Dionysis Manousakas*, Cecilia Mascolo, Alastair R. Beresford, Dennis Chan, and Nikhil Sharma

Quantifying Privacy Loss of Human Mobility Graph Topology

Abstract: Human mobility is often represented as a mobility network, or graph, with nodes representing places of significance which an individual visits, such as their home, work, places of social amenity, etc., and edge weights corresponding to probability estimates of movements between these places. Previous research has shown that individuals can be identified by a small number of geolocated nodes in their mobility network, rendering mobility trace anonymization a hard task. In this paper we build on prior work and demonstrate that even when all location and timestamp information is removed from nodes, the graph topology of an individual mobility network itself is often uniquely identifying. Further, we observe that a mobility network is often unique, even when only a small number of the most popular nodes and edges are considered. We evaluate our approach using a large dataset of cell-tower location traces from 1 500 smartphone handsets with a mean duration of 430 days. We process the data to derive the top$-N$ places visited by the device in the trace, and find that 93% of traces have a unique top-10 mobility network, and all traces are unique when considering top-15 mobility networks. Since mobility patterns, and therefore mobility networks for an individual, vary over time, we use graph kernel distance functions, to determine whether two mobility networks, taken at different points in time, represent the same individual. We then show that our distance metrics, while imperfect predictors, perform significantly better than a random strategy and therefore our approach represents a significant loss in privacy.

Keywords: Mobility privacy; De-anonymization; $k-$anonymity; Graph kernels.

*Corresponding Author: Dionysis Manousakas:
University of Cambridge, dm754@cam.ac.uk.
Cecilia Mascolo: University of Cambridge and
the Alan Turing Institute, cm542@cam.ac.uk.
Alastair R. Beresford: University of Cambridge,
arb33@cam.ac.uk.
Dennis Chan: University of Cambridge, dc598@cam.ac.uk.
Nikhil Sharma: University College London,
nikhil.sharma@ucl.ac.uk.

1 Introduction

Our mobile devices collect a significant amount of data about us and location data of individuals are particularly privacy sensitive. Furthermore, previous work has shown that removing direct identifiers from mobility traces does not provide anonymity: users can easily be reidentified by a small number of unique locations that they visit frequently [6, 43].

Consequently, some approaches have been proposed that protect location privacy by replacing location coordinates with encrypted identifiers, using different encryption keys for each location trace in the population. This preprocessing results in locations that are strictly user-specific and cannot be cross-referenced between users. Examples include the research track of the Nokia Mobile Data Challenge,[1] where visited places were represented by random integers [14]; and identifiable location information collected by the Device Analyzer dataset,[2] including WiFi access point MAC addresses and cell tower identifiers, are mapped to a set of pseudonyms defined separately for each handset [35]. Moreover, temporal resolution may also be deliberately decreased to improve anonymization [11] since previous work has demonstrated that sparsity in the temporal evolution of mobility can cause privacy breaches [6].

In this paper, *we examine the degree to which mobility traces without either semantically-meaningful location labels, or fine-grained temporal information, are identifying.* To do so, we represent location data for an individual as a mobility network, where nodes correspond to abstract locations and edges to their connectivity, i.e. the respective transitions made by an individual between locations. We then examine whether or not these graphs reflect user-specific behavioural attributes that could act as a fingerprint, perhaps allowing the reidentification of the individual they represent. In particular, we show how graph kernel distance functions [34] can be used to assist reidentification of anonymous mobility networks. This opens up new opportunities for both attack and defense. For example, patterns found

---

1 http://www.idiap.ch/project/mdc
2 https://deviceanalyzer.cl.cam.ac.uk

in mobility networks could be used to support automated user verification where the mobility network acts as a behavioural signature of the legitimate user of the device. However the technique could also be used to link together different user profiles which represent the same individual.

Our approach differs from previous studies in location data deanonymization [7, 9, 10, 21], in that *we aim to quantify the breach risk in preprocessed location data that do not disclose explicit geographic information*, and where instead locations are replaced with a set of user-specific pseudonyms. Moreover, we also do not assume specific timing information for the visits to abstract locations, *merely ordering.*

We evaluate the power of our approach over a large dataset of traces from 1 500 smartphones, where cell tower identifiers (*cid*s) are used for localization. Our results show that the data contains structural information which may uniquely identify users. This fact then supports the development of techniques to efficiently re-identify individual mobility profiles. Conversely, our analysis may also support the development of techniques to cluster into larger groups with similar mobility; such an approach may then be able to offer better anonymity guarantees.

A summary of our contributions is as follows:

– We show that network representations of individual longitudinal mobility display distinct topology, even for a small number of nodes corresponding to the most frequently visited locations.
– We evaluate the sizes of identifiability sets formed in a large population of mobile users for increasing network size. Our empirical results demonstrate that all networks become quickly uniquely identifiable with less than 20 locations.
– We propose kernel-based distance metrics to quantify mobility network similarity in the absence of semantically meaningful spatial labels or fine-grained temporal information.
– Based on these distance metrics, we devise a probabilistic retrieval mechanism to reidentify pseudonymized mobility traces.
– We evaluate our methods over a large dataset of smartphone mobility traces. We consider an attack scenario where an adversary has access to historical mobility networks of the population she tries to deanonymize. We show that by informing her retrieval mechanism with structural similarity information computed via a deep shortest-path graph kernel, the adversary can achieve a median deanonymization probability 3.52 times higher than

a randomised mechanism using no structural information contained in the mobility networks.

