Revisiting Location Privacy from a Side-Channel Analysis Viewpoint (Extended Version)

Clément Massart, François-Xavier Standaert

ICTEAM - Crypto Group, Université catholique de Louvain, Belgium.

Abstract. Inspired by the literature on side-channel attacks against cryptographic implementations, we describe a framework for the analysis of location privacy. It allows us to revisit (continuous) re-identification attacks with a combination of information theoretic and security metrics. Our results highlight conceptual differences between re-identification attacks exploiting leakages that are internal or external to a pseudonymised database. They put forward the amount of data to collect in order to estimate a predictive model as an important – yet less discussed – dimension of privacy assessments. They finally leverage recent results on the security evaluations/certification of cryptographic implementations to connect information theoretic and security metrics, and to formally bound the risk of re-identification with external leakages.

1 Introduction

Location privacy has become an important concern with the advent of pervasive computing: we refer to [2] for one of the first studies motivating this active line of research. In this paper, we are interested in the quantification of location privacy in a setting where an adversary can access a database with location information about different users, together with their pseudonyms. We focus in particular on the risks of re-identification attacks, where an adversary tries to exploit leakages (i.e., the location data of some individuals supposedly in the pseudonymized database) to re-identify users. Such re-identification attacks are continuous (i.e., it is possible for the adversary to accumulate leakages for the same user). In this context, our starting observation is that leakages can be internal (i.e., part of the data collected in the database) or external (i.e., fresh observations).

Our first contribution is a consolidating one. We revisit re-identification attacks with a combination of information theoretic and security metrics, as usually considered in the evaluation of leaking cryptographic implementations against side-channel attacks [20]. We put forward that re-identification attacks with internal leakages can be captured with information theoretic metrics similar to Diaz et al.'s anonymity degree [7], and that re-identification attacks with external leakages can be captured with security metrics (similar to Maouche et al.'s re-identification rate [15]). We consolidate these results by connecting both types of metrics thanks to established results in the worst-case evaluation of cryptographic implementations [8], which prove that the success rate of a worst-case side-channel attack is (under some assumptions) proportional to the mutual information between its target key and the leakages it exploits.

We then show that this consolidating effort can lead to both new observations / refined intuitions and technical advances in the analysis of location privacy.

A first novel observation is that the database size has opposite effects on attacks using internal and external leakages: attacks with internal (resp., external) leakages become more challenging (resp., easier) when the database grows.

A second novel observation is that most existing location privacy metrics tailored for the evaluation of external leakages quantify (to some extent) the success of an attack given a statistical model for the collected data. We argue that the convergence of the model is also interesting to analyze since it determines the amount of data needed to infer something about a user's behavior.

Based on these observations, we conclude that evaluating the risks of re-identifications with external leakages is in general more challenging since they increase when collecting more data (or merging databases). In this respect, our third and most important contribution is to show that such risks can be formally bounded. For this purpose, we leverage a recent result in the leakage certification of cryptographic implementations [3], which shows that our information theoretic metrics evaluated with internal leakages are (in expectation) an upper bound of these metrics evaluated with external observations. As a result, we can bound the risk of re-identification attacks with external leakages independent of the database size (with the bound becoming tight as this size increases).

Finally, we additionally show that localization attacks where an adversary tries to predict the position of a user (as in [19]) can be captured in a similar framework, and critically depends on the time dimension of the observations which determines the adversary's efforts to intercept a user.

2 Definitions and framework

In this section, we specify the location data that we aim to analyze and the estimation tools used to characterize statistical distributions.

Data specification. The location data we consider is based on spatial coordinates: longitude x and latitude y (possibly with a time component t). In general, we will consider two types of observations. First, "independent observations" where every triple (x, y, t) is analyzed independently.[1] These could for example correspond to the location data of a mobile phone application (where consecutive observations are distant in space and time), as pictured in Figure 1 (left). We will next refer to this data as "positions" and denote them as $\boldsymbol{p} = (x, y, t)$. Second, correlated observations for which the joint analysis is expected to lead to improved characterization. It could for example correspond to the GPS data of a jogger (where consecutive observations are close in space and time), as pictured in Figure 1 (right). We will denote them as "routes" in the following.

[1] The word "independent" does not refer to the fact that these observations are truly independent, but only to the fact that such observations are exploited assuming it.

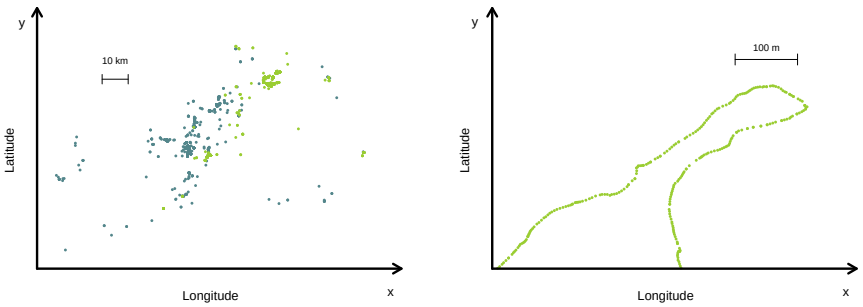

Fig. 1. Left: positions for two users. Right: route for a single user.

More formally, we first define a set of n users:

$$\mathcal{U} = \{u_1, u_2, \ldots, u_n\}.$$

Assuming that some location data has been collected for each of the users, we denote the jth route of user i as:

$$\boldsymbol{r}_{ij} = \{\boldsymbol{p}_{ij}^1, \boldsymbol{p}_{ij}^2, \ldots, \boldsymbol{p}_{ij}^k, \ldots, \boldsymbol{p}_{ij}^{N_{\boldsymbol{p}}^{ij}}\},$$

with $N_{\boldsymbol{p}}^{ij}$ the number of positions in the route, and denote the number of routes collected for a user i as $N_{\boldsymbol{r}}^i$. Note that independent observations can be considered as single-position routes. In the latter case, the number of routes collected per user equals the number of positions collected per user, next denoted as $N_{\boldsymbol{p}}^i$.

The collection of location data is then formalized as follows. We first assume that the routes are sampled from an unknown statistical distribution that reflects the true users' behavior. This true distribution can be continuous, in which case we denote its Probability Density Function (PDF) as $\mathsf{f}(\boldsymbol{r}|u)$, or discrete, in which case we denote its Probability Mass Function (PMF) as $\mathsf{g}(\boldsymbol{r}|u)$. Next, the sampling process giving rise to a set of $N_{\boldsymbol{r}}^i$ routes for user i is written as:

$$\mathcal{S}_i \xleftarrow{N_{\boldsymbol{r}}^i} \mathsf{f}(\boldsymbol{r}|u_i) \text{ or } \mathcal{S}_i \xleftarrow{N_{\boldsymbol{r}}^i} \mathsf{g}(\boldsymbol{r}|u_i),$$

for the continuous and discrete cases, respectively.

Eventually, even if routes can be sampled from a continuous distribution, their storage is generally discrete. Besides, it is usually convenient for exploitation purposes (or necessary for privacy purposes) to further truncate the data. We reflect this process with a discretization function. Ignoring the time component for simplicity, it decomposes the location space into Δ ($=\delta_1 \times \delta_2$) cells, illustrated in Figure 2 (left) and denoted as $\mathsf{D}_1 : \mathcal{R} \to \{0,1\}^{\Delta}$, with $\mathcal{R}$ the set of all possible routes. In the basic setup of this section, it is computed by assigning a one to each cell where at least one observation of the route falls.

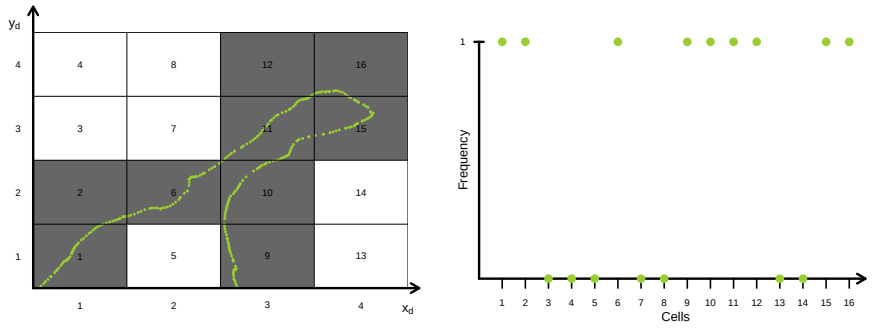

Fig. 2. Left: discretization of a route with D_1 & $\Delta = 16$. Right: D_1-discretized route $\boldsymbol{d}$.

Based on this discretization function, we represent a discretized route as in Figure 2 (right), which we define as $\boldsymbol{d} = \mathsf{D}_1(\boldsymbol{r})$. For conciseness, we only consider the case where the true (unknown) user distribution is continuous and then discretized. Directly discrete user distributions can be formalized identically.

