CORA Cork Open Research Archive
Cartlann Taighde Oscailte Chorcaí

Title	Two de-anonymization attacks on real-world location data based on a hidden Markov model
Authors	Eshun, Samuel N.;Palmieri, Paolo
Publication date	2022-06
Original Citation	Eshun, S. N. and Palmieri, P. (2022) 'Two de-anonymization attacks on real-world location data based on a hidden Markov model ', 7th IEEE European Symposium on Security and Privacy (IEEE EuroS&P 2022), Genoa, Italy, June 6-10, co-located Workshop Proceedings. https://doi.org/10.1109/EuroSPW55150.2022.00062
Type of publication	Conference item
Link to publisher's version	https://ieeexplore.ieee.org/abstract/document/9799345 - 10.1109/EuroSPW55150.2022.00062
Rights	© 2022, Samuel N. Eshun. Under license to IEEE.
Download date	2024-04-19 09:05:35
Item downloaded from	https://hdl.handle.net/10468/13352

UCC
University College Cork, Ireland
Coláiste na hOllscoile Corcaigh

Two de-anonymization attacks on real-world location data based on a hidden Markov model

Samuel N. Eshun
School of Computer Science & IT
University College Cork
Cork, Ireland
s.eshun@cs.ucc.ie

Paolo Palmieri
School of Computer Science & IT
University College Cork
Cork, Ireland
p.palmieri@cs.ucc.ie

***Abstract*—The increasing demand for smart context-aware services and the widespread use of location-based services (LBS) have resulted in the proliferation of mobile devices equipped with geolocation sensors (including GPS, geomagnetic field sensor, accelerometer, proximity sensor, et cetera). As a result, service providers and telecommunications companies can collect massive mobility datasets, often for millions of subscribers. To provide a degree of privacy, dataset owners normally replace personal identifiers such as name, address, and social security number (SSN) with pseudorandom identifiers prior to publication or sale. However, it has been repeatedly shown how sensitive information can be easily extracted or inferred from individuals' mobility data even when personal identifiers are removed. Knowledge of the extent to which location data can be de-anonymized is therefore crucial, in order to design appropriate privacy mechanisms that can prevent re-identification.**

In this paper, we propose and implement two novel and highly effective de-anonymization techniques: the Forward, and the KL algorithms. Our work utilizes a hidden Markov model (which incorporates spatio-temporal trajectories) in a novel way to generate user mobility profiles for target users. Using a real-world reference dataset containing mobility trajectories from the city of Shanghai (GeoLife, a reference dataset also used in previous studies), we evaluate the robustness of the proposed attack techniques. The results show that our attack techniques successfully re-identify up to 85% anonymized users. This significantly exceeds current comparable de-anonymization techniques, which have a success rate of 40% to 45%.

***Index Terms*—Location privacy, De-anonymization, DBSCAN Clustering, Hidden Markov model**

1. Introduction

The widespread adoption of smart devices equipped with geolocation sensors has led to the increased availability and use of location-based services (LBS). Such services enable service providers (SP) and telecommunication companies to collect large mobility datasets, often spanning millions of users who subscribe to these services. These mobility datasets are particularly valuable, as they can be used for urban planning, traffic forecasting, network optimization, targeted advertising, and a variety of other purposes, including contact tracing in the recent Covid-19 pandemic [1], [2]. Service providers have therefore an incentive in publishing or selling such datasets. A 2020 Grand View Research report on the market for personal location dataset estimates the market value at USD 10.6 billion in 2019, and predicts an annual growth rate of 15.2% over the forecast period (2020-27) [3]. The growth of open data policies across the world has also led to an increase in the freely available mobility datasets [4].

Due to privacy concerns regarding sensitive information that can be extracted from individuals' mobility data in published datasets, owners of mobility dataset are often required to replace personal identifiers such as name, address, and social security numbers (SSN) with pseudorandom identifiers to make the dataset (pseudo) anonymous. However, privacy risks still exists even if personal identifiers are removed before the mobility dataset is published or sold, and a large body of research has demonstrated that personal data of individuals is still at risk of exposure, even if their personal identifiers are removed [5]. For example, the analysis of places and geographical areas that are frequently visited by pseudonymous individuals in a dataset (areas of interest) can reveal their home, workplace, and place of worship [5]. Moreover, side information such as social networks geo-tagged posts (on platforms such as Twitter and Instagram), and publicly available mobility trajectories databases can easily be used to de-anonymize individuals whose data is included in the mobility dataset [6]. This is true even in the case of very large scale datasets: the work of Kondor et al. [7] presented a large-scale re-identification attack using mobile networks and transportation card usage, achieving a success rate of about 55% over four weeks of activities. Recent work by Farzanehfar et al. [8] confirms that the threat of re-identification persists even in a country-scale mobility dataset.

Even without secondary information, depending on how the personal identifiers are removed or how the dataset is sanitized, retrieving the auxiliary information from the same target dataset is sometimes feasible. This was illustrated in the work of De Montjoye et al. [9], who mined anonymized users' mobility records to uncover mobility patterns associated with the individuals in the target dataset. Another technique uses the home/work locations couple as a quasi-identifier to re-identify the users in the anonymized mobility dataset [5].

Given the various re-identification attacks proposed in the relevant literature, it is difficult to estimate the

actual level of privacy (if any) achieved by anonymizing a dataset. For this purpose, Shokri et al. [10] developed a framework for quantifying location privacy. Their works address the difficulties inherent in comparing various location privacy protection mechanisms (LPPMs) due to the lack of a systematic process.The work of Wang et al. [11] makes a case for a performance mismatch between algorithms and theoretical privacy bound. They ascribed the discrepancy to an underestimation of the impact of spatial-temporal data from a variety of sources. The performance of these algorithms was evaluated using two real-world datasets and then proposed algorithms to improve the performance.

Most de-anonymization attacks on geo-located databases, including those discussed previously, usually consider the user's temporal or spatial mobility trajectories (or both) in de-anonymizing or inferring sensitive information from the target database. These attacks are generally categorized as either linkage attacks or inference attacks. In the former, separate user accounts that belong to the same individual but are either anonymous or under different usernames are *re-linked* through data analysis. In the latter, an adversary tries to reveal personal identifiers or sensitive information of anonymized records in a target database using secondary information (auxiliary information) different from the target database or hidden values from the same target database.