2 Related Work

2.1 Mobility Deanonymization

Protecting the anonymity of personal mobility is notoriously difficult due to sparsity [1] and hence mobility data are often vulnerable to deanonymization attacks [22]. Numerous studies into location privacy have shown that even when an individual's data are anonymized, they continue to possess unique patterns that can be exploited by a malicious adversary with access to auxiliary information. Zang et al. analysed nationwide call-data records (*CDR*s) and showed that the N most frequently visited places, so called $top-N$ data, correlated with publicly released side information and resulted in privacy risks, even for small values of Ns [43]. This finding underlines the need for reductions in spatial or temporal data fidelity before publication. De Montjoye et al. quantified the unicity of human mobility on a mobile phone dataset of approximately $1.5M$ users with intrinsic temporal resolution of one hour and a 15-month measurement period [6]. They found that four random spatio-temporal points suffice to uniquely identify 95% of the traces. They also observe that the uniqueness of traces decreases as a power law of spatio-temporal granularity, stressing the hardness of achieving privacy via obfuscation of time and space information.

Several inference attacks on longitudinal mobility are based on probabilistic models trained on individual traces and rely on the regularity of human mobility. Mulder et al. developed a re-identification technique by building a Markov model for each individual in the training set, and then using this to re-identify individuals in the test set by likelihood maximisation [7]. Similarly, Gambs et al. used Markov chains to model mobility traces in support of re-identification [9].

Naini et al. explored the privacy impact of releasing statistics of individuals mobility traces in the form of histograms, instead of their actual location information [21]. They demonstrated that even this statistical information suffices to successfully recover the identity of individuals in datasets of few hundred people, via jointly matching labeled and unlabeled histograms of a population. Other researchers have investigated the privacy threats from information sharing on location-based social networks, including the impact of location

semantics on the difficulty of re-identification [27] and location inference [2].

All the above previous work assumes locations are expressed using a universal symbol or global identifier, either corresponding to (potentially obfuscated) geographic coordinates, or pseudonymous stay points. Hence, cross-referencing between individuals in the population is possible. This is inapplicable when location information is anonymised separately for each individual. Lin et al. presented a user verification method in this setting [16]. It is based on statistical profiles of individual indoor and outdoor mobility, including cell tower ID and WiFi access point information. In contrast, we employ network representations based solely on cell tower ID sequences without explicit time information.

Often, studies in human mobility aim to model properties of a population, thus location data are published as aggregate statistics computed over the locations of individuals. This has traditionally been considered a secure way to obfuscate the sensitive information contained in individual location data, especially when released aggregates conform to $k-$anonymity [32] principles. However, recent results have questioned this assumption. Xu et al. recovered movement trajectories of individuals with accuracy levels of between 73% and 91% from aggregate location information computed from cellular location information involving 100 000 users [38]. Similarly, Pyrgelis et al. performed a set of inference attacks on aggregate location time-series data and detected serious privacy loss, even when individual data are perturbed by differential privacy mechanisms before aggregation [26].

2.2 Anonymity of Graph Data

Most of the aforementioned data can be represented as *microdata* with rows of fixed dimensionality in a table. Microdata can thus be embedded into a vector space. In other applications, datapoints are *relational* and can be naturally represented as *graphs*. Measuring the similarity of such data is significantly more challenging, since there is no definitive method. Deanonymization attacks on graphs have mostly been studied in the context of social networks and aimed to either align nodes between an auxiliary and an unknown targeted graph [23, 29], or quantify the leakage of private information of a graph node via its neighbors [44].

In the problem studied here, *each individual's information is an entire graph*, rather than a node in a graph or a node attribute, and thus deanonymization is

reduced to a graph matching or classification problem. To the best of our knowledge, this is the first attempt to deanonymize an individual's structured data by applying graph similarity metrics. Since we are looking at relational data, not microdata, standard theoretical results on microdata anonymization, such as differential privacy [8], are not directly applicable. However, metrics related to structural similiarity, including $k-$anonymity, can be generalized in this framework.

3 Proposed Methodology

In this section, we first adapt the privacy framework of $k-$anonymity to the case of graph data (Section 3.1). Next we introduce our methodology: We assume that all mobility data are initially represented as a sequence of pseudonymous locations. We also assume the pseudonymisation process is distinct per user, and therefore locations cannot be compared between individuals. In other words, it is not possible to determine whether pseudonymous location l_u for user u is the same as (or different from) location l_v for user v. We convert a location sequence for each user into a mobility network (Section 3.2). We then extract feature representations of these networks and embed them into a vector space. Finally, in the vector space, we can define pairwise distances between the network embeddings (Section 3.3) and use them in a deanonymization scenario (Section 3.4).

Our methodology is, in principle, applicable to many other categories of recurrent behavioural trajectories that can be abstracted as graphs, such web browsing sessions [24, 42] or smartphone application usage sequences [37]. We leave such analysis as future work.

3.1 $k-$anonymity on Graphs

Anonymity among networks refers to topological (or structural) equivalence. In our analysis we adopt the privacy framework of $k-$anonymity [32] which we summarize as follows:

Definition 3.1. ($k-$**anonymity**) A microdata release of statistics containing separate entries for a number of individuals in the population satisfies the $k-$anonymity property if the information for each individual contained in the release is indistinguishable from at least $k - 1$

other individuals whose information also appears in the release.

Therefore we interpret $k-$anonymity in this paper to mean that the mobility network of an individual in a population should be identical to the mobility network of at least $k-1$ other individuals. Recent work casts doubt on the protection guarantees offered by $k-$anonymity in location privacy [31], motivating the definition of $l-$diversity [17] and $t-$closeness [15]. Although $k-$anonymity may be insufficient to ensure privacy in the presence of adversarial knowledge, $k-$anonymity is a good metric to use to measure the uniqueness of an individual in the data. Moreover, this framework is straightforwardly generalizable to the case of graph data.