Types of estimations. Given a set of $N^i_{\boldsymbol{r}}$ discretized routes $\mathcal{D}_i$ obtained for a user i denoted as: $\mathcal{D}_i \xleftarrow{N^i_r} \mathsf{D}_1(\mathsf{f}(\boldsymbol{r}|u_i))$, the evaluation of our metrics will be based on the estimation of a model for the true (unknown) distribution. We will consider two types of modeling phases. The first one, next denoted as the direct (DI) estimation process, uses all the samples in the set $\mathcal{D}_i$ and is written as:

$$\tilde{\mathsf{f}}(\boldsymbol{d}|u_i) \xleftarrow{\text{di}} \mathcal{D}_i \text{ or } \tilde{\mathsf{g}}(\boldsymbol{d}|u_i) \xleftarrow{\text{di}} \mathcal{D}_i,$$

for continuous and discrete models. In the following, we will always refer to "DI-estimated" models with the tilde notation. Note that independent of the nature of the true distribution (i.e., continuous or discrete), it is always possible to model it as continuous and discrete. In the second type of modeling, next denoted as the (k-fold) cross–validated (CV) estimation process, the set $\mathcal{D}_i$ is first split into k non-overlapping sets $\mathcal{D}_i^{(j)}$ with $1 \leq j \leq k$. We then define k model building sets $\mathcal{B}_i^{(j)} = \mathcal{D}_i \setminus \mathcal{D}_i^{(j)}$ and the corresponding test sets $\mathcal{T}_i^{(j)} = \mathcal{D}_i^{(j)}$ so that we write the model estimation for all (j)'s as:

$$\left\{\hat{\mathsf{f}}^{(1:k)}(\boldsymbol{d}|u_i), \mathcal{T}_i^{(1:k)}\right\} \xleftarrow{\text{cv}} \mathcal{D}_i,$$

or:

$$\left\{\hat{\mathsf{g}}^{(1:k)}(\boldsymbol{d}|u_i), \mathcal{T}_i^{(1:k)}\right\} \xleftarrow{\text{cv}} \mathcal{D}_i,$$

where the hat notation is for "CV-estimated" models. As usual in statistics, the difference between these estimations is that direct estimation may suffer from overfitting (i.e., the characterization of features that are specific to the collected data $\mathcal{D}_i$ rather than the distribution $\mathsf{f}(\boldsymbol{r}|u_i)$), which cross-validation aims to limit. In this work, we will need both types of estimation in order to capture both internal and external leakages (details are given in Section 3).

Estimation tools. For both types of estimation, we then need to define how the models are built. In the basic setup of this section, we will exploit two (discrete) estimation tools: an exhaustive one and a simplifying one.

The exhaustive model, denoted as $\tilde{\mathsf{g}}_{\text{ex}}(\boldsymbol{d}|u_i)$ for the discrete and direct estimation case (the variant with cross–validation is referred to with the hat notation): it corresponds to a histogram with 2^{Δ} bins corresponding to all the routes. Note that despite the support of this model grows exponentially in Δ, its memory complexity (and the time needed to evaluate it) is bounded by the amount of collected data (i.e., we only need to store routes with non-zero probabilities).

The 1st-order independent model, denoted as $\tilde{\mathsf{g}}_1(\boldsymbol{d}|u_i)$ for the discrete and direct estimation case (the variant with cross–validation is denoted with the hat notation): it corresponds to the independent estimation of $\tilde{\mathsf{g}}_1(\boldsymbol{d}(c)|u_i)$ for the Δ cells, with $\boldsymbol{d}(c)$ a cell of the discretized route $\boldsymbol{d}$. More precisely, for each cell c, this model is computed as follows:

$$\tilde{\mathsf{g}}_1(\boldsymbol{d}(c)|u_i) = \frac{1}{N^i_{\boldsymbol{r}}} \sum_{\boldsymbol{d'} \in \mathcal{D}_i} \boldsymbol{d'}(c).$$

In the case of the exhaustive model, the probability of a user u_i given a discretized route $\boldsymbol{d}$ is directly obtained thanks to Bayes' formula, assuming a uniform distribution for the users. For example, in the DI estimation case it yields:

$$\tilde{\Pr}_{\text{ex}}[u_i|\boldsymbol{d}] = \frac{\tilde{\mathsf{g}}_{\text{ex}}(\boldsymbol{d}|u_i)}{\sum_{j=1}^{n} \tilde{\mathsf{g}}_{\text{ex}}(\boldsymbol{d}|u_j)}.$$

In the 1st-order independent case, it is derived similarly by first computing the 1st-order likelihood as follows:

$$\tilde{\mathsf{q}}_1(\boldsymbol{d}|u_i) = \prod_{c \in \boldsymbol{d}} \tilde{\mathsf{g}}_1\big(\boldsymbol{d}(c)|u_i\big) \cdot \prod_{c \notin \boldsymbol{d}} \Big(1 - \tilde{\mathsf{g}}_1\big(\boldsymbol{d}(c)|u_i\big)\Big),$$

where $c \in \boldsymbol{d}$ denotes the cells that are part of the route $\boldsymbol{d}$ and $c \notin \boldsymbol{d}$ the ones that are not. The probability $\tilde{\Pr}_1[u_i|\boldsymbol{d}]$ is derived thanks to Bayes as:

$$\tilde{\Pr}_1[u_i|\boldsymbol{d}] = \frac{\tilde{\mathsf{q}}_1(\boldsymbol{d}|u_i)}{\sum_{j=1}^{n} \tilde{\mathsf{q}}_1(\boldsymbol{d}|u_j)}.$$

In this case, the probability of a route is estimated by assuming the independence of the observations in each cell, which is obtained by multiplying the probabilities of all the cells in the route. Summarizing, we so far defined models estimated directly and with cross–validation, that correspond to an exhaustive characterization of the routes, or are based on a 1st-order independence assumption. Other options could be considered. For example, Gambs et al. used a modeling based on Markov chains which could also be analyzed with the following tools [13].

We detail in Appendix A how this 1st-order independent model can be generalized to an *oth-order independent model* which can capture higher-order correlations in the distribution by first "extending" the data towards higher-orders

and then using estimation tools similar to the ones described in this section. In this respect, we note that an oth-order independent model is not equivalent to the exhaustive model since the knowledge of a statistical distribution is not equivalent to the knowledge of its moments. So while increasing o can be used to characterize higher-order dependencies of the user's behavior, it cannot lead to an optimal model.

3 Threat models and metrics

Our threat model is depicted in Figure 3 and formalized as follows.

First, as in Section 2, we have a number of users (i.e., Alice, Bob, Carol, David, ... on the figure). For each of them, a number of routes have been collected and stored in a database under different pseudonyms. Pseudonyms are user IDs reorganized according to a secret permutation. Other data may be collected (e.g., performance data for sport applications, preferences for cultural applications, ...). Second, we mostly (yet, as will be clear next, not only) consider an open data scenario where the collected data is anonymized thanks to pseudonyms, and then made public, e.g., to facilitate the investigations of social scientists. Third, we assume that the adversary can have access to two types of leakages. The first one is an "internal leakage". That is, the adversary learns that some route(s) in the database correspond(s) to a user. This typically happens by spying on a user while data is collected. The second one is an "external leakage". That is, the adversary learns that some fresh route(s) correspond(s) to a user. This typically happens by spying on a user after data has been collected.

Concretely, there are two important quantities that impact privacy in this threat model. First, the size of the anonymized database, which we will denote with a number of routes collected per user $N_{\boldsymbol{r}}^i$ (as in Section 2). Second, the number of leakages obtained per user, that we will next denote as $M_{\boldsymbol{r}}^i$. In general, we expect that the number of leakages collected is significantly smaller than the size of the anonymized database.

Based on this setup, the goal of the adversary is to re-identify the users thanks to their leakages, which may for example allow him to gain access to other sensitive data. This can be achieved both with internal and external leakages. Yet, the meaning of successful re-identification with these two types of leakages is quite different. In the latter case (i.e., with external leakages), it implies that the collected data is representative of the true users' distributions. That is, external leakages can only be linked to the collected data of their originating user if this data can be used to predict fresh routes to some extent. By contrast re-identification with internal leakages does not imply anything regarding the representativity of the collected data. That is, since the leakages come from the database, they are guaranteed to be linkable to their originating user (possibly with other users if they are found in the observations of multiple users).