In general, to carry out a de-anonymization attack, one must first decide on the most appropriate strategy for gathering background information on the target individuals (user fingerprinting) in the target database. The adversary then decides on the type of de-anonymization technique (for example, feature matching, graph-based, mobility models matching, statistical matching, etc.) that best suits the background knowledge modelled after the user fingerprinting. Finally, the de-anonymization attack is performed using the choosen de-anonymization technique. The success rates of these techniques are normally quantifiable and, as discussed in [12], the combination of the type of attack technique chosen and the background knowledge has an impact on the success of the de-anonymization attack.

Among the various de-anonymization techniques that have been presented in the research literature, a number are based on the hidden Markov model (HMM), a statistical approach that is frequently used in linear sequence 'labeling' problems [13]. In the context of location, the HMM approach has proved effective in extracting hidden information from complex trajectories [11], [14]. Generally, modelling HMM to build background knowledge on the target database mainly involves two phases. First, is the use of a clustering algorithm to identify the areas of interest to feed into the model. For instance, if the areas of interest represents the hidden states, identifying them from the raw mobility dataset usually requires a clustering algorithm (such as K-Means, DBSCAN, Hierarchical, etc.). In line with this, the Density-based spatial clustering of applications with noise (DBSCAN) algorithm is used in this work. The second phase is to compute the transition probabilities based upon the hidden states transitions, and the emission probabilities using the observation sequence.

A successful de-anonymization threat usually leads to identity disclosure, linking two or more de-anonymized accounts of different networks to the same user and disclosure of sensitive information such as an address, email, SSN etc. [15] identifies and classifies the above three primary threats emanating from a successful de-anonymization attack, and names them *content* (disclosure of sensitive data inferred from the location), *linkage* and *identity* (full re-identification of the user).

1.1. Contribution

In this work, we propose two novel Hidden Markov Model-based attacks that take into account both the temporal and spatial influences on user mobility trajectories. Our model, in particular, provides initial predictions about the user's regions of interest and, based on these predictions, constructs a hidden Markov model that output the probabilities of a user visiting any of these areas of interest for some observed given days of the week. Together, this process sums up the user's profile (which we describe as user fingerprinting) to re-identify the target individuals. Based on this user profiling, we propose and implement two de-anonymization attack mechanisms: the *Forward* algorithm and a divergence measure algorithm (based on the Kullback-Leibler divergence).

Finally, we evaluate the performance of the de-anonymization algorithms using real-world mobility traces from publicly available dataset (GeoLife [16]. Experimental results show that our proposed de-anonymization algorithms achieve a re-identification success rate (over 80%) that is significantly higher than those achieved by comparable studies, over the same dataset.

2. Preliminaries

This section presents preliminary information that is useful in the understanding of the remainder of the paper. In particular, we present the main techniques that are used as a basis for the construction of the novel de-anonymizers proposed in this work.

Two main techniques are used in this paper for building the mobility behaviour model of users: DBSCAN (presented in Section 2.1) and HMM (Section 2.2). In particular, the DBSCAN, a clustering algorithm, is used to cluster each individual's areas of interest. In other words, it is an unsupervised learning problem in which we seek to discover some structures (interesting locations) in an unlabeled dataset. The hidden Markov model (HMM) is a statistical model whose ability to modelled data without prior knowledge of the hidden states makes it useful in a variety of settings. It is used in this work to model the mobility behaviour of each user considering the days of the week and areas of interest predicted by the DBSCAN. Thus, these two techniques are combined to understand mobility behaviour for each user in the dataset.

The term de-anonymization is sometimes referred to as re-identification. Throughout this paper, both terms may be used interchangeably.

2.1. The DBSCAN

The Density-based spatial clustering of applications with noise (DBSCAN) is a well-known algorithm in data

TABLE 1: Notation: acronyms and symbols

Symbol	Meaning
N	the total number of states
n	the total number of users
π	initial probability distribution over all states
∇	area of Interest or states
δ	the days observation sequence of size T
$\beta_{i,j}$	the emission probabilities of moving to state ∇_i of type δ_j
γ	the transition probability matrix with members represented as γ_{ij}
T	maximum size of observation sequence

mining [17]. The main idea is to use a distance and points threshold input to group data points closer together as a cluster. It excels at detecting data patterns and can easily differentiate dense clusters from low clusters. Moreover, the DBSCAN can form clusters of varying shapes and is very robust to outliers and noise. Its robustness to outliers makes it applicable to mobility datasets, as they generally have many outliers and noise. The DBSCAN algorithm has two main parameters input:

- Epsilon(eps)- It defines the radius distance between two data points. Thus, the distance between two points less than or equal to *eps* belongs to the same cluster.
- minPoints- This input defines the minimum number of data points that can form a cluster.

After eps and minPoints are determined (supplied by the user), the algorithm randomly picks a starting point by creating a circle (using *eps*). Following that, a distance measurement is used to determine the distance between two data points (for example, the Euclidean distance). The starting points are then categorized as core points (greater than minPoints), Border points (contains at least one data point but less than minPoints) and finally noise (has zero data points within epsilon distance). The clusters are finally computed using the concept of reachability (within epsilon distance) and connectivity (defines if two pints belongs to the same cluster). For a full explanation of DBSCAN, we refer the reader to [17]

2.2. Hidden Markov Model (HMM)

A hidden Markov model is a statistical tool used in modelling sequential observations (visible) that probabilistically depend on a hidden sequence of events (hidden states). Our work considers both the spatial (areas of interest, or AoI) as well as temporal (days of the week) factors in the formation of these clusters. We propose a hidden Markov model [18], [19] to define a user's underlying AoI (e.g. home, place of work or worship etc.) by observing the sequences of days of the week. This allows us to build a mobility profile for each anonymous user to aid us in de-anonymizing them.