Structural equivalence in the space of graphs corresponds to isomorphism and, based on this, we can define $k-anonymity\ on\ unweighted\ graphs$ as follows:

Definition 3.2. (**Graph Isomorphism**) Two graphs $G = (V, E)$ and $G' = (V', E')$ are *isomorphic* (or *belong to the same isomorphism class*) if there exists a bijective mapping $g : V \to V'$ such that $(v_i, v_j) \in E$ iff $\big(g(v_i), g(v_j)\big) \in E'$.

Definition 3.3. (**Graph** $k-$**anonymity**) *Graph* $k-anonymity$ is the minimum cardinality of isomorphism classes within a population of graphs.

After clustering our population of graphs into isomorphism classes, we can also define the *identifiability set* and *anonymity size* [25] as follows:

Definition 3.4. (**Identifiability Set**) *Identifiability set* is the percentage of the population which is uniquely identified given their top$-N$ network.

Definition 3.5. (**Anonymity Size**) The *anonymity size* of a network within a population is the cardinality of the isomorphism class to which the network belongs.

3.2 Mobility Information Networks

To study the topological patterns of mobility, we represent user movements by a mobility network. A preliminary step is to check whether a first-order network is a reasonable representation of movement data, or whether a higher-order network is required.

First-order network representations of mobility traces are built on the assumption of a *first-order temporal correlation* among their states. In the case of mobility data, this means that the transition by an individual to the next location in the mobility network can be accurately modelled by considering only their current location. For example, the probability that an individual visits the shops or work next depends only on where they are located now, and a more detailed past history of places recently visited does not offer significant improvements to the model. The alternative is that a sequences of the states are better modelled by higher-order Markov chains, namely that transitions depend on the current state and one or more previously visited states. For example, the probability that an individual visits the shops or work next depends not only on where they are now, but where they were earlier in the day or week. If higher-order Markov chains are required, we should assume a larger state-space and use these states as the nodes of our individual mobility networks. Recently proposed methods on optimal order selection of sequential data [28, 39] can be directly applied at this step.

Let us assume a mobility dataset from a population of users $u \in U$. We introduce two network representations of user's mobility.

Definition 3.6. (**State Connectivity Network**): A **state connectivity network** for u is an unweighted directed graph $C^u = \big(V^u, E^u\big)$. Nodes $v_i \in V^u$ correspond to states visited by the user throughout the observation period. An edge $e_{ij} = \big(v_i^u, v_j^u\big) \in E^u$ represents the information that u had at least one recorded transition from v_i^u to v_j^u.

Definition 3.7. (**Mobility Network**): A **mobility network** for u is a weighted and directed graph $G^u = \big(V^u, E^u, W^u\big) \in \mathcal{G}$, with the same topology as the state connectivity network and additionally an edge weight function $W^u : E^u \to \mathbb{R}^+$. The weight function assigns a frequency w_{ij}^u to each edge e_{ij}^u, which corresponds to the number of transitions from v_i^u to v_j^u recorded throughout the observation period.

To facilitate comparisons of frequencies across networks of different sizes in our experiments, we normalize edge weights on each mobility network to sum to 1.

In first-order networks, nodes correspond to distinct places the user visits. Given a high-frequency, timestamped, sequence of location events for a user, distinct places can be extracted as small geographic regions where a user stays longer than a defined time interval, using existing clustering algorithms [13]. Nodes in the

mobility network have no geographic or timing information associated with them. Nodes may have *attributes* attached to them reflecting additional side information. For example, in this paper we consider whether attaching the frequency of visits a user makes to a specific node aids an attacker attempting to deanonymize the user.

In some of our experiments, we prune the mobility networks of users by reducing the size of the mobility network to the N most frequent places and rearranging the edges in the network accordingly. We refer to these networks as **top$-N$ mobility networks**.

3.3 Graph Similarity Metrics

It is not possible to apply a graph isomorphism test to two mobility networks to determine if they represent the same underlying user because a user's mobility network is likely to vary over time. Therefore we need distance functions that can measure the degree of similarity between two graphs. Distance functions decompose the graph into feature vectors (smaller substructures and pattern counts), or histograms of graph statistics, and express similarity as the distance between those feature representations. In the following, we introduce the notion of graph kernels and describe the graph similarity metrics used later in our experiments.

We wish to compute the similarity between two graphs $G, G' \in \mathcal{G}$. Kernel functions [34], or *kernels*, are symmetric positive semidefinite functions, where $K(G, G') : \mathcal{G} \times \mathcal{G} \rightarrow \mathcal{R}^+$, meaning that for all $n > 1$, $G_1, ..., G_n \in \mathcal{G}$, and $c_1, ..., c_n \in \mathcal{R}$, we have $\sum_{i,j=1}^{n} c_i c_j K(G_i, G_j) \geq 0$. Each kernel function corresponds to some feature map $\phi(G)$, where the kernel function can be expressed as the inner product between feature maps, i.e., $K(G, G') = \langle \phi(G), \phi(G') \rangle$.

In order to ensure the result from the kernel lies in the range from -1 to 1 inclusive, we apply *cosine normalization* as follows:

$$K(G, G') = \left\langle \frac{\phi(G)}{||\phi(G)||}, \frac{\phi(G')}{||\phi(G')||} \right\rangle. \qquad (1)$$

One interpretation of this function is as the *cosine similarity of the graphs in the feature space* defined by the map of the kernel.

In our experiments we apply a number of scalable kernel functions on our mobility datasets and assess their suitability for deanonymization applications on mobility networks. We note in advance that as the degree distribution and all substructure counts of a graph

remain unchanged under structure-preserving bijection of the vertex set, all examined graph kernels are invariant under isomorphism. We briefly introduce these kernels in the remainder of the section.