In statistical terms, these two types of leakages therefore directly correspond to our two types of estimations. Internal leakages can be captured with the

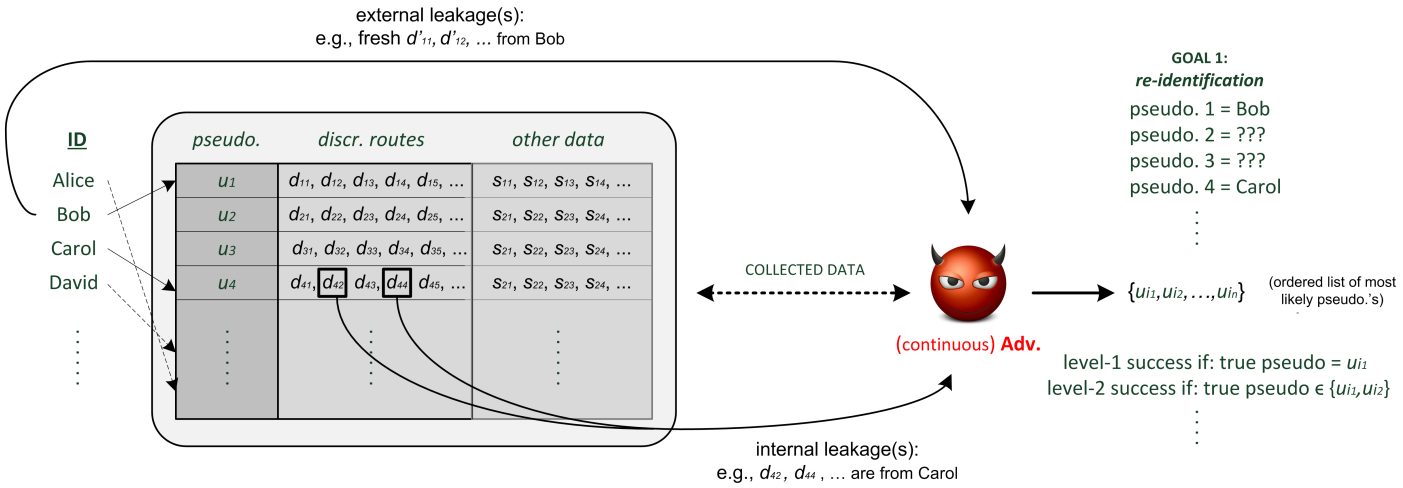

Fig. 3. Re-identification attack threat model.

direct estimation process (with overfitting). External leakages can be captured with the cross-validated one (which prevents overfitting and therefore can be used to assess how predictive the model is against fresh routes).

Eventually, and as illustrated on the figure, given a set of leakages for a user, the adversary outputs an ordered list of most likely pseudonyms. We then say that the attack is a level-1 (resp., level-2, ...) success if the first element of this list is the correct pseudonym (resp., the correct pseudonym is among the first two elements of the list, ...). Unless mentioned otherwise, we assume level-1 successes in the next experimental sections.

Re-identification metrics. One important feature of our threat model is that it corresponds to a continuous attack. That is, it is possible for the adversary to obtain multiple leakages corresponding to the same target user and to combine these leakages. As already mentioned, the attack will therefore depend on two main quantities: the amount of collected data $N_{\boldsymbol{r}}^i$ and the number of leakages $M_{\boldsymbol{r}}^i$. Therefore, a first natural metric would be to compute the success rate of the re-identification attack over randomly generated leakages. For each user u_i, this corresponds to a probability of level-l success $\mathrm{SR}_l^i(N_{\boldsymbol{r}}^i, M_{\boldsymbol{r}}^i)$. Note that it is natural to consider the probability of success over randomly generated leakages in our setting since (i) the leakage generation process is typically not under adversarial control, and (ii) what we want to measure is the number of leakages leading to a high success rate. One can also consider the average success rate (over all the users), which we denote as $\mathrm{SR}_l(N_{\boldsymbol{r}}, M_{\boldsymbol{r}})$.

A limitation of the previous metric is that it is intensive to estimate, since it requires building 3-dimensional "evaluation plots" with the amount of collected data $N_{\boldsymbol{r}}^i$ as X axis, the number of leakages $M_{\boldsymbol{r}}^i$ as Y axis and the success rate as Z axis. As a result, and inspired by results in side-channel analysis, we will consider easier-to-estimate information theoretic metrics based on the mutual information between the target random variable U corresponding to the users and a random variable D corresponding to (discretized) routes. As shown in [8], in case the observations can be viewed as "noisy leakages", computing this metric in function of the amount of collected data is an excellent predictor of the attack data complexity (i.e., the number of leakages needed to reach a given success rate). It allows simplifying the evaluations to 2-dimensional plots, with the amount of collected data as X axis & the information theoretic metric as Y axis.

In this respect, the final difficulty to compute the metrics relates to the fact that in practice, neither the evaluator of a database nor the adversary trying to exploit it know the true users's distributions from which the data originates (i.e., the PDFs $\mathsf{f}(\boldsymbol{r}|u_i)$ or the PMFs $\mathsf{g}(\boldsymbol{r}|u_i)$ defined in Section 2). The only thing that can be analyzed and exploited is the sampled data $\mathcal{D}_i$. As already mentioned, this is where the two types of (internal and external ones) leakages considered and the two types of estimations described in Section 2 come into play and allow defining metrics that can capture these two contexts.

Internal leakages evaluation. In this setting, we use the direct estimation process of Section 2 to obtain discrete models $\tilde{\mathsf{g}}(\boldsymbol{d}|u_i)$ and the probabilities

$\tilde{\Pr}[u_i|\boldsymbol{d}]$ and compute the Hypothetical Information (HI):

$$\tilde{\mathrm{HI}}(U;D) = \mathrm{H}[U] + \sum_{u_i \in \mathcal{U}} \Pr[u_i] \cdot \sum_{\boldsymbol{d} \in \mathcal{D}} \tilde{\mathsf{g}}(\boldsymbol{d}|u_i) \cdot \log_2 \tilde{\Pr}[u_i|\boldsymbol{d}],$$

where we use the short notation $\Pr[X = x] := \Pr[x]$ and we assume uniform users $\big($i.e., $\mathrm{H}[U] = \log_2(n)\big)$. As discussed in [9], the HI corresponds to the amount of information that would be extracted from the observations of (hypothetical) users behaving exactly according to the models $\tilde{\mathsf{g}}(\boldsymbol{d}|u_i)$. The higher the HI, the more different are the model distributions of the users and the easier the re-identification attack will be. Intuitively, the word hypothetical is used to make explicit that we do not know the true users' distributions. Concretely, it corresponds to the context of internal leakages since in this case, the collected data exactly defines the PDFs and PMFs that the adversary exploits. Note that when considering internal leakages, the adversary has no incentive to exploit a simplified model (since this model is guaranteed to be correct by definition). Hence, in the following, we will only consider the HI with exhaustive model.

External leakages evaluation. In this setting, the situation significantly differs since in order to be successful, the adversary has to build from the collected data a model that can be used to predict fresh routes. In order to capture this goal, we will therefore use the cross–validation estimation process of Section 2. Indeed, it typically reflects situations where a part of the observations are used to build a model that is then tested with another part of the observations (which actually corresponds to the leakages in an actual attack).

Importantly, successful attacks in this setting require that the true distributions are stationary to some extent (i.e., that $\mathsf{f}(\boldsymbol{d}|u_i)$ or $\mathsf{g}(\boldsymbol{d}|u_i)$ do not vary too much over time), so we need a metric that captures this requirement. For this purpose, we compute the Perceived Information (PI) as described next. First, we use the cross–validated estimation process in order to generate the models and test samples as in Section 2:

$$\left\{\hat{\mathsf{g}}^{(1:k)}(\boldsymbol{d}|u_i), \mathcal{T}_i^{(1:k)}\right\} \xleftarrow{\text{cv}} \mathcal{D}_i,$$

where any model $\hat{\mathsf{g}}^{(j)}(\boldsymbol{d}|u_i)$ can be used to define probabilities $\hat{\Pr}^{(j)}[u_i|\boldsymbol{d}]$. We next evaluate the models by using the test samples and deriving the estimates:

$$\hat{\mathrm{PI}}^{(j)}(U;D) = \mathrm{H}[U] + \sum_{u_i \in \mathcal{U}} \Pr[u_i] \cdot \sum_{\boldsymbol{d}' \in \mathcal{T}_i^{(j)}} \frac{1}{|\mathcal{T}_i^{(j)}|} \cdot \log_2 \hat{\Pr}^{(j)}[u_i|\boldsymbol{d}']. \quad (1)$$

The latter equation actually corresponds to an estimation by sampling, where we assign a probability $\frac{1}{|\mathcal{T}_i^{(j)}|}$ to each test sample which (by definition) directly originates from the true user distribution. The k outputs of this process (for a

k–fold cross–validation) are finally averaged in order to obtain a more precise estimate $\hat{\mathrm{PI}}(U;D)$. As discussed in [10], the PI corresponds to the amount of information that is extracted from the observations of actual users, by using models that are potentially biased by estimation and assumption errors. The PI is related to the success rate of an adversary using the same models. Intuitively, a positive PI means that the collected data reflects the differences between the true users' distributions to some extent. Incidentally, it also means that the user's distributions have been somewhat stationary during the data collection. The more positive the PI, the more successful are re-identification attacks with external leakages. By contrast, a negative PI means that the collected data is not sufficiently reflective of the true user's distribution (i.e., the models built from the collected data are not sufficiently predictive, due to insufficient data, wrong assumptions or strong model drift). Eventually, the PI converges towards Shannon's standard definition of Mutual Information (MI) if the models $\hat{\mathbf{g}}(\boldsymbol{d}|u_i)$ are perfect (i.e., if they are equal to the true $\mathbf{g}(\boldsymbol{d}|u_i)$).