3. Hidden Markov Model (HMM) for location trajectories

In order to model trajectories, in this work we formally parameterize a hidden Markov model as $HMM = \{N, \Omega, \gamma, \beta, \pi\}$, where:

- $\boldsymbol{N}$, is the number of states (∇). The states represent the area of interest (AoI) visited by the user on any particular day as identified by the clustering algorithm (based on the assigned parameters) used. The area of interest (hidden states) is basically where a user frequently visits and spends considerable time. Such places, typically generated using a clustering algorithm, are usually sensitive, and the user may not wish to disclose them. These locations include but are not limited to one's home, work, place of worship, and shopping centre. The place of visits or area of interest together makes up the state set; $\nabla = \{\nabla_1, \nabla_2, \cdots, \nabla_N\}$.
- $\boldsymbol{\Omega}$, is the number of distinct symbols observed. For example, the observed symbols in this work are the set of distinct days (δ) of the week. Thus, the maximum number of Ω is 7. These observations are captured when the user visits any of these states. Hence, if a user visits more than one area of interest (state) on a particular day, say Tuesday, then this day would be recorded multiple times in line with the number of states visited. The sequence of observation symbols, $\delta_1 \cdots \delta_T$, has a maximum size of T.
- $\boldsymbol{\gamma}$, is an $N \times N$ transition probability matrix with members represented as γ_{ij}. It basically represents the probability distribution of transiting from one area of interest (state) to the other. Each member γ_{ij} is the probability of moving from state ∇_i to state ∇_j, i.e., the number of times a user moves from state ∇_i to state ∇_j over the number of times a user moves from state ∇_i. We formally define γ_{ij} as:

$$\gamma_{i,j} = \frac{\sigma(i,j)}{\sum_{j=0}^{N-1} \sigma(i,j)}$$

  where $\sigma(i,j)$ is the number of transitions from state (∇_i) to state ∇_j such that $\sum_j \gamma_{i,j} = 1 \; \forall \; i$.

- $\boldsymbol{\beta}$, is the emission probabilities with members represented as $\beta_{i,j}$, which is an $N \times \Omega$ matrix. It is the probability of a given state (∇) generating a particular observation symbol (δ). That is the probability that a user visited an area of interest (∇) on a particular day (δ).

$$\beta_{i,j} = \frac{\tau(i,j)}{\sum_{j=0}^{\Omega-1} \tau(i,j)}, \quad \text{where} \quad \begin{matrix} 1 \leq i \leq N \\ 1 \leq j \leq \Omega \end{matrix}$$

  $\tau(i,j)$ is the number of times a user visits an area of interest (∇_i) on a given day (δ_j), also $\sum_j \beta_{i,j} = 1 \; \forall \; i$

- $\boldsymbol{\pi}$, is the initial probability distribution over all states (∇), where

$$\pi_i = \frac{\sigma_i}{\sum_{j=0}^{N-1} \sigma_j}, \quad \text{and} \quad 1 \leq i \leq N$$

  where σ_i is the total number of times, a user first start from a state (∇_i) or a place of interest (AoI) for all given days.

4. Experimental Setup and Pre-processing

In preparing the target dataset for profiling of users, we sort the dataset by the number of days spent by each

user and divide it into two non-overlapping datasets: the training dataset, which accounts for approximately 80% of the total ground-truth dataset, and the auxiliary dataset (testing dataset), which accounts for the remaining 20%. This is in deviation from other setups in the literature [5], [20], where the testing dataset is not disjoint from the training dataset (and in fact is a subset of it). As argued in [21], this leads to a testing dataset that is heavily biased, and to artificially high de-anonymization results. Similarly to [21], our dataset is instead non-overlapping (trained and testing datasets are disjoint) to avoid bias. Unlike [21], however, where the dataset was split into two halves, one for training and one for testing, in this work the former makes up the bigger (80%) portion of the dataset. This, in a real-world scenario, would constitute anonymous users mobility datasets published by a service provider. The testing dataset is the adversary knowledge of a known user mobility traces, which should constitute a small fraction of the published mobility dataset and not necessarily be the same as the published datasets. The assumption is that, the adversary may only gain limited knowledge about the mobility behaviour of the target user.

Furthermore, for each user in the anonymous dataset (training datasets), a clustering algorithm is used to group each user's mobility traces (longitude and latitude) into clusters (Area of Interest, or AoI). The number of clusters (AoI) output depends on the clustering algorithm and the temporal and spatial resolution of the raw mobility traces. For instance, selecting smaller values for the selected clustering algorithm parameters may return more areas of interest (states) per user, which, in effect, may impact the success of de-anonymization.

These clusters (AoI), generated by the clustering algorithm, acts as the hidden states in the HMM as defined in Section 3. Generally, these clusters or areas of interest contain sensitive information (like home, place of work, place of worship, the type of health facility the user usually visits, etc.) that the adversary is likely to discover for each anonymous user in the target database (training dataset). Figure 1 depicts an example of a user whose mobility trajectories are classified into five clusters (areas) by the clustering algorithm, with the black points representing noise (outliers). The user in the picture is part the GeoLife dataset, described in Section 6.1. The specific days of the week during which users entered these clusters (areas of interest) represent the observational sequence used to learn about these areas of interest (clusters).

Then, using the HMM algorithm, we create a profile of each anonymous user's mobility traces (Markov model) from the training data. These profiles act as the fingerprint or unique identifier (trajectory model) for each anonymous user in the target database. An adversary can then easily associate one of these profiles with a target user whose identity and a limited set of spatiotemporal trajectories (testing dataset) are known to the adversary.

To speed up the re-identification process, we created a database (Λ) to hold the model (λ) of each anonymous user such that: $\Lambda = \{\lambda_j\}_{j=1}^{n}$, where $\lambda_j = \{N, \Omega, \gamma, \beta, \pi\}$ is the model parameters for each anonymous user, and n is the total number of anonymous users. The database (Λ) is updated each time the temporal and spatial resolutions are modified. Depending on the selected resolution parameters, the clustering algorithm may identify more or fewer areas of interest and accordingly affect the overall computation of the model parameters $(N, \Omega, \gamma, \beta, \pi)$ and consequently affect the re-identification process. The effect of this modification is illustrated in the analysis in Section 6.2

The testing data, which is just a fraction of the training dataset, is assumed to be the auxiliary knowledge (Spatio-temporal points) the adversary has about the target user, and wants to link to one of the anonymous users in the target database (training dataset). These spatiotemporal datasets (testing dataset) in the sense of auxiliary information is similar in granularity to that of the target database (training dataset) but not a subset of the target database. Thus, the adversary knows the target user but owns very limited Spatio-temporal trajectories (testing dataset) about this user and wants to link this auxiliary information to one of the anonymous users in the database. Though these users in the target database are unknown, their Markov mobility patterns containing sensitive information are known to the adversary.