3.3.1 Kernels on degree distribution

The degree distribution of nodes in the graph can be used to quantify the similarity between two graphs. For example, we can use a histogram of weighted or unweighted node degree as a feature vector. We can then compute the pairwise distance of two graphs by taking either the inner product of the feature vectors or passing them through a Gaussian radial basis function (RBF) kernel:

$$K(G, G') = \exp\left(- \frac{||\phi(G) - \phi(G')||^2}{2\sigma^2} \right).$$

Here, the parameters of the kernel are the variance σ (in case RBF is used) and the number of bins in the histogram.

3.3.2 Kernel on graph atomic substructures

Kernels can use counts on substructures, such as subtree patterns, shortest paths, walks or limited-size subgraphs. This family of kernels are called $R-convolution$ *graph kernels* [12]. In this way, graphs are represented as vectors with elements corresponding to the frequency of each such substructure over the graph. Hence, if $s_1, s_2, ... \in \mathcal{S}$ are the substructures of interest and $\#(s_i \in G)$ the counts of s_i in graph G, we get as feature map vectors

$$\phi(G) = [\#(s_1 \in G), \#(s_2 \in G), ...]^T \qquad (2)$$

with dimension $|\mathcal{S}|$ and kernel

$$K(G, G') = \sum_{s \in \mathcal{S}} \#(s \in G) \#(s \in G'). \qquad (3)$$

In the following, we briefly present some kernels in this category and explain how they are adapted in our experiments.

Shortest-Path Kernel

The Shortest-Path (SP) graph kernel [5] expresses the similarity between two graphs by counting the co-occuring shortest paths in the graphs. It can be written in the form of equation (3) where each element $s_i \in \mathcal{S}$ is a triplet $(a_{\text{start}}^i, a_{\text{end}}^i, n)$, where n is the length of the
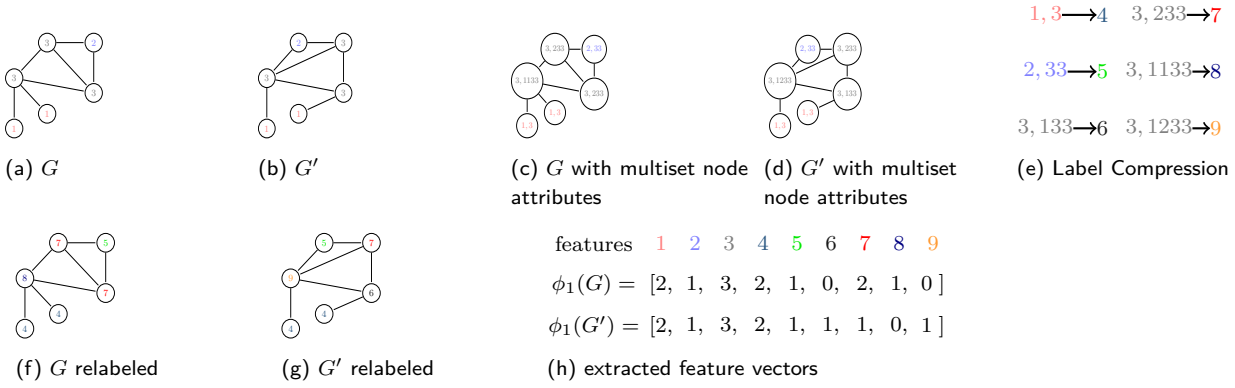
(a) G (b) G' (c) G with multiset node attributes (d) G' with multiset node attributes (e) Label Compression

(f) G relabeled (g) G' relabeled (h) extracted feature vectors

$$
\begin{array}{cccccccccc}
\text{features} & 1 & 2 & 3 & 4 & 5 & 6 & 7 & 8 & 9 \\
\phi_1(G) = & [\,2, & 1, & 3, & 2, & 1, & 0, & 2, & 1, & 0\,] \\
\phi_1(G') = & [\,2, & 1, & 3, & 2, & 1, & 1, & 1, & 0, & 1\,]
\end{array}
$$

Fig. 1. Computation of the Weisfeiler-Lehman subtree kernel of height $h = 1$ for two attributed graphs.

path and $a_{\text{start}}^i, a_{\text{end}}^i$ the attributes of the starting and ending nodes. The shortest path set is computable in polynomial time using, for example, the Floyd-Warshall algorithm, with complexity $O(|V|^4)$, where $|V|$ is number of nodes in the network.

Weisfeiler-Lehman Subtree Kernel

Shervashidze et al. proposed an efficient method [30] to construct a graph kernel utilizing the Weisfeiler-Lehman (WL) test of isomorphism [36]. The idea of the WL kernel is to measure co-occurences of subtree patterns across node attributed graphs.

Computation progresses over iterations as follows:
1. each node attribute is augmented with a multiset of attributes from adjacent nodes;
2. each node attribute is then compressed into a single attribute label for the next iteration; and
3. the above steps are repeated until a specified threshold h is reached.

An example is shown in Figure 1.

If G and G' are the two graphs, the WL subtree kernel is defined as follows:

$$
K_{WL}^h(G, G') = \langle \phi_h(G), \phi_h(G') \rangle,
$$

where $\phi_h(G)$ and $\phi_h(G')$ are the vectors of labels extracted after running h steps of the computation (Figure 1h). They consist of h blocks, where the i−th component of the j−th block corresponds to the frequency of label i at the j−th iteration of the computation. The computational complexity of the kernel scales *linearly* with the number of edges $|E|$ and the length h of the WL graph sequence.