Note that when considering external leakages, the adversary has incentives to exploit simplified models (such as a 1st-order independent model). Indeed, the PI is a tradeoff between estimation and assumption errors, i.e., between the speed of convergence and the asymptotic informativeness of a model. This tradeoff will be discussed in the experimental sections.

Remark. As the amount of collected data increases, it reflects more and more the true users' distributions. Hence, if both the HI and the PI metrics are estimated with the same model family and based on data discretized with the same function (and Δ), they both converge towards the same value. In case an exhaustive model is used, they additionally converge towards the true mutual information $\mathrm{MI}(U;D)$. Based on this intuition, it was demonstrated in [3] that the HI is (in expectation) an upper bound of the MI and PI.

Besides, we note that the HI and PI metrics are averages over the users. Yet, it can happen that the success rate of the re-identification attacks highly vary in function of the users. The latter is easily analyzed by computing those metrics for fixed users, i.e., by evaluating $\tilde{\mathrm{HI}}(U=u;D)$ or $\hat{\mathrm{PI}}(U=u;D)$.

Links to other metrics. We now discuss the links between the HI and PI metrics and related metrics introduced in the privacy literature.

***k*-anonymity and related metrics.** The re-identification threat model *with internal leakages* (where the collected data is available to the adversary) shares similarities with the problem of privacy-preserving data publishing for which various metrics have been introduced, some of them surveyed in [12]. In this setting, the k-anonymity is among the simplest (and most popular) solutions [17]. Informally, k-anonymity guarantees that a leakage does not allow to (strictly) distinguish (i.e., with probability one) a user from at least $k-1$ other users. As discussed in [14], this may not be enough to prevent all types of linking attacks. A typical example is the case where all the users that remain indistinguishable have the same "sensitive data" that the linking attack aims to recover (i.e., referred to as "other data" in Figure 3). In other words, a lack of diversity in

the sensitive data may allow the adversary to deduce private information for a database that ensures k-anonymity. The main limitation of the k-anonymity and refinements such as the l-diversity in our context is that they only consider the strict indistinguishability of the users, and ignore the possibility that the list of users for which a leakage is possible may have different conditional probabilities. The latter possibility is particularly relevant in a continuous attack setting (since these probabilities may be combined in maximum likelihood attacks).

Other information theoretic metrics. In order to mitigate the previous limitation, a usual solution is to consider information theoretic metrics, for example such as the anonymity degree introduced by Diaz et al. [7] or variations thereof [18]. The anonymity degree can be viewed as similar to the HI metric, since it is also computed from a model built thanks to a direct estimation process. The only differences are that (i) the anonymity degree considers the normalized conditional entropy rather than the mutual information for the HI (hence its link with the success rate is less direct), and (ii) as already mentioned, the HI metric makes explicit that it is based on a hypothetical model.

Location privacy metrics. By contrast, neither the k-anonymity nor the anonymity degree can be used to evaluate re-identification attacks *with external leakages*. This was argued in a paper by Shokri et al. [19] (yet, for a different attack goal than re-identification, namely the localization attacks that we discuss in Section 5). The authors identified three types of metrics (namely the uncertainty, accuracy and correctness) and argued that correctness is the appropriate way to quantify attacks aiming at predicting new events. Informally, the correctness is correlated to the probability of error of an adversary trying to predict fresh positions. In our re-identification context, it therefore captures a similar intuition as the success rate of an attack exploiting external leakages, the data complexity of which being itself correlated with the PI metric. The accuracy then measures the convergence of the model estimate and can be analyzed based on the convergence of the PI metric (thanks to estimation plots or confidence intervals). We will argue next that it is also relevant to the evaluation of re-identification attacks with external leakages. Eventually, the uncertainty corresponds to the HI metric and is indeed irrelevant to analyze attacks using external leakages. Even closer to our framework, the attacks described in [15] consider re-identification with external leakages quantified with a re-identification rate (which is directly equivalent to our success rate).

Unicity. In yet another line of papers, de Montjoye et al. introduced the concept of unicity, which captures re-identification attacks based on location data [5, 6]. Their analysis uses a direct estimation process and is therefore linked to the HI. Informally, assuming independent leakages so that each observation reduces $\mathrm{H}[U]$ by $\tilde{\mathrm{HI}}(U; D)$, unicity corresponds to a case where the number of leakages is such that users have no entropy left.

Differential privacy. Eventually, we mention that preventing re-identification attacks could be achieved thanks to differential privacy [11], yet in a different setting. Namely, in differential privacy, one aims to guarantee that a few queries

made to a database do not reveal private information. In our setting, we rather make this database fully available to the adversary and allow multiple leakages. As will be clear in the experimental sections, obtaining privacy in this setting is extremely challenging (if possible at all). Our quantitative tools can therefore be viewed as a motivation for differential privacy (or similar frameworks aiming at restricting the adversary's power in a relevant manner). See for example the discussion about geo-indistinguishability in [1, 16].

4 Experimental validation and discussion

The following experiments are based on 4 different data sets (all of them discretized thanks to the D_1 process). Our first data set is a simulated one where the space is discretized into $\Delta = 128$ cells and we generated 1000 routes for 5 users according to chosen distributions. The experiments additionally consider a more discretized process with only $\Delta = 16$ cells. This setting is only used in order to put forward the general intuitions of the metrics (since it allows generating sufficient number of observations from stable distributions so that all metrics perfectly converge). Our second data set comes from *Brightkite*, an application enabling to share visited places with friends. It provides global coordinates that we reduce in two steps [4]. First we only consider the San Francisco Area. Second, we discretize the space into $\Delta = 16$ cells. It then remained 302 users with at least 50 single-position routes.[2] Our third data set is based on *jogging records* obtained with smart watches. We followed 7 users with at least 100 routes, discretized in respectively 16, 32 and 121 cells.[3] Eventually, our last data set comes from the *BikeShare* stations (also in the San Francisco area), publicly available for a contest about data visualization.[4] We consider 27 groups of users (since the database has been anonymized by grouping users according to their ZIP code) and $\Delta = 33$ cells which correspond to different BikeShare stations.

We start by evaluating the simulated data set to put forward general intuitions that can be extracted from our framework and metrics. We then analyze the different real-life data sets and discuss their interpretation.

General metric intuitions. The metrics estimated from our simulated data set are given in Figure 4. The left plot reports information theoretic metrics in function of the number of collected routes per user N. The right plot reports the success rate in function of the number of (external) leakages obtained by the adversary M, for various N values. They lead to the next observations.

Starting with the IT metrics, a first noticeable fact is that the HI decreases with N while the PI increases with N. This is theoretically expected in both cases. For the HI, the reduction intuitively depends on the number of collisions in a data set (just as would be observed with the k-anonymity metric). As a result, larger numbers of routes per user N imply more risks of collisions for our

[2] `https://snap.stanford.edu/data/loc-brightkite.html` (4/2008 - 10/2010).
[3] This data set is not publicly available (1/2010 - 2/2016).
[4] `https://www.fordgobike.com/system-data` (8/2013 - 8/2016).

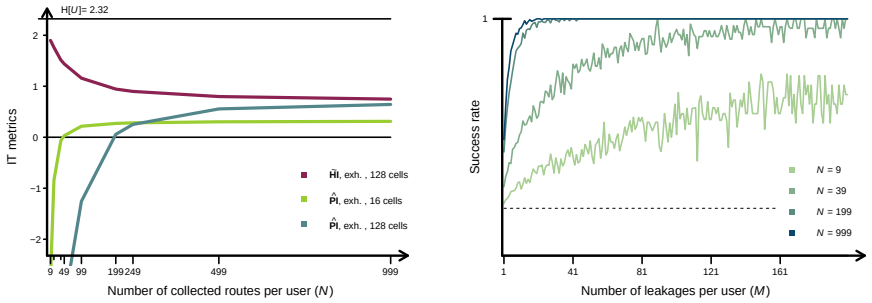

Fig. 4. Simulated data. Left: IT metrics. Right: success rate of attacks with external leakages (with the collected routes discretized in $\Delta = 128$ cells).

data set (where some routes are possible for all users). For the PI, it reflects the fact that by increasing N, one builds more accurate models, with less estimation errors, saturating when N is sufficient to perfectly estimate the model.