For this analysis, and in order to be able to quantify the success rate, we assume the number of users in both the training and testing datasets is equal. Although this may not be the case in a real-world scenario, where the adversary is likely to have just a fraction of the number of people published in the target dataset (training dataset), this is a common approach in the literature. For this reason, let $\lambda^{\star}$ be the model (built from the auxiliary knowledge) for the target user the adversary knows about. This would be needed in performing the re-identification attack. A preprocessing such as encoding the user's observational sequences $(\delta_1 \cdots \delta_T)$ is performed before feeding the testing dataset into the de-anonymizing algorithm. This helps the model to understand better to assign a weight to the dataset correctly. Depending on the selected de-anonymizing algorithm and the dataset in question, the testing data is tuned by modifying the Spatio-temporal resolutions (in some cases, a new model is built on the testing data) before performing the re-identification attack.

5. De-anonymizers

In this section, two de-anonymization attack techniques (de-anonymizers) are presented: the *Forward* algorithm and a *Kullback-Leibler divergence*-based algorithm. The Forward algorithm is presented in Section 5.2. As it relies on the likelihood estimator by generating sequence of conditional probabilities in computing the likelihood, we also present the likelihood estimator in Section 5.1. The Kullback-Leibler divergence-based algorithm (Section 5.3), on the other hand, relies on the divergence between the two probabilistic distributions by quantifying their distance. Both techniques allow to re-identify users in the target dataset: experimental results of their application to two real-world datasets are discussed in Section 6.

5.1. Likelihood

One of the problems a hidden Markov model can solve in a real-world scenario is the probability for an observation sequence of size T (assumed to be generated by N-states) to belong to the model, or for the model results in such an observation sequence, i.e., $P(\delta|\lambda)$. The

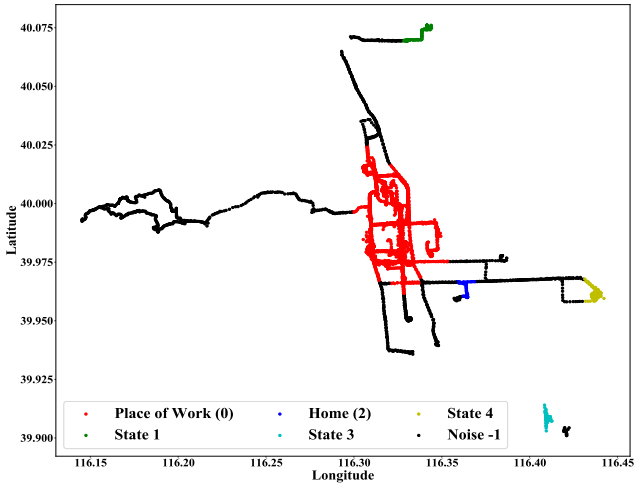

Figure 1: The spatial representation of the movements of a user in the GeoLife dataset: 5 areas of interest (clusters) are recognized, with black points being noise. Where an area is not connected to the rest of the graph, this could indicate that the location is reached via means that prevent GPS signal from being received (for instance, underground metro, or a tunnel).

intuition behind the likelihood score is that if a model (λ) is built based on observation sequence; $\delta = \delta_1 \cdots \delta_T$ (from the target database) similar to the one created by an individual (target user), this should result in a higher likelihood score, $P(\delta|\lambda)$. Ideally, computing the probability of such an observation sequence $\delta = \delta_1 \cdots \delta_T$ can occur when we know the states sequence $\nabla = \nabla_1 \cdots \nabla_T$. Thus computing the probability of the observational sequence considering all possible states sequence is given by:

$$\begin{aligned} P(\delta|\nabla,\lambda) &= \prod_{t=1}^{T} P(\delta_t|\nabla_t,\lambda) \\ &= \beta_1(\delta_1)\cdot\beta_2(\delta_2)\cdots\beta_T(\delta_T) \end{aligned} \tag{1}$$

Moreover, the probability of states sequence given the model is:

$$P(\nabla|\lambda) = \pi_1 \cdot \gamma_{12} \cdot \gamma_{21} \cdots \gamma_{T-1T} \tag{2}$$

Hence the joint probability of being in a state (∇) of a specific type of observation is given by:

$$P(\delta,\nabla) = P(\delta|\nabla)P(\nabla) \tag{3}$$

Therefore, the probability of the observation sequence assumed to be generated by the T states sequence is by summing the joint probability over all possible hidden states:

$$P(\delta) = \sum_{\nabla_T} P(\delta,\nabla) = \sum_{\nabla_T} P(\delta|\nabla)P(\nabla) \tag{4}$$

With the given model λ, the observational sequence probability becomes:

$$\begin{aligned} P(\delta|\lambda) &= \sum_{\nabla_T} P(\delta|\nabla,\lambda)P(\nabla|\lambda) \\ &= \sum_{\nabla_T} \pi_1\beta_1(\delta_1)\cdot\gamma_{1,2}\beta_2(\delta_2)\cdot\cdots\gamma_{T-1,T}\beta_T(\delta_T) \end{aligned} \tag{5}$$

A direct computation of equation (5) for N states and T observations is computationally infeasible due to N^T possible states sequence, especially when both N and T are large numbers. Thus having a computational complexity of approx $2T \cdot N^T$ for such computations is not ideal in calculating the likelihood. A more reliable and efficient algorithm to avoid such an exponential algorithm is the *Forward* algorithm [19], [22]. The derivation is similar to the one presented in [22].

5.2. The Forward de-anonymizer

The brute-force summation procedure for every possible state is not feasible in practice. Thus the use of the *Forward* algorithm [19], [22] to make such computations. Like dynamic programming, the *Forward* algorithm reduces the number of calculations when computing the observational probability. It computes the observational probability by summing the probabilities for all paths (hidden states), resulting in such an observational sequence.