Deep Graph Kernels

Deep graph kernels (DKs) are a unified framework that takes into account similarity relations at the level of atomic substructures in the kernel computation [41]. Hence, these kernels can quantify *similar substructure* co-occurence, offering more robust feature representations. DKs are based on computing the following inner product:

$$
K(G, G') = \phi(G)^T M \phi(G'),
$$

where ϕ is the feature mapping of a classical R-convolution graph kernel.

In the above, $M : |\mathcal{V}| \times |\mathcal{V}|$ is a positive-definitive matrix encoding the relationships between the atomic substructures and $\mathcal{V}$ is the vocabulary of the observed substructures in the dataset. Here, M can be defined using the edit distance of the substructures, i.e. the number of elementary operations to transform one substructure to another; or M can be learnt from the data, applying relevant neural language modeling methods [19].

3.4 Deanonymization of User Mobility Networks and Privacy Leakage Evaluation

3.4.1 Hypothesis

The basic premise of our deanonymization approach can be postulated as follows:

The mobility of a person across different time periods is stochastic, but largely recurrent and stationary, and its expression at the level of the individual mobility network is discriminative enough to reduce a person's privacy within a population.

For example, the daily commute to work corresponds to a relatively stable sequence of cell towers. This can be expressed in the mobility network of the user as a persistent subgraph and forms a characteristic behavioural pattern that can be exploited
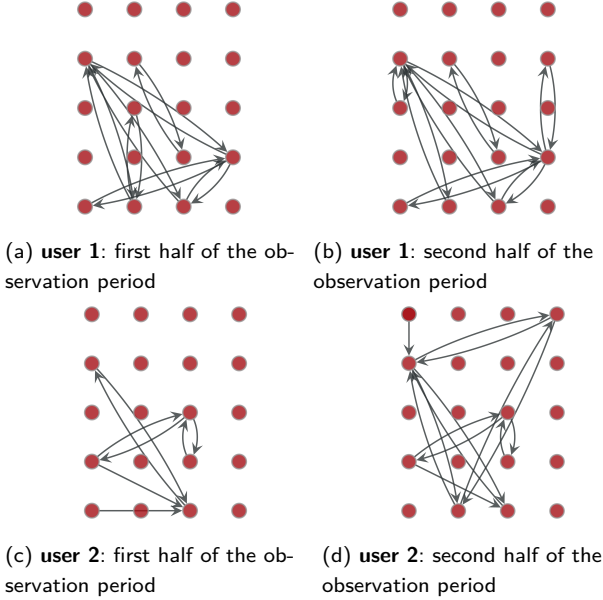
(a) **user 1**: first half of the observation period

(b) **user 1**: second half of the observation period

(c) **user 2**: first half of the observation period

(d) **user 2**: second half of the observation period

Fig. 2. Top−20 networks for two random users from the Device Analyzer dataset. Depicted edges correspond to the 10% most frequent transitions in the respective observation window. The networks show a high degree of similarity between the mobility profiles of the same user over the two observation periods. Moreover, the presence of single directed edges in the profile of **user 2** forms a discriminative pattern that allows us to distinguish **user 2** from **user 1**.

for deanonymization of mobility traces. Empirical evidence for our hypothesis is shown in Figure 2. For ease of presentation, in the figure, nodes between the disparate observation periods of the users can be cross-referenced. We assume that cross-referencing is not possible in our attack scenario, as locations are independently pseudonymized.

3.4.2 Threat Model

We assume an adversary has access to a *set of mobility networks* $G \in \mathcal{G}_{\text{training}}$ *with disclosed identities (or labels)* $l_G \in \mathcal{L}$ and a *set of mobility networks* $G' \in \mathcal{G}_{\text{test}}$ *with undisclosed identities* $l_{G'} \in \mathcal{L}$.[3]

We define a normalised similarity metric among the networks $K : \mathcal{G}_{\text{training}} \times \mathcal{G}_{\text{test}} \to \mathcal{R}^+$. We hypothesize

---

3 Generally we can think of $l_{G'} \in \mathcal{J} \supset \mathcal{L}$ and assign some fixed probability mass to the labels $l_{G'} \in \mathcal{J} \setminus \mathcal{L}$. However, here we make the *closed world assumption* that the training and test networks come from the same population. We make this assumption for two reasons: first, it is a common assumption in works on deanonymization and, second, we cannot directly update our beliefs on $l_{G'} \in \mathcal{J} \setminus \mathcal{L}$ by observing samples from $\mathcal{L}$.

that a training and test mobility network belonging to the same person have common or similar connectivity patterns, thus a high degree of similarity.

The intention of an adversary is to deanonymize a given test network $G' \in \mathcal{G}_{\text{test}}$, by appropriately defining a vector of probabilities over the possible identities in $\mathcal{L}$.

An **uninformed adversary** has *no information* about the networks of the population and in the absence of any other side knowledge, the prior belief of the adversary about the identity of G' is a uniform distribution over all possible identities:

$$P\big(l_{G'} = l_{G_i}\big) = 1/|\mathcal{L}|, \text{ for every } G_i \in \mathcal{G}_{\text{training}}. \quad (4)$$

An **informed adversary** has *access to the population of training networks* and can compute the pairwise similarities of G' with each $G_i \in \mathcal{G}_{\text{training}}$ using a kernel function K. Hence the adversary can update her belief for the possible identities in $\mathcal{L}$ according to the values of K. Therefore, when the adversary attempts to deanonymize identities in the data, she assigns probabilities that follow a *non-decreasing function* of the computed pairwise similarity of each label. Denoting this function by f, we can write the updated adversarial probability estimate for each identity as follows:

$$P\big(l_{G'} = l_{G_i} | \mathcal{G}_{\text{training}}, K\big) = \frac{f\big(K(G_i, G')\big)}{\sum\limits_{j \in \mathcal{L}} f\big(K(G_j, G')\big)}, \quad (5)$$

for every $G_i \in \mathcal{G}_{\text{training}}$.