A second observation (specific to the PI and the evaluation of external leakages) is that the level of discretization brings a tradeoff between the speed of convergence of the model and the informativeness of the (fully characterized) model. For example, the model estimated with 16 cells converges faster (and is more rapidly informative) than the one with 128 cells, but less informative for $N = 999$. The latter suggests that the speed of convergence of a model is a relevant evaluation metric since it determines the amount of observations that a malicious adversary would require to build a database that is sufficient to infer something about a user. This could for example be useful in a (non open data) scenario where routes are maliciously collected.

Eventually, a third observation is that since we only consider exhaustive models in this simulated setting, the HI and the PI with 128 cells converge towards the same value (equal to the MI), as expected [3].

An open source tool in R language allowing the generation of HI/PI plots to confirm these intuitions is available in complement to this work.

As for the success rate plots, we first observe that the number of leakages M required to reach high success rates is proportional to the value of the PI for the corresponding N value, as theoretically predicted in [8]. This makes it an interesting alternative to the success rate since it is typically easier to sample, as reflected by the less smooth success rate curves when M increases (since there are less sets of M traces based on which we can estimate the success rate). Also, the PI plots allow easy comparisons between two models (e.g., the ones with 16 and 128 cells) in order to determine the number of collected traces such that one becomes more informative than the other, which happens when their respective PI curves intersect. Besides, we also note that the number of leakages M required to reach a high success rate is usually lower than the number of routes that must be collected to build an informative model. That is, building

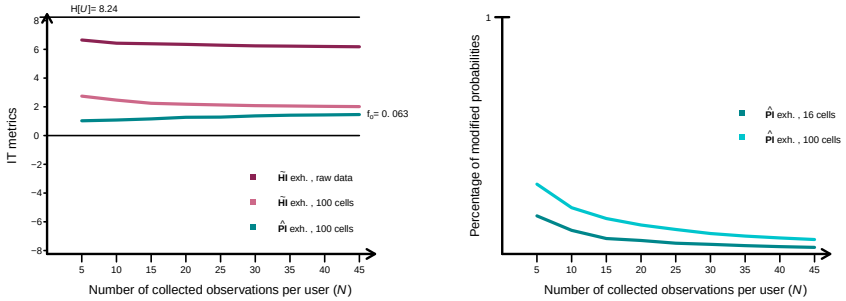

Fig. 5. Left: Brightkite data, IT analysis. Right: outliers correction

a model enabling statistical inference attacks with external leakages is more difficult than mounting the attack once a well estimated model is known.

Brightkite data set. We use this second data set to illustrate the more challenging nature of real observations from the privacy viewpoint. The latter can be confirmed from the IT metrics estimations in Figure 5, and the large gap existing between the HI computed from the raw data (from which no positive PI could be extracted) and the HI and PI computed after discretization. It reflects the general fact that location privacy rarely comes for free (i.e., without sanitization). For example, the high HI value for the raw data suggests that re-identification with internal leakages is trivial (i.e., successful after a couple of raw leakages). Since this value bounds the PI for a larger database, it means that with the amount of collected data, we can only bound the risks of re-identification with external leakages very conservatively with the raw HI value. We also notice that by discretizing the data we can reduce the HI while enabling the estimation of predictive models, as witnessed by the PI curve that converges to positive values. The latter values ($> \frac{1}{10}$) suggest that a handful of external leakages is sufficient for re-identification. Indeed, a simple bound for the re-identification complexity is given by $\frac{c}{\text{PI}}$ [8].[5] Stronger discretizations (e.g., with 16 cells) show a faster convergence of the PI at the cost of a reduction of its asymptotic value (i.e., a speed of convergence vs. informativeness tradeoff).

Another observation of interest is that in the case of real data, it frequently happens that the modeling phase is made more difficult due to outliers (i.e., user's observations that only happen rarely, possibly once). Those can lead to prohibitively low probabilities (e.g., zero probabilities that make the estimation of the PI impossible). We deal with these outliers by "correcting" the observations for which the probability is lower than $\frac{1}{N}$ and setting them to this minimum value, while also counting the fraction of corrected probabilities. This fraction f_o is always given for the final value of our PI estimates in the figures (e.g., $f_o = 0.063$ in the left part of Figure 5). As illustrated the right part of the figure, it generally decreases with the size of the profiling set.

[5] With c a small constant depending on H[U] and the target success rate (e.g., $c = \text{H[U]}$ is a usual heuristic that corresponds to a success rate of approximately 80%).

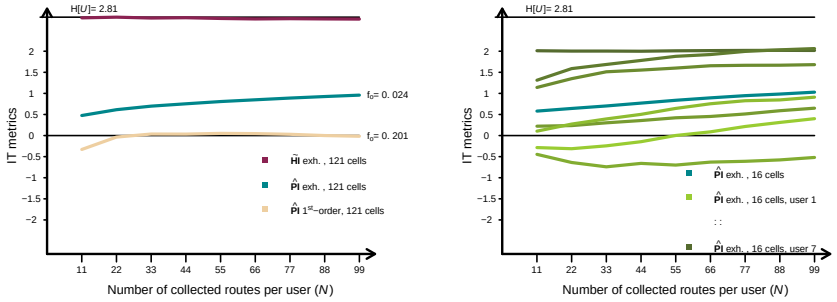

Fig. 10. Jogging data. Left: IT analysis. Right: IT analysis per user.

Jogging data set. We use this third data set and the results in (the left part of) Figure 10 to illustrate a case where a simplified modeling exploiting an independence assumption does not allow a better model convergence. The figure clearly shows that the PI extracted by using an exhaustive model (with sufficiently discretized routes) is significantly higher than the PI extracted by using a 1st-order independent model. It corresponds to the intuition that in the case of jogging data (e.g., the ones in right part of Figure 1), the consecutive observations of a route are highly correlated and therefore an independence assumption during the modeling is unlikely to bring any significant gain.

Another observation of interest in this context is that the routes of the joggers we analyzed are (on average) very discriminating, both with respect to internal leakages (with an HI value stuck to $\mathrm{H}[U]$) and external leakages (with a PI value close to $\frac{\mathrm{H}[U]}{4}$). We further use this context to put forward the differences that can occur in the characterization of the models for different users, as illustrated in the right part of Figure 10. The latter recalls that the HI and PI are average metrics and are handy to have a quick "privacy overview" of a data set. However, a rigorous analysis of the leakages' informativeness has to be performed at the user level. For example, the aforementioned connection between information theoretic metrics and the success rate only holds per user [8].

BikeShare data set. We finally use this fourth data set to illustrate a case where a simplified modeling exploiting an independence assumption is needed for re-identification with external leakages, and to discuss the addition of the time component of the observation's in our reasoning.

Starting with the modeling issue, we first provide some additional intuition about the BikeShare data set based on Figure 18 (given in Appendix B). It represents the daily usage of different bike stations by different users with different ZIP codes (i.e., 100% means the station is used everyday by the user). More precisely, it corresponds to the daily usage of sets of users living in the same area, which have been grouped in order to preserve their anonymity. We will denote these sets as users for simplicity. One can clearly see that depending on the area a user lives in, his most used BikeShare stations vary significantly.

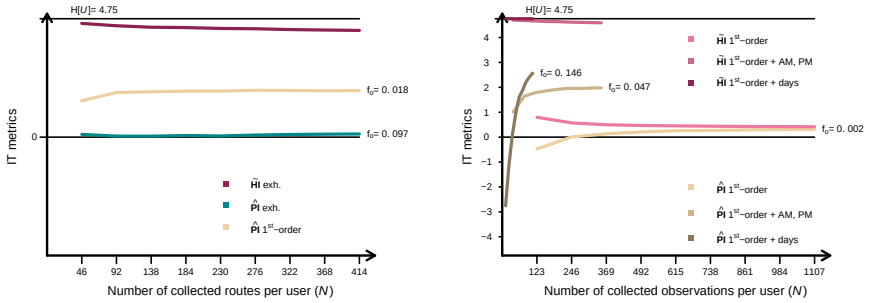

Fig. 13. BikeShare data. Left: IT analysis. Right: IT analysis with time component.

We performed the same IT analysis as in the previous sections and report it in Figure 13. This time, and in contrast with the jogging case study, the 1st-order independence assumption is needed to build a predictive model allowing re-identification with external leakages. This result can be explained by considering the nature of the data represented in Figure 18. Namely, contrary to the case of the jogging data where the consecutive observations in a route are very correlated, the use of BikeShare can be interleaved with other public or private transports, hence creating a good level of independence between the observations in a route. Based on this conclusion, our following analyzes of the BikeShare data set will systematically split all the routes into several independent observations. Note that contrary to the case of jogging data, the definition of a route is more difficult in the BikeShare context. In Figure 13 we arbitrarily defined it as the consecutive observations of one day.