To compute the *Forward* algorithm, let's define a forward probability (*forward parameter*); $\alpha_t(i) = P(\delta_1 \cdots \delta_t, \nabla_t = i|\lambda)$, which is the probability of the observational sequence, $\delta_1 \cdots \delta_t$ after being in state ∇_i at time t for the given model (λ).

In computing the helper vector (forward parameter) inductively:
First step:

$$\alpha_1(i) = \pi_i \cdot \beta_i(\delta_1) \quad \text{where} \quad 1 \le i \le N \ , \tag{6}$$

Inducting step gives:

$$\alpha_{t+1}(j) = \left[\sum_{i=1}^{N} \alpha_t(i)\cdot\gamma_{ij}\right]\cdot\beta_j(\delta_{t+1}) \ , \tag{7}$$

where

$$\begin{aligned} &1 \le j \le N \\ &1 \le t \le T-1 \end{aligned} \ ,$$

and where:

- $\alpha_t(i)$ is the probability of observing $\delta_1 \cdots \delta_t$ up to being in state ∇_i at time t ,
- γ_{ij} is the transition probability of moving from state ∇_i to state ∇_j ,
- $\beta_j(\delta_{t+1})$ is the probability of observing δ_{t+1} at state ∇_j.

Finally, summing all $\alpha_T(i)'s$ at final time step $t = T$ for all possible previous states gives:

$$P(\delta|\lambda) = \sum_{i=1}^{n} \alpha_T(i) \tag{8}$$

Comparing this computation to the previous brute-force summation of the likelihood for all possible states

(5) reduces the computation to order $O(N^2T)$, where N^2 is as a result of computing N previous states (N elements of α_t) to all N states for each T observations. Thus, with this forward computation, we can generate a score for each model (anonymous users whose mobility traces are known) to aid us in the re-identification (de-anonymization) process.

We explore this solution by building a scoring algorithm (maximum likelihood score) to de-anonymize (re-identify) the anonymous users (whose mobility traces are known) from the training dataset with a certain threshold. Computing such a score, all the anonymous users' models in the target database (Λ) are each called for each target user's observational sequence and the results (the likelihood score) stored in the form of a matrix for the re-identification analysis. Thus, let $\{\tilde{\Omega}_i\}_{i=1}^{n}$ be the total target users with a known observational sequence in the testing dataset and n be the total number of users where $\tilde{\Omega}_i = \delta_1^i \cdots \delta_T^i$ is the observational sequence for target user i. Hence, from equation (8), we compute $P(\tilde{\Omega}_i|\lambda_j)$, the likelihood scores for user i from the testing dataset given models λ_j from the training dataset (anonymous users). Thus, from the likelihood scores, the model λ_j (an anonymous user) that generates the maximum likelihood score is likely to be the corresponding user from the testing dataset (known users).

5.3. Kullback-Leibler divergence de-anonymizer

The next de-anonymization algorithm relies on a divergence measure or relative entropy based on the Kullback-Leibler divergence [23], which in itself is a generalization of Shannon's entropy. This divergence measure was first proposed by [24] as a distance measure for any two Markov models. It measures the difference between the two models' log probabilities based on the observations generated by one of the models. Technically, it's not a distance measure as it does not satisfy the symmetric or triangular inequality property. The divergence measure $\Gamma(\cdot,\cdot)$, between two models λ_1 and λ_2 is defined as:

$$\Gamma(\lambda_1, \lambda_2) = \frac{1}{T}[log\ P(\delta^{(2)}|\lambda_1) - log\ P(\delta^{(2)}|\lambda_2)] \quad (9)$$

where $\delta^{(2)} = \delta_1 \cdots \delta_T$, is the observation sequence of size T generated by the model λ_2. Looking at equation 9, one can see that it is just the difference in the likelihood computed in equation 8 between the models $\lambda_1\ and\ \lambda_2$. The interpretation of equation 9 is how well is model λ_1 different from model λ_2 or how divergent is model λ_1 from λ_2 based on the observation sequence generated by one of the models.

As this score measure is not symmetric, we employ the Jensen-Shannon divergence (JS) to make it symmetrized and smoothened. The Jensen-Shannon divergence (JS) between models λ_1 and λ_2 is given by:

$$JS(\lambda_1||\lambda_2) = \frac{1}{2}\left[\Gamma\left(\lambda_1, \frac{(\lambda_1+\lambda_2)}{2}\right) + \Gamma\left(\lambda_2, \frac{(\lambda_2+\lambda_1)}{2}\right)\right] \quad (10)$$

Based on this divergent measure, we are able to build an evaluation score for all the trained models to make predictions. A divergent score between λ_1 and λ_2 closer to 0 simply means the two models, λ_1 and λ_2, are likely to be identical.

The adversary explores this algorithm by constructing a probability model of the target user's Spatio-temporal trajectories, similar to the one built for the users in the target database. This model comprises transition probabilities (the likelihood of moving from one area of interest to the other) and emission probabilities representing the days the users were seen in these locations (areas of interest). Consequently, for each target user $(\lambda^\star)$, the adversary computes the divergence measure between $(\lambda^\star)$ and each anonymous users in the target database, $\Lambda = \{\lambda_j\}_{j=1}^{n}$. From the divergent scores between the target user and the anonymous users, the pair models (target user and any of the anonymous users) that result in the minimum score; $\min(\lambda^\star, \lambda)$ for a given threshold indicates a corresponding match in the target database (Λ), i.e., the two models are likely to be identical.

6. Experimental Evaluation

We evaluate the performance of our de-anonymization attacks based on individual-level ground truth mobility trajectories from two different groups of users. The two proposed de-anonymization attacks are both based on the target user's blueprint (unique Spatio-temporal profile) built using the HMM. Parameters such as spatial and temporal values were varied to test the robustness of the proposed de-anonymizers. In addition, the success of the de-anonymizers was compared with the state of the art de-anonymizers in the literature. Specifically, we compare our work to other attacks that used a similar dataset to compute the accuracy of their algorithms.

6.1. Dataset

In evaluating the efficiency of the model and the de-anonymization attacks, we exploit a publicly available real-life mobility traces of users in the city of Shanghai (GeoLife) [16].