3.4.3 Privacy Loss

In the case of the uninformed adversary, the true label for any user is expected to have rank $|\mathcal{L}|/2$. Under this policy, the amount of privacy for each user is proportional to the size of the population.

In the case of the informed adversary, knowledge of $\mathcal{G}_{\text{training}}$ and the use of K will induce some positive *privacy loss* which will result in the expected rank of user to be smaller than $|\mathcal{L}|/2$. The privacy loss can be quantified as follows:

$$PL\big(G'; \mathcal{G}_{\text{training}}, K\big) = \frac{P\big(l_{G'} = l_{G'_{\text{true}}} | \mathcal{G}_{\text{training}}, K\big)}{P\big(l_{G'} = l_{G'_{\text{true}}}\big)} - 1 \quad (6)$$

A privacy loss equal to zero reflects no information gain compared to an uninformed adversary with no access to graphs with disclosed identities.

Let us assume that the users of our population generate distinct mobility networks. As will be supported with empirical evidence in the next section, this is often the case in real-world *cid* datasets of few thousand users even for small networks sizes (e.g. for top-20 networks in our dataset). Under the above premise, the maximal privacy loss occurs when the presented test network is an identical copy of a training network of the same user which exists in the data of the adversary, i.e. $G' \in \mathcal{G}_{\text{training}}$. This corresponds to a user deterministically repeating her mobility patterns over the observation period recorded in the test network. In such a scenario, we could think that isomorphism tests are the most natural way to compute similarity; however, isomorphism tests will be useless in real-world scenarios, since the stochastic nature and noise inherent in the mobility networks of a user would make them non-isomorphic. Maximal privacy loss reflects the discriminative ability of the kernel and cannot be exceeded in real-world datasets, where the test networks are expected to be noisy copies of the training networks existing in our system. The step of comparing with the set of training networks adds computational complexity of $\mathcal{O}(|\mathcal{G}_{\text{training}}|)$ to the similarity metric cost.

Moreover, our framework can naturally facilitate incorporating new data to our beliefs when multiple examples per individual exist in the training dataset. For example, when multiple instances of mobility networks per user are available, we can use $k-$nearest neighbors techniques in the comparison of distances with the test graph.

4 Data for Analysis

In this section we present an exploratory analysis of the dataset used in our experiments, highlighting statistical properties of the data and empirical results regarding the structural anonymity of the generated state connectivity networks.

4.1 Data Description

We evaluate our methodology on the Device Analyzer dataset [35]. Device Analyzer contains records of smartphone usage collected from over $30\,000$ study participants around the globe. Collected data include information about system status and parameters, running background processes, cellular connectivity and wireless connectivity. For privacy purposes, released *cid* information is given a unqiue pseudonym separately for each user and contains no geographic, or semantic, information concerning the location of users. Thus we cannot determine geographic proximity between the nodes and the location data of two users cannot be directly aligned.

For our experiments, we analysed *cid* information collected from $1\,500$ handsets with the largest recorded location datapoints in the dataset. Figure 4a shows the observation period for these handsets; note that the mean is greater than one year but there is lot of variance across the population. We selected these $1\,500$ handsets in order to examine the re-identifiability of devices with rich longitudinal mobility profiles. This allowed us to study the various attributes of individual mobility affecting privacy in detail. As mentioned in the previous section, the cost of computing the adversarial posterior probability for the deanonymization of a given unlabeled network scales linearly with the population size.

4.2 Mobility Networks Construction

We began by selecting the optimal order of the network representations derived from the mobility trajectories of the $1\,500$ handsets selected from the Device Analyzer dataset. We first parsed the *cid* sequences from the mobility trajectories into mobility networks. In order to remove *cid*s associated with movement, we only defined nodes for *cid*s which were visited by the handset for at least 15 minutes. Movements from one *cid* to another were then recorded as edges in the mobility network.

As outlined in Section 3.1, we analysed the pathways of the Device Analyzer dataset during the entire observation period, applying the model selection method [28] of Scholtes.[4] This method tests graphical models of varying orders and selects the optimal order by balancing the model complexity and the explanatory power of observations.

We tested higher-order models up to order three. In the case of top-20 mobility networks, we found routine patterns in the mobility trajectories were best explained with models of order two for more than 20% of the users. However, when considering top-100, top-200, top-500 and full mobility networks, we found that the optimal model for our dataset has order one for more than 99% of the users; see Figure 3. In other words, when considering mobility trajectories which visit less frequent locations in the graph, the overall increase in likelihood of the
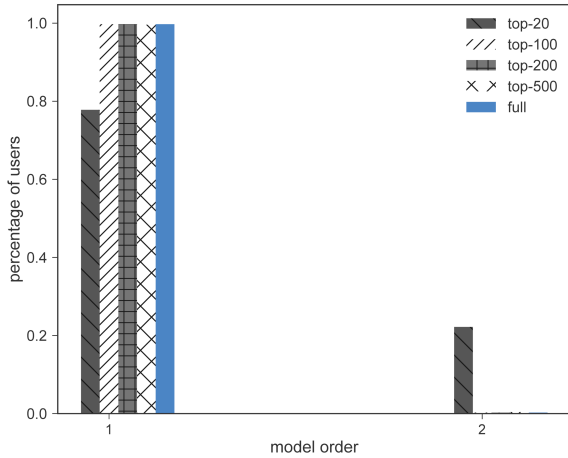
---

4 https://github.com/IngoScholtes/pathpy

Fig. 3. Optimal order for increasing number of locations.

data for higher-order models cannot compensate for the complexity penalty induced by the larger state space. So while there might still be regions in the graph which are best represented by a higher-order model, the optimal order describing the entire graph is one. Therefore we use a model of order one in the rest of this paper.