This analysis leads to important conclusions from a risk assessment viewpoint. Indeed, it highlights that the possibility to mount re-identification attacks exploiting external leakages against the privacy of some users in a database does not only depend on the amount of data collected but also on the assumptions that an adversary can make about them. In this respect, the possibility to bound this risk thanks to the HI estimated with an exhaustive model is a useful tool for privacy assessments. As already observed, for many real-world data sets, this bound is unlikely to be tight due to a lack of data (since for an infinite amount of data, it is proven to be tight). For example, in the simple case of Figure 13, we tested a 1st-order independent model which leads to a better PI than the exhaustive model, but still falls far away from the HI bound. This gap captures the risk that some non-obvious assumption about the data set (or simply more data in case a malicious database owner is hiding a part of the collected data) would significantly improve the model informativeness and/or convergence: combined with the previous experimental observation that building a model is usually more data consuming than exploiting it, it implies that the risks of statistical inference attacks are in general hard to bound tightly, unless some specific mechanisms prevent the unrestricted use of the data.

Including the time component. We conclude the paper by showing that the time component of location observations can be used to further improve the attacks with external leakages. This fact is illustrated by the information theoretic analysis in the right part of Figure 13 where the PI is estimated with and without time component, considering two granularities: AM/PM (in which case the total number of bins of the independent model is doubled) and daily (in which case this total number of bins is multiplied by seven).

Two preliminary remarks resulting from the figure are: (i) that the value of the PI without time component in the left part of Figure 13 is reduced compared to the one in the right plot. The latter derives from the fact that we now estimate $\mathrm{PI}(U;P)$ (since, as mentioned earlier, we split all the routes in independent observations) rather than $\mathrm{PI}(U;D)$, and routes contain several positions: roughly, the average number of observations per route can be approximated by the ratio between the two PI values on these figures; and (ii) that the maximum size of the profiling set decreases when considering the time component (since we now need a sufficient amount of observations for all time values).

The figure highlights the significant gain of information that is obtained by characterizing the time component of the users' observations, hence revealing that their biking habits differ depending on the days and time of the days. It also confirms the aforementioned fact that the HI bound becomes tighter when a large database with more observations is available.[6]

5 Localization attacks

As a complement to our previous discussions, we next describe how a second threat model, where the adversary's goal is to predict the position of a user, can be captured in a similar framework as for the re-identification threat model. Such so-called localization attacks, depicted in Figure 14, are correspond to the ones discussed in [19] that we (re)formalize as follows.

5.1 Threat model and metrics

First, we again have a number of users for which a number of positions have been collected. For simplicity, we now consider positions rather than routes since the (basic) adversarial goal in a localization attack is to predict such locations. Yet, one could also consider more elaborated attacks trying to predict a sequence of (possibly correlated) locations (called tracking attacks in [19]) which would then require to consider routes as in the previous section. Second, we consider challenge observations that the adversary tries to predict and correspond to fresh positions of a target user. Based on this setup, the goal of the adversary is to predict the fresh position thanks to a model built from his collected data. This threat model shares similarities with the re-identification attack based on

[6] Note that the bound is here given for 1st-order independent models, as shown in the left part of the figure, the bound for the exshaustive models is stuck at $\mathrm{H}[U]$.

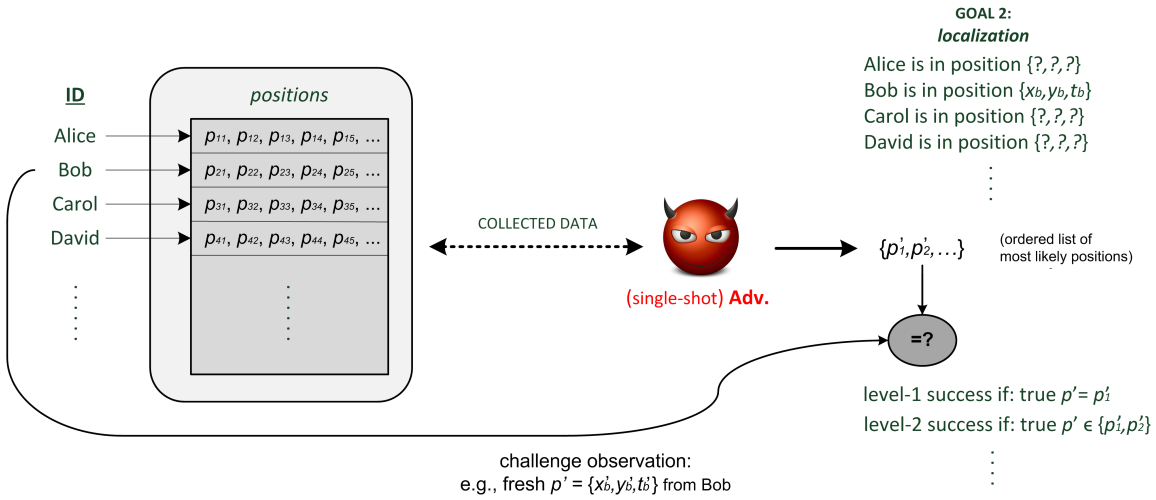

Fig. 14. Localization attack threat model.

external leakages since here as well, the adversary's success depends on how well the collected data represents the true users's distributions. Yet, it has a different (prediction) goal and target random variable (i.e., positions rather than user IDs for re-identification attacks). It also differs due to its "single-shot" nature (rather than continuous for re-identification attacks): while it is possible to continuously accumulate leakages for the same target user in a re-identification attack, the target localization is fresh in each challenge of the localization attack.

Localization metrics. Due to its single-shot nature, this second threat model is easier to evaluate. Namely, it is now possible to compute the success rates $\hat{\mathrm{SR}}_l^i(N_{\boldsymbol{r}}^i, 1)$ only in function of the amount of collected data $N_{\boldsymbol{r}}^i$ (since the $M_{\boldsymbol{r}}^i$ leakages are turned into single "challenge positions"). We will use these success rates (or their average $\hat{\mathrm{SR}}_l(N_{\boldsymbol{r}}, 1)$) as our main metric. Note that the fact that attacks are not continuous does not prevent the adversary to try succeeding on multiple independent challenges, leading to a multi-challenge success rate:

$$\hat{\mathrm{SR}}_l(N_{\boldsymbol{r}}, M) = 1 - (1 - \mathrm{SR}_l(N_{\boldsymbol{r}}, 1))^M.$$

As previously, we define a level-1 success as the situation where the challenge position corresponds to the most likely position output by the adversary, a level-2 success when it is among the first two positions, ...

Links to other metrics As mentioned in Section 3, metrics to quantify the localization attack threat model have been introduced in [19]. The SR can be used as an alternative to their main (correctness) metric. Besides, the convergence of the SR metric can be used as an alternative to their accuracy metric (which is also relevant to the evaluation of localization attacks). Similarly, the concrete efficiency of the localization attacks described in [13] is assessed thanks to another accuracy metric which directly corresponds to our success rate.[7]

5.2 Experimental validation and discussion

We use two data sets from the experiments in Section 4 to illustrate our second threat model. We start by discussing its specificities and then exhibit general conclusions that can be extracted from our framework.

BikeShare data set. We first recall that the goal of a localization attack is to predict the position of a user. So for the attack to make sense in practice, it is necessary to include the time in our reasoning, since the time an adversary needs to wait at a certain position to meet some user with good probability is part of the risk analysis. Furthermore, and taking the example of our previous

[7] Localization attacks are computationally faster than re-identification attacks with external leakages since they target users independently and exploit the conditional distributions $\hat{\mathbf{g}}(\boldsymbol{p}|u_i)$ directly, while re-identification attacks need to compute $\hat{\Pr}[u_i|\boldsymbol{p}]$ via Bayes. Note also that a user can be localized even if $\mathrm{MI}(U;P) = 0$ (e.g., if all users have identical conditional distributions that are not uniform over the positions, which would then be reflected by the entropy of the observations $\mathrm{H}[P]$).

time division in AM/PM, it is of course not guaranteed that the data set contains observations every morning and afternoon. For example, the left part of Figure 15 shows the "daily coverage" of the BikeShare data set (defined as the percentage of days for which at least one station is visited). In order to deal with these important features of the threat model, our following experiments will systematically compute the "single-shot" success rate metric for various levels of success l and various number of observations per user N. The level of the success is relevant to capture the possibility that an adversary uses a team of people to intercept his target user (with each team member waiting at a different place). Furthermore, we will add the coverage of the attack (e.g., for the AM/PM attack, the percentage of mornings and afternoons with data collected, for the daily attack, the percentage of days with data collected). The product of the coverage and the level-l success rate gives a lower bound for the concrete feasibility of a localization attack (the better the coverage, the tighter the bound).