The GeoLife [16] dataset, consisting of 182 users' GPS trajectories, was collected by Microsoft Research Asia spanning from April 2007 to August 2012 in Shanghai. The dataset collected at a high rate of 1 to 5 s covers users' daily routines such as shopping, sporting activities, workplace, going home, etc. Users whose mobility traces covers less than 8 days of GPS trajectories were filtered out for this evaluation.

6.2. Experimental results

Our experiment considers how the de-anonymizers perform separately on the real-life mobility dataset (GeoLife [16]) and how the performance rates are affected by varying spatial resolution. As presented in figure 2, the success rate of the *Forward* deanonymizer (D_{FW}) on the GeoLife dataset ranges from 72% to 65%, depending on the chosen radius distance. This performance is a significant improvement when compared to the performance of other de-anonymizers in the literature (as discussed in Section 6.3). From the figure, the radius distance is inversely proportional to the success rate of the *Forward* de-anonymizer.

Successful re-identification of an anonymous user in the target database implies that there is a similarity in

observational mobility trajectories used to build the model and that of the known user, resulting in the algorithm returning the highest score.

The D_{KL} de-anonymizer was also tested on GeoLife datasets, with even stronger results. Figure 2 depicts the success rates of the D_{KL} de-anonymizer on the GeoLife; from the figure, the prediction rate increases from 74% to 85% as the radius distance decreases.

When it comes to the performance of both de-anonymizers, the D_{KL} de-anonymizer has a higher prediction rate than the *Forward* algorithm. Figure 2 depicts the superiority of the D_{KL} algorithm over that of the *Forward* algorithm (D_{FW}) with varying spatial resolution. However, the rates of increase within both de-anonymizers were almost the same as the radius distance decreases. This presupposes that the characteristics of the dataset in question have an impact on the performances of the de-anonymizer.

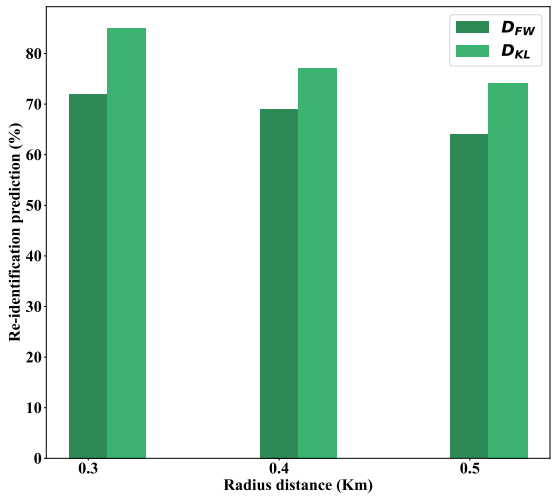

Figure 2: Performance rate of the de-anonymizers on the GeoLife dataset [16].

A property that makes the D_{KL} de-anonymizer more effective is the observational property matrix (emission probabilities). Deviations in $\boldsymbol{\beta}$ return a higher dissimilarity score D_{KL} than in $\boldsymbol{\gamma}$ (transition between areas of interest). Consequently, users are more likely to be re-identified if they normally visit areas of interest on particular days. Naturally users gravitate to places of interest on specific days, like going to the theatre on Wednesdays or to a stadium on Saturdays. For example, consider two users who usually visit the same location, say a shopping mall; the algorithm easily re-identifies them in a target database as they normally make this visit on different days. Thus even if the trajectories are similar, they are still likely to be re-identified because of the weight placed on the identity matrix.

From the analysis, nearly 66% of the user profiles (Spatio-temporal user patterns) in the GeoLife dataset built by the HMM were so distinct that both the *Forward* and D_{KL} de-anonymizers could re-identify these same individuals. As a result of these users' unique mobility patterns based on the HMM model, the attacker could re-identify them regardless of the de-anonymizer used in this work. Furthermore, one intriguing finding from the study is that when both de-anonymizers predict the same user, as in 66% of the cases in the GeoLife dataset, then it is almost a certainty. At no point did the two algorithms (*Forward* & D_{KL}) predicted the same user that turned out to be false. This gives the assurance that combining these two algorithms in the de-anonymization attacks will yield a result with a high degree of accuracy.

6.3. Comparison to previous attacks

As the GeoLife dataset [16] is frequently used in the literature, it is commonly regarded as a benchmark. For this reason, we compare our results to previous results on the basis of this dataset.

To make a fair comparisons between our work and other known de-anonymization attacks, we only consider previous works that used a non-overlapping training and the testing dataset. As discussed in Section 4, and to avoid ambiguity, the training and the testing dataset should not be overlapping, but rather disjoint (non-overlapping) to avoid bias. As evidenced in [21], such bias results in artificially high success rates for the de-anonymization.

The works of MMC [21] and UHMM [14] best met the criteria outlined above. Both schemes re briefly outlined in Section 7.

Figure 3 shows how our algorithm compares to MMC [21] and UHMM [14] in terms of success rate. As illustrated in Figure 3, our de-anonymization attack outperforms other algorithms when using the same parameters; our attacks D_{FW} and D_{KL} succeed at a rate of 72% and 88%, respectively, compared to 40 to 45 per cent for UHMM [14] and MMC [21], demonstrating the effectiveness of our attack tools.

Further proof of the effectiveness of the proposed Forward and KL algorithms is that they surpass even attacks evaluated over biased testing sets. Speifically, UHMM's achieved attack success rate is still lower than that of our algorithms even when the training and testing datasets overlapped (were not disjoint), at approximately 70%. We however discarded this result from the comparison for the reasons outlined above. Similarly, [20], which achieves a performance rate of up to around 90% based on the cabs dataset in San Francisco, comes closest to our algorithm's performance, but was again heavily biased because the testing dataset was directly extracted from the training dataset.