4.3 Data Properties and Statistics

In Table 1 we provide a statistical summary of the original and the pruned versions of the mobility networks. We observe that allowing more locations in the network implies an increase in the variance of their statistics and leads to smaller density, larger diameter and larger average shortest-path values.

A *recurrent edge traversal* in a mobility network occurs when a previously traversed edge is traversed for a second or subsequent time. We then define *recurrence rate* as the percentage of edge traversals which are recurrent. We find that mobility networks display a high recurrence rate, varying from 68.8% on average for full networks to 84.7% for the top-50 networks, indicating that the mobility of the users is mostly comprised of

repetitive transitions between a small set of nodes in a mobility network.

Figure 4b displays the normalized histogram and probability density estimate of network size for full mobility networks. We observe that sizes of few hundred nodes are most likely in our dataset, however mobility networks of more than 1 000 nodes also exist. Reducing the variance in network size will be proved helpful in cross-network similarity metrics, hence we also consider truncated versions of the networks.

As shown in Figure 4c, the parsed mobility network of a typical user is characterized by a *heavy-tailed degree distribution*. We observe that a small number of locations have high degree and correspond to dominant states for a person's mobility routine, while a large number of locations are only visited a few times throughout the entire observation period and have a small degree.

Figure 4d shows that the estimated probability distribution of average edge weight. This peaks in the range from two to four, indicating that many transitions captured in the full mobility network are rarely repeated. However, most of the total weight of the network is attributed to the tail of this distribution, which corresponds to the edges that the user frequently repeats.

4.4 Anonymity Clusters on Top$-N$ Networks

We examine to what extent the heterogeneity of users mobility behaviour can be expressed in the topology of the state connectivity networks. For this purpose, we generate the isomorphism classes of the top$-N$ networks of our dataset for increasing network size N. We then compute the graph $k-$anonymity of the population and the corresponding identifiability set. This analysis demonstrates empirically the privacy implications of releasing anonymized users pathway information at increasing levels of granularity.

Before presenting our findings on the Device Analyzer dataset, we will perform a theoretical upper bound analysis on the identifiability of a population, by finding the maximum number of people that can be distin-

| Networks	# of networks	Num. of nodes, avg.	Edges, avg.	Density, avg.	Avg. clust. coef.	Diameter, avg.	Avg. short. path	Recurrence rate (%)
top-50 locations	1500	49.92 ± 1.26	236.55 ± 78.14	0.19 ± 0.06	0.70 ± 0.07	3.42 ± 0.86	1.93 ± 0.20	84.7 ± 5.6
top-100 locations	1500	98.33 ± 7.93	387.05 ± 144.73	0.08 ± 0.03	0.60 ± 0.10	4.67 ± 1.48	2.33 ± 0.40	78.3 ± 7.8
top-200 locations	1500	179.23 ± 37.82	548.21 ± 246.11	0.04 ± 0.02	0.47 ± 0.12	7.52 ± 4.21	3.07 ± 1.18	73.0 ± 9.9
full	1500	334.60 ± 235.81	741.64 ± 527.28	0.02 ± 0.02	0.33 ± 0.09	15.98 ± 10.18	4.84 ± 2.93	68.8 ± 12.3

Table 1. Summary statistics of mobility networks in the Device Analyzer dataset.
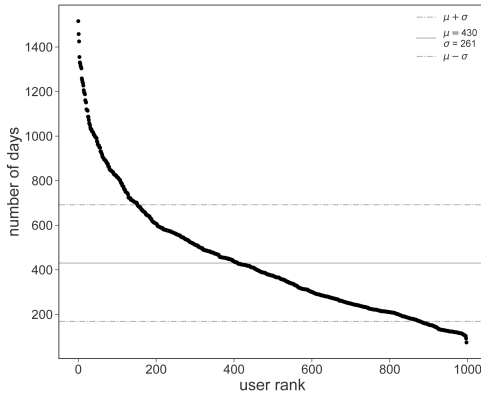
(a) Observation period duration distribution.
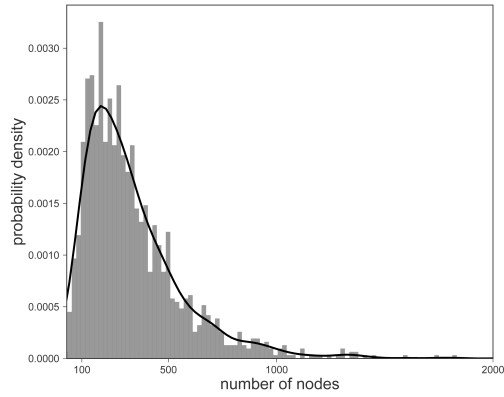
(b) Normalized histogram and probability density estimate of network size for the full mobility networks over the population.
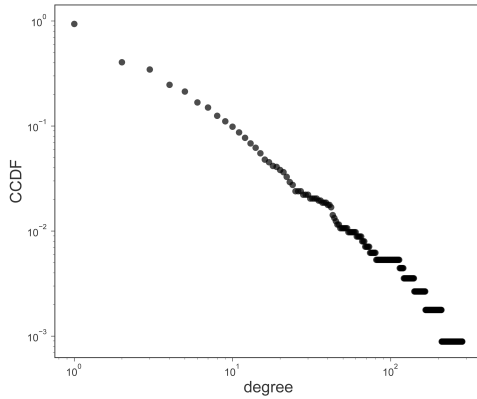
(c) Complementary cumulative distribution function ($CCDF$) for the node degree in the mobility network of a typical user from the population, displayed on log-log scale.
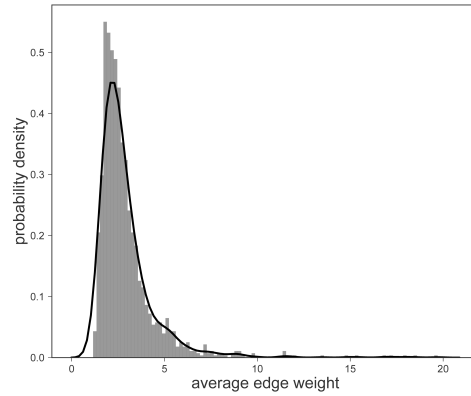
(d) Normalized histogram and probabilty density of average edge weight over the networks.