The results of a localization attack in the afternoons (with 75% coverage) are in the right part of Figure 15. One can see that for all levels l, the success rate of the attack is significantly better than a random guess, with success rate close to one after $l = 15$ (which concretely means that a team of 15 people has high probability to meet a target user for 75% of the afternoons). We also notice that increasing the size of the data set from $N = 30$ to $N = 270$ does not lead to a major improvement of the success rate, suggesting that the model has converged (i.e., the fact that success rate curves are not improved by increasing N carries the same intuition as the saturation of the PI curves). A similar plot is given in the left part of Figure 16 for the Thursdays. The more limited number of observations per day is exhibited by a smaller maximum value for the number of observations per user N and a more visible gain when N grows.

Jogging data set We complete our localization experiments by showing one example of attack against the jogging data set in the right part of Figure 16. The main observation here is that despite the smaller data set (we could only collect sufficient amounts of routes for four users out of the seven), the data leads to

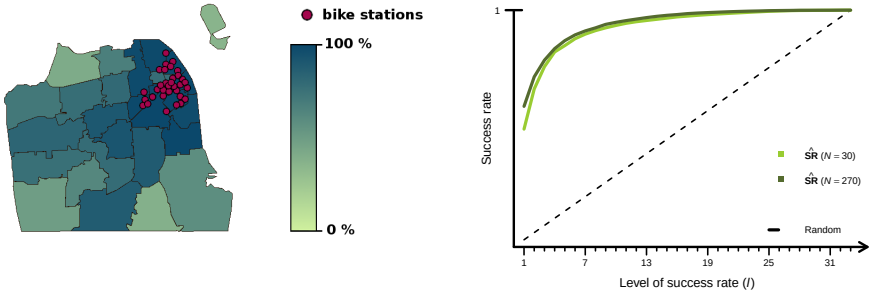

Fig. 15. BikeShare data. Left: "daily coverage" per user / ZIP code (i.e., % of days for which at least one station is visited). Right: success rate on PMs with 75% coverage.

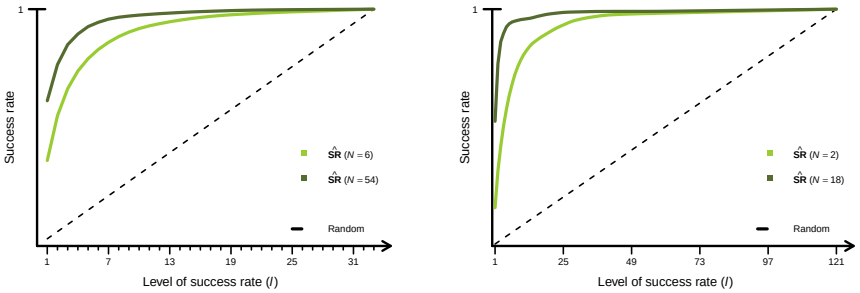

Fig. 16. Left: BikeShare data – success rate on Thursdays with a coverage of 88%. Right: jogging data – success rate on Tuesdays with a coverageof 23% (4 users)

easier predictions (e.g., for the 4 users considered, the success rate gets close to one for $l = 3$). As for the BikeShare data, we considered a simple adversarial strategy waiting for the target user at the l most likely positions given by a 1st-order independent model. However, in this case more advanced strategies could be considered (e.g., one could estimate the probability that the user is either at position A or position B by capturing higher-order correlations in the model). We leave such investigations as a scope for further research.

Acknowledgments. François-Xavier Standaert is a Senior Research Associate of the Belgian Fund for Scientific Research (FNRS-F.R.S.). This work has been funded in parts by the ERC project SWORD (Consolidator Grant 724725).

References

1. Miguel E. Andrés, Nicolás Emilio Bordenabe, Konstantinos Chatzikokolakis, and Catuscia Palamidessi. Geo-indistinguishability: differential privacy for location-based systems. In Ahmad-Reza Sadeghi, Virgil D. Gligor, and Moti Yung, editors, *ACM SIGSAC*, pages 901–914. ACM, 2013.
2. Alastair R. Beresford and Frank Stajano. Location privacy in pervasive computing. *IEEE Pervasive Computing*, 2(1):46–55, 2003.
3. Olivier Bronchain, Julien M. Hendrickx, Clément Massart, Alex Olshevsky, and François-Xavier Standaert. Leakage certification revisited: Bounding model errors in side-channel security evaluations. *IACR Cryptology ePrint Archive*, 2019:132, 2019.
4. Eunjoon Cho, Seth A. Myers, and Jure Leskovec. Friendship and mobility: user movement in location-based social networks. In Chid Apté, Joydeep Ghosh, and Padhraic Smyth, editors, *ACM SIGKDD*, pages 1082–1090. ACM, 2011.
5. Yves-Alexandre de Montjoye, César A Hidalgo, Michel Verleysen, and Vincent Blondel. Unique in the crowd: The privacy bounds of human mobility. *Nature Scientific reports*, 3(1376):5, 2013.
6. Yves-Alexandre de Montjoye, Laura Radaelli, Vivek Kumar Singh, and Alex Sandy Pentland. Unique in the shopping mall: On the reidentifiability of credit card metadata. *Science*, 347(6221):536–539, 2015.

7. Claudia Díaz, Stefaan Seys, Joris Claessens, and Bart Preneel. Towards measuring anonymity. In Roger Dingledine and Paul F. Syverson, editors, *PET*, volume 2482 of *LNCS*, pages 54–68. Springer, 2002.
8. Alexandre Duc, Sebastian Faust, and François-Xavier Standaert. Making masking security proofs concrete - or how to evaluate the security of any leaking device. In Elisabeth Oswald and Marc Fischlin, editors, *EUROCRYPT*, volume 9056 of *LNCS*, pages 401–429. Springer, 2015.
9. François Durvaux, François-Xavier Standaert, and Santos Merino Del Pozo. Towards easy leakage certification: extended version. *J. Cryptographic Engineering*, 7(2):129–147, 2017.
10. François Durvaux, François-Xavier Standaert, and Nicolas Veyrat-Charvillon. How to certify the leakage of a chip? In Phong Q. Nguyen and Elisabeth Oswald, editors, *EUROCRYPT*, volume 8441 of *LNCS*, pages 459–476. Springer, 2014.
11. Cynthia Dwork. Differential privacy: A survey of results. In Manindra Agrawal, Ding-Zhu Du, Zhenhua Duan, and Angsheng Li, editors, *TAMC*, volume 4978 of *LNCS*, pages 1–19. Springer, 2008.
12. Benjamin C. M. Fung, Ke Wang, Rui Chen, and Philip S. Yu. Privacy-preserving data publishing: A survey of recent developments. *ACM Comput. Surv.*, 42(4):14:1–14:53, 2010.
13. Sébastien Gambs, Marc-Olivier Killijian, and Miguel Nunez del Prado Cortez. Next place prediction using mobility markov chains. In *Proceedings of the First Workshop on Measurement, Privacy, and Mobility*, MPM '12, pages 3:1–3:6, 2012.
14. Ashwin Machanavajjhala, Daniel Kifer, Johannes Gehrke, and Muthuramakrishnan Venkitasubramaniam. *L*-diversity: Privacy beyond *k*-anonymity. *TKDD*, 1(1):3, 2007.
15. Mohamed Maouche, Sonia Ben Mokhtar, and Sara Bouchenak. Ap-attack: A novel re-identification attack on mobility datasets. In Dali Kaafar and Gang Zhou, editors, *MobiQuitous*, pages ?–? ACM, 2017.
16. Simon Oya, Carmela Troncoso, and Fernando Pérez-González. Is geo-indistinguishability what you are looking for? In Bhavani M. Thuraisingham and Adam J. Lee, editors, *Proceedings of the 2017 on Workshop on Privacy in the Electronic Society*, pages 137–140. ACM, 2017.
17. Pierangela Samarati and Latanya Sweeney. Generalizing data to provide anonymity when disclosing information (abstract). In Alberto O. Mendelzon and Jan Paredaens, editors, *ACM SIGACT-SIGMOD-SIGART*, page 188. ACM Press, 1998.
18. Andrei Serjantov and George Danezis. Towards an information theoretic metric for anonymity. In Roger Dingledine and Paul F. Syverson, editors, *PET*, volume 2482 of *LNCS*, pages 41–53. Springer, 2002.
19. Reza Shokri, George Theodorakopoulos, Jean-Yves Le Boudec, and Jean-Pierre Hubaux. IEEE s&p. pages 247–262. IEEE Computer Society, 2011.
20. François-Xavier Standaert, Tal Malkin, and Moti Yung. A unified framework for the analysis of side-channel key recovery attacks. In Antoine Joux, editor, *EUROCRYPT 2009*, volume 5479 of *LNCS*, pages 443–461. Springer, 2009.

A An extended setup

It is easy to see the pros and cons of the two proposed estimation tools in Section 2. The exhaustive model potentially captures all the correlations within different cells, but due to the large cardinality of its support it potentially converges very slowly. By contrast, the 1st-order independent model does not capture any correlation between the cells, but due to its simplicity it potentially converges very fast. It is also easy to see that ignoring correlations can be detrimental to the understanding of location privacy.

