R. Srikant. Privacy-preserving data mining. *SIGMOD Rec.*, 29(2), 2000.

[2] R. J. Bayardo and R. Agrawal. Data privacy through optimal k-anonymization. In *IEEE ICDE*, Tokyo, Japan, April 2005.

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

[4] Dartmouth College. CRAWDAD. Online http://crawdad.cs.dartmouth.edu/.

[5] 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 *ACM MobiSys*, Breckenridge, CO, June 2008.

[6] B. Fristedt, N. Jain, and N. Krylov. *Filtering and Prediction: A Primer*. Providence, R.I.: American Mathematical Society, 2007.

[7] B. Gedik and L. Liu. Location privacy in mobile systems: A personalized anonymization model. In *IEEE ICDCS*, Columbus, OH, June 2005.

[8] M. Gruteser and D. Grunwald. Anonymous usage of location-based services through spatial and temporal cloaking. In *ACM MobiSys*, San Francisco, CA, May 2003.

[9] D. S. Hochbaum and B. Fishbain. Nuclear threat detection with mobile distributed sensor networks. *Annals of Operations Research*, 2009.

[10] B. Hoh, M. Gruteser, H. Xiong, and A. Alrabady. Preserving privacy in gps traces via uncertainty-aware path cloaking. In *ACM CCS*, Alexandria, VA, October 2007.

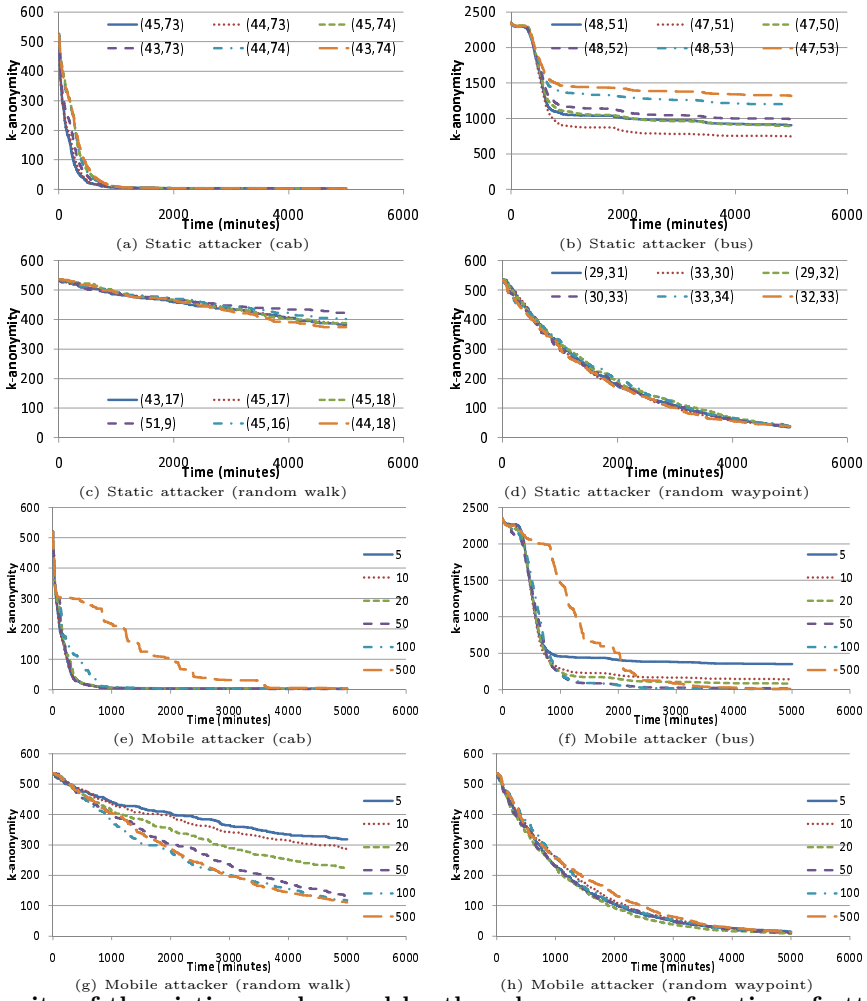


Figure 11: k-anonymity of the victim as observed by the adversary as a function of attack time, when the adversary is (a)-(d) static (Problem B2) and (e)-(h) mobile within a pre-determined path (Problem B3).

[11] R. Jain, D. Lelescu, and M. Balakrishnan. Model T: An Empirical Model for User Registration Patterns in a Campus Wireless LAN. In *ACM MobiCom*, Cologne Germany, August 2005.

[12] T. Karagiannis, J.-Y. L. Boudec, and M. Vojnovic. Power Law and Exponential Decay of Inter Contact Times between Mobile Devices. In *ACM MobiCom*, Montreal, Quebec, Canada, September 2007.

[13] M. Kim, D. Kotz, and S. Kim. Extracting a mobility model from real user traces. In *IEEE Infocom*, Barcelona, Spain, April 2006.

[14] G. Lin, G. Noubir, and R. Rajaraman. Mobility Models for Ad hoc Network Simulation. In *IEEE Infocom*, Hong Kong, China, March 2004.

[15] D. J. Martin, D. Kifer, A. Machanavajjhala, J. Gehrke, and J. Y. Halpern. Worst-case background knowledge for privacy-preserving data publishing. In *IEEE ICDE*, Istanbul, Turkey, February 2007.

[16] J. Meyerowitz and R. R. Choudhury. Hiding stars with fireworks: location privacy through camouflage. In *ACM MobiCom*, Beijing, China, September 2009.

[17] P. Mohan, V. N. Padmanabhan, and R. Ramjee. Nericell: Rich monitoring of road and traffic conditions using mobile smartphones. In *ACM SenSys*, Raleigh, NC, November 2008.

[18] A. Narayanan and V. Shmatikov. Robust de-anonymization of large sparse datasets. In *Proc. of 29th IEEE Symposium on Security and Privacy*, Oakland, CA, May 2008.

[19] NASA Ames Research Center. NASA Ames scientist develops cell phone chemical sensor. Online http://www.nasa.gov/centers/ames/news/features/2009/cell_phone_sensors.html.

[20] M. Piorkowski, N. Sarafijanovic-Djukic, and M. Grossglauser. CRAWDAD data set epfl/mobility (v. 2009-02-24). Downloaded from http://crawdad.cs.dartmouth.edu/epfl/mobility, February 2009.

[21] ShangHai Grid. Online http://www.cse.ust.hk/dcrg.

[22] L. Sweeney. k-anonymity: a model for protecting privacy. *International Journal on Uncertainty, Fuzziness and Knowledge-based Systems*, 10(5), 2002.

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

[24] X. Xiao and Y. Tao. M-invariance: towards privacy preserving re-publication of dynamic datasets. In *ACM SIGMOD*, Beijing, China, June 2007.