Fig. 4. Empirical statistical findings of the Device Analyzer dataset.

guished by networks of size N. This corresponds to the number of non-isomorphic graphs with N nodes.

Currently the most of efficient way of enumerating non-isomorphic graphs is by using McKay's algorithm [18], implemented in the package nauty.[5] Table

N	4	5	6	7	8	9
# undirected	11	34	156	1044	12346	274668
N	4	5		6		7
# directed	2128	9608		1540944		882033440

Table 2. Sequences of non-isomorphic graphs for undirected and directed graphs of increasing size.

5 http://pallini.di.uniroma1.it/

2 presents the enumeration for undirected and directed non-isomorphic graphs of increasing size. We observe that there exist 12 346 undirected graphs with 8 nodes and 9 608 directed graphs with 5 nodes. In other words, finding the top−8 places for each person is the smallest number which could produce unique graphs for each person in our sample of 1 500 individuals; this reduces to 5 when directionality is taken into account. Moreover, we find that top−12 undirected and top−8 directed networks are sufficient to enable each human on the planet to be represented by a different graph, assuming world population of 7.6B.

Next we present the results of our analysis on the Device Analyzer data. As observed in Figure 5, *sparsity arises in a mobility network even for very small N*. In particular, in the space of undirected *top*−4 location
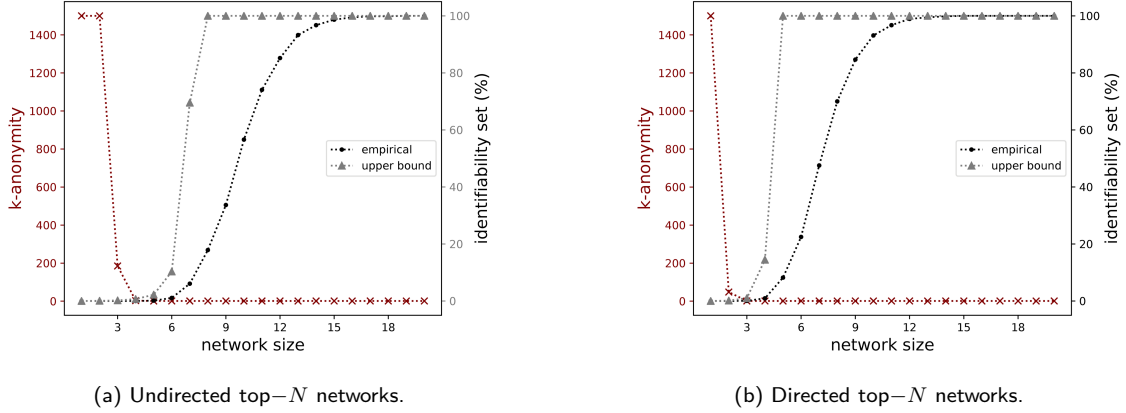
(a) Undirected top$-N$ networks.

(b) Directed top$-N$ networks.

Fig. 5. Identifiability set and $k-$anonymity for undirected and directed top$-N$ mobility networks for increasing number of nodes. Displayed is also the theoretical upper bound of identifiability for networks with N nodes.
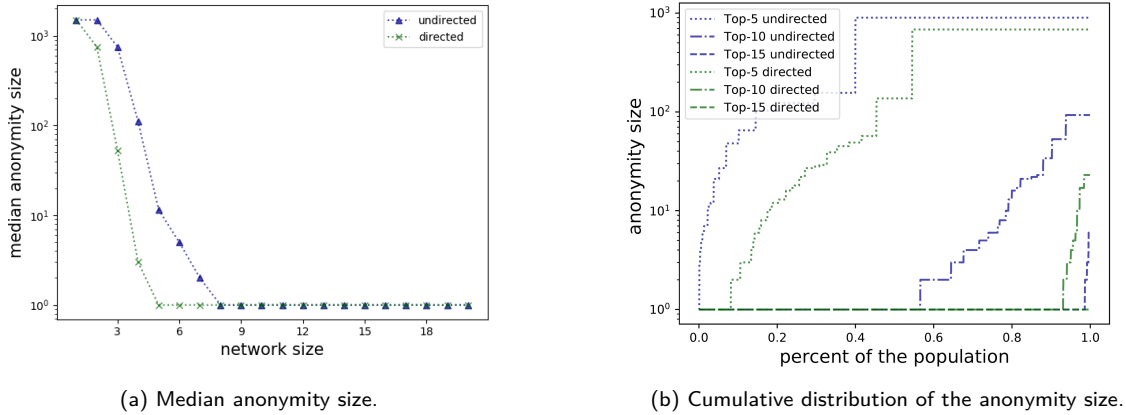

(a) Median anonymity size.