Consider two users whose routes are both starting from point A and ending at point B in the left part of Figure 17. The space of these routes is discretized in 4 cells and only 4 (discretized) routes are possible, namely $\{1,3\}$, $\{1,4\}$, $\{2,3\}$ and $\{2,4\}$. Now assume that the first user only takes routes $\{1,3\}$ and $\{2,4\}$ – each of them with probability $1/2$, and the second user only takes routes $\{1,4\}$ and $\{2,3\}$ – each of them with probability $1/2$. Then, the 1st-order independent model is identical for both users, and does not enable re-identification.

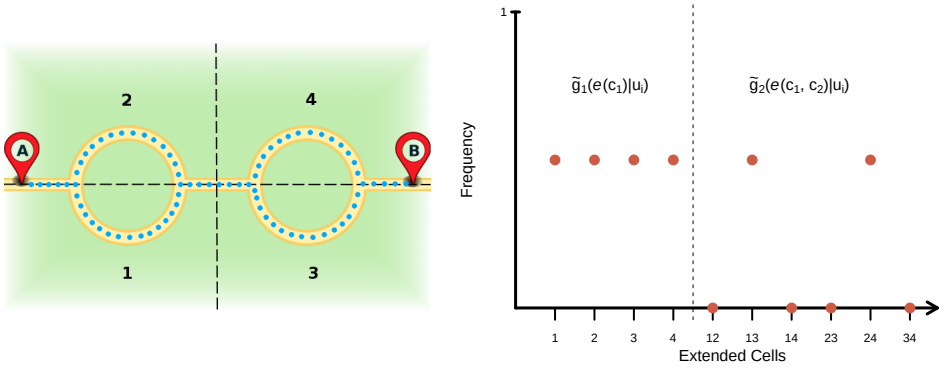

Fig. 17. Left: example of (correlated) routes. Right: 2nd-order indep. model.

We next detail how a 1st-order independent model can be generalized to an *oth-order independent model* which can capture higher-order correlations in the distribution by first "extending" the data towards higher-orders and then using estimation tools similar to the ones described Section 2. In this respect, we insist that while it may sound counter-productive to combine a discretization process with an extension one, it may happen – at least theoretically – that reducing Δ and increasing o is a better (adversarial) strategy than just increasing Δ. Note also that a oth-order independent model is never equivalent to the exhaustive model since the knowledge of a statistical distribution (characterized by the exhaustive model) is not equivalent to the knowledge of its moments. So while increasing o can be used to characterize higher-order dependencies of the user's behavior, it cannot lead to an optimal model.

A.1 Extended data specification

In order to capture the possible correlations between several cells without directly moving to the (usually more expensive to estimate) exhaustive model, one option is to consider "extended routes" of which the support moves from all possible cells to all possible pairs of cells (or triples of cells, ...). For this purpose, we define a oth-order extension function (for $1 \leq o \leq \Delta$) as:

$$\mathsf{E}_o : \{0,1\}^{\Delta} \rightarrow \{0,1\}^{\binom{\Delta}{1}+\binom{\Delta}{2}+\ldots+\binom{\Delta}{o}},$$

where the $\binom{\Delta}{1}$ exponent corresponds to the cells, as already considered in Section 2, the $\binom{\Delta}{2}$ one to pairs of cells, and the $\binom{\Delta}{o}$ one to o-tuples of cells.

In the setup of this section, it is computed by assigning a one to each o-tuple of cells such that at least one observation of the route falls in each of its o cells. In other words, the value of each "extended cell" corresponding to an o-tuple of cells is the product of the individual cell values. We next denote such oth-order extended routes as:

$$\boldsymbol{e} = \mathsf{E}_o(\boldsymbol{d}).$$

A.2 Estimation tools

One advantage of the previous extension is that it can be plugged into the second estimation tool of Section 2. We refer to the models estimated by first extending the routes to exploit correlations at order up to o and relying on an independence assumption for the correlations at order larger than o as *oth-order independent models*. For example, a 2nd-order independent model for the discrete and (e.g.,) direct estimation case, that we denote as $\tilde{\mathsf{g}}_{1:2}(\boldsymbol{e}|u_i)$, is illustrated in the right part of Figure 17 and already allows detecting user-specific features for our 4-route example. This model corresponds to the concatenation of the previous 1st-order independent model $\tilde{\mathsf{g}}_1(\boldsymbol{d}|u_i) := \tilde{\mathsf{g}}_1(\boldsymbol{e}(c_1)|u_i)$ and the independent estimation of $\tilde{\mathsf{g}}_2(\boldsymbol{e}(c_1,c_2)|u_i)$ for each of the $\binom{\Delta}{2}$ pairs of cells, with $\boldsymbol{e}(c_1,c_2)$ a pair of cells of the discretized route $\boldsymbol{d}$ (that we next denote as an extended cell for short).

The probability of a user u_i given an extended route $\boldsymbol{e}$ capturing correlations at order up to 2, next denoted as $\tilde{\Pr}_{1:2}[u_i|\boldsymbol{e}]$, is then computed as previously. Namely, we first compute the 2nd-order likelihood as:

$$\begin{aligned}\tilde{\mathsf{q}}_2(\boldsymbol{e}|u_i) = &\prod_{c_1,c_2 \in \boldsymbol{e}} \tilde{\mathsf{g}}_2[\boldsymbol{e}(c_1,c_2)|u_i] \\ &\cdot \prod_{c_1,c_2 \notin \boldsymbol{e}} \left(1 - \tilde{\mathsf{g}}_2[\boldsymbol{e}(c_1,c_2)|u_i]\right),\end{aligned}$$

where $c_1, c_2 \in \boldsymbol{e}$ this times denotes the extended cells that are part of the extended route $\boldsymbol{e}$ and $c_1, c_2 \notin \boldsymbol{e}$ the ones that are not in the extended route. The probability $\tilde{\Pr}_{1:2}[u_i|\boldsymbol{e}]$ is finally derived thanks to Bayes' formula:

$$\tilde{\Pr}_{1:2}[u_i|\boldsymbol{e}] = \frac{\tilde{\mathsf{q}}_1(\boldsymbol{e}|u_i) \cdot \tilde{\mathsf{q}}_2(\boldsymbol{e}|u_i)}{\sum_{j=1}^{n} \tilde{\mathsf{q}}_1(\boldsymbol{e}|u_j) \cdot \tilde{\mathsf{q}}_2(\boldsymbol{e}|u_j)}.$$

Models capturing correlations up to larger orders are computed similarly by multiplying the likelihoods at each order, and denoted as $\tilde{\Pr}_{1:o}[u_i|\boldsymbol{e}]$. (One could also build models capturing only specific correlations at order o_1 and o_2, denoted as $\tilde{\Pr}_{o_1,o_2}[u_i|\boldsymbol{e}]$). Computing the probability of a user given an extended route rapidly becomes computationally expensive: for a discretized route $\boldsymbol{d}$ of size $N_{\boldsymbol{o}}$, there are $\binom{N_{\boldsymbol{o}}}{o}$ extended cells at order o.

B Additional figure

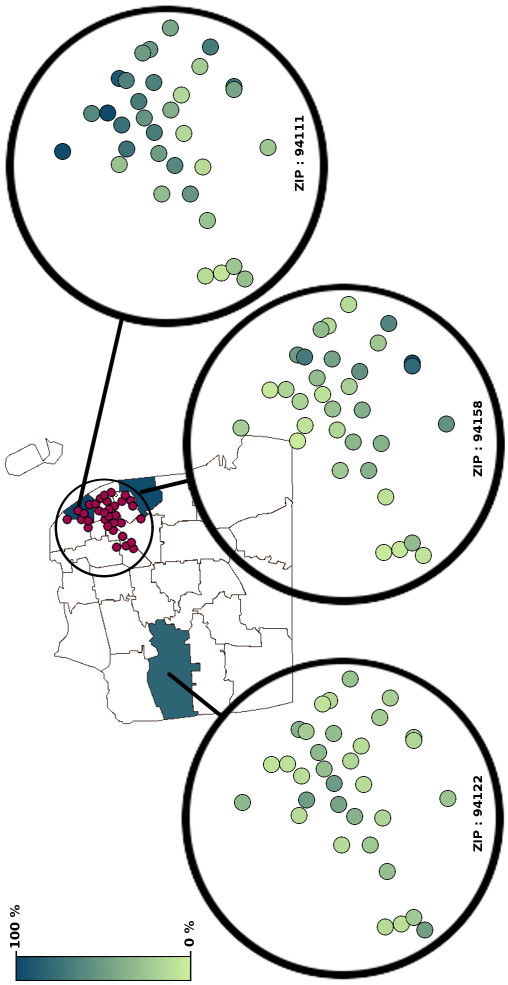

Fig. 18. Daily usage of BikeShare stations for three users (ZIP codes).