7. Related Work

De-anonymization attacks in the context of a geo-located dataset have been in the study for a while. However, prior research has demonstrated that deleting personal identifiers like names, and email addresses etc., is insufficient to anonymize individuals in the target mobility dataset. According to the work of Mulder et al. [25], individual behaviour in a Global System for Mobile communication (GSM) network could be used to re-identify users. They employed two different methods of identification: sequence of cells ID and Markovian model. The latter used static cells (GSM cells) as the states of the

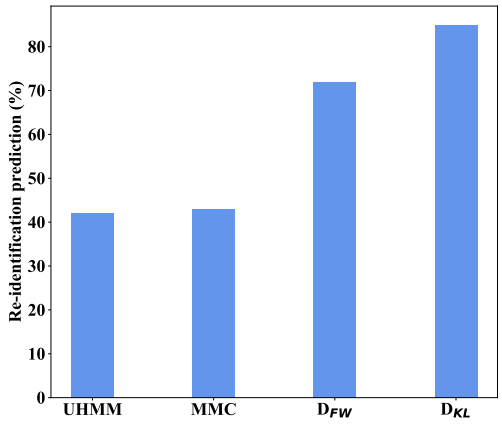

Figure 3: Comparing the performance of our method and that of MMC [21] and UHMM [14] on the GeoLife [16] dataset

model, which made it almost unlikely to form an ergodic (every state could be connected in a single step) model, i.e., the transition between cells is unlikely to happen without neighbouring cells. Gambs et al. [21] also build a Mobility Markov Chain (MMC) to de-anonymize users in a mobility dataset. In their work, a state represents each point of interest (frequently visited locations). As with our work, each MMC is associated with a user, which represents mobility behaviour. In addition, they proposed distance metrics to measure the similarity between two MMC where the minimum distance is the re-identified user. One of the datasets used to evaluate the performance of their algorithm is the GeoLife dataset. Freudiger et al. [5] worked on re-identification attacks in the context of geolocation. Their work considers both spatial and temporal influences for de-anonymizing the target users. In particular, they create a pair, home/work as an area of interest, where a home is a location where the user typically stays between 9 p.m. and 9 a.m., and work is a location where the user typically stays between 9 a.m. and 5 p.m. The home/work pair is then used as a quasi-identifier to perform the de-anonymization attacks. Wang et al. [14] proposed a user hidden Markov model that took spatial and temporal influences into account when de-anonymizing targeted users in a geo-located dataset. Each HMM constitutes the mobility behaviour of the user in the target dataset. Their technique is similar to our hidden Markov model but with a number of differences. First, they divided a day into a 24-time span, which corresponds to the set of states while a place of visit is regarded as the observational sequence. This is in sharp contrast to our approach, which uses frequently visited locations (AoI) as the set of states and the days of the week as the observational sequence. In addition, their attack was evaluated based on the GeoLife dataset using the ranking and voting step as the output metric.

8. Conclusion

In this work, we proposed and implemented a de-anonymization model for geolocation data. To de-anonymize the target users, our attack employs the hidden Markov model (by taking into account Spatio-temporal trajectories) to create user mobility profiles (user fingerprinting).

We build the HMM by first predicting the users' initial areas of interest or frequently visited locations (known as hidden states) and the days (observations) on which users visited these locations before feeding into the model. We assess the robustness of attack techniques based on ground truth mobility trajectories from two different cities (Shanghai and Rome). Based on the evaluation score, our attack techniques successfully re-identify up to 85% anonymized users in the GeoLife dataset. The proposed algorithms significantly outperform other comparable de-anonymization attack techniques in the literature.

Acknowledgment

The authors would like to express their gratitude to Ivan Kayongo at IBM Kenya for his assistance in optimizing the code.

References

[1] L. Calderoni, D. Maio, and P. Palmieri, "Location-aware mobile services for a smart city: Design, implementation and deployment," *J. Theor. Appl. Electron. Commer. Res.*, vol. 7, no. 3, pp. 74–87, 2012. [Online]. Available: http://www.jtaer.com/dec2012/Calderoni_p7.pdf

[2] P. Lohar, G. Xie, M. Bendechache, R. Brennan, E. Celeste, R. Trestian, and I. Tal, "Irish attitudes toward COVID tracker app & privacy: Sentiment analysis on twitter and survey data," in *ARES 2021: The 16th International Conference on Availability, Reliability and Security, Vienna, Austria, August 17-20, 2021*, D. Reinhardt and T. Müller, Eds. ACM, 2021, pp. 37:1–37:8. [Online]. Available: https://doi.org/10.1145/3465481.3469193

[3] "Location intelligence market size, share & trends analysis report," Grand View Research, Tech. Rep. GVR-2-68038-401-7, Feb 2020. [Online]. Available: https://www.grandviewresearch.com/industry-analysis/location-intelligence-market

[4] M. Fiore, P. Katsikouli, E. Zavou, M. Cunche, F. Fessant, D. L. Hello, U. M. Aïvodji, B. Olivier, T. Quertier, and R. Stanica, "Privacy in trajectory micro-data publishing: a survey," *Trans. Data Priv.*, vol. 13, no. 2, pp. 91–149, 2020. [Online]. Available: http://www.tdp.cat/issues16/tdp.a363a19.pdf

[5] J. Freudiger, R. Shokri, and J. Hubaux, "Evaluating the privacy risk of location-based services," in *Financial Cryptography and Data Security - 15th International Conference, FC 2011, Gros Islet, St. Lucia, February 28 - March 4, 2011, Revised Selected Papers*, ser. Lecture Notes in Computer Science, G. Danezis, Ed., vol. 7035. Springer, 2011, pp. 31–46. [Online]. Available: https://doi.org/10.1007/978-3-642-27576-0_3

[6] M. Srivatsa and M. Hicks, "Deanonymizing mobility traces: using social network as a side-channel," in *the ACM Conference on Computer and Communications Security, CCS'12, Raleigh, NC, USA, October 16-18, 2012*, T. Yu, G. Danezis, and V. D. Gligor, Eds. ACM, 2012, pp. 628–637. [Online]. Available: https://doi.org/10.1145/2382196.2382262

[7] D. Kondor, B. Hashemian, Y. de Montjoye, and C. Ratti, "Towards matching user mobility traces in large-scale datasets," *IEEE Trans. Big Data*, vol. 6, no. 4, pp. 714–726, 2020. [Online]. Available: https://doi.org/10.1109/TBDATA.2018.2871693

[8] A. Farzanehfar, F. Houssiau, and Y. de Montjoye, "The risk of re-identification remains high even in country-scale location datasets," *Patterns*, vol. 2, no. 3, p. 100204, 2021. [Online]. Available: https://doi.org/10.1016/j.patter.2021.100204

[9] Y.-A. De Montjoye, C. A. Hidalgo, M. Verleysen, and V. D. Blondel, "Unique in the crowd: The privacy bounds of human mobility," *Scientific reports*, vol. 3, no. 1, pp. 1–5, 2013.

[10] R. Shokri, G. Theodorakopoulos, J. L. Boudec, and J. Hubaux, "Quantifying location privacy," in *32nd IEEE Symposium on Security and Privacy, S&P 2011, 22-25 May 2011, Berkeley, California, USA*. IEEE Computer Society, 2011, pp. 247–262. [Online]. Available: https://doi.org/10.1109/SP.2011.18

[11] H. Wang, C. Gao, Y. Li, G. Wang, D. Jin, and J. Sun, "De-anonymization of mobility trajectories: Dissecting the gaps between theory and practice," in *25th Annual Network and Distributed System Security Symposium, NDSS 2018, San Diego, California, USA, February 18-21, 2018*. The Internet Society, 2018. [Online]. Available: http://wp.internetsociety.org/ndss/wp-content/uploads/sites/25/2018/02/ndss2018_06B-3_Wang_paper.pdf

[12] D. Al-Azizy, D. E. Millard, I. Symeonidis, K. O'Hara, and N. Shadbolt, "A literature survey and classifications on data deanonymisation," in *Risks and Security of Internet and Systems - 10th International Conference, CRiSIS 2015, Mytilene, Lesbos Island, Greece, July 20-22, 2015, Revised Selected Papers*, ser. Lecture Notes in Computer Science, C. Lambrinoudakis and A. Gabillon, Eds., vol. 9572. Springer, 2015, pp. 36–51. [Online]. Available: https://doi.org/10.1007/978-3-319-31811-0_3

[13] R. Das, C. W. Cairo, and D. Coombs, "A hidden markov model for single particle tracks quantifies dynamic interactions between LFA-1 and the actin cytoskeleton," *PLoS Comput. Biol.*, vol. 5, no. 11, 2009. [Online]. Available: https://doi.org/10.1371/journal.pcbi.1000556

[14] R. Wang, M. Zhang, D. Feng, Y. Fu, and Z. Chen, "A de-anonymization attack on geo-located data considering spatio-temporal influences," in *Information and Communications Security - 17th International Conference, ICICS 2015, Beijing, China, December 9-11, 2015, Revised Selected Papers*, ser. Lecture Notes in Computer Science, S. Qing, E. Okamoto, K. Kim, and D. Liu, Eds., vol. 9543. Springer, 2015, pp. 478–484. [Online]. Available: https://doi.org/10.1007/978-3-319-29814-6_41

[15] S. Sharma, P. Gupta, and V. Bhatnagar, "Anonymisation in social network: a literature survey and classification," *Int. J. Soc. Netw. Min.*, vol. 1, no. 1, pp. 51–66, 2012. [Online]. Available: https://doi.org/10.1504/IJSNM.2012.045105

[16] Y. Zheng, L. Zhang, X. Xie, and W. Ma, "Mining interesting locations and travel sequences from GPS trajectories," in *Proceedings of the 18th International Conference on World Wide Web, WWW 2009, Madrid, Spain, April 20-24, 2009*, J. Quemada, G. León, Y. S. Maarek, and W. Nejdl, Eds. ACM, 2009, pp. 791–800. [Online]. Available: https://doi.org/10.1145/1526709.1526816

[17] M. Ester, H. Kriegel, J. Sander, and X. Xu, "A density-based algorithm for discovering clusters in large spatial databases with noise," in *Proceedings of the Second International Conference on Knowledge Discovery and Data Mining (KDD-96), Portland, Oregon, USA*, E. Simoudis, J. Han, and U. M. Fayyad, Eds. AAAI Press, 1996, pp. 226–231. [Online]. Available: http://www.aaai.org/Library/KDD/1996/kdd96-037.php

[18] L. E. Baum and T. Petrie, "Statistical inference for probabilistic functions of finite state markov chains," *The annals of mathematical statistics*, vol. 37, no. 6, pp. 1554–1563, 1966.

[19] L. E. Baum and G. Sell, "Growth transformations for functions on manifolds," *Pacific Journal of Mathematics*, vol. 27, no. 2, pp. 211–227, 1968.

[20] C. Y. T. Ma, D. K. Y. Yau, N. K. Yip, and N. S. V. Rao, "Privacy vulnerability of published anonymous mobility traces," in *Proceedings of the 16th Annual International Conference on Mobile Computing and Networking, MOBICOM 2010, Chicago, Illinois, USA, September 20-24, 2010*, N. H. Vaidya, S. Banerjee, and D. Katabi, Eds. ACM, 2010, pp. 185–196. [Online]. Available: https://doi.org/10.1145/1859995.1860017

[21] S. Gambs, M. Killijian, and M. N. del Prado Cortez, "De-anonymization attack on geolocated data," *J. Comput. Syst. Sci.*, vol. 80, no. 8, pp. 1597–1614, 2014. [Online]. Available: https://doi.org/10.1016/j.jcss.2014.04.024

[22] L. Rabiner, "A tutorial on hidden markov models and selected applications in speech recognition," *Proceedings of the IEEE*, vol. 77, no. 2, pp. 257–286, 1989.

[23] S. Kullback and R. A. Leibler, "On information and sufficiency," *The annals of mathematical statistics*, vol. 22, no. 1, pp. 79–86, 1951.

[24] B. Juang and L. R. Rabiner, "A probabilistic distance measure for hidden markov models," *AT&T Tech. J.*, vol. 64, no. 2, pp. 391–408, 1985. [Online]. Available: https://doi.org/10.1002/j.1538-7305.1985.tb00439.x

[25] Y. D. Mulder, G. Danezis, L. Batina, and B. Preneel, "Identification via location-profiling in GSM networks," in *Proceedings of the 2008 ACM Workshop on Privacy in the Electronic Society, WPES 2008, Alexandria, VA, USA, October 27, 2008*, V. Atluri and M. Winslett, Eds. ACM, 2008, pp. 23–32. [Online]. Available: https://doi.org/10.1145/1456403.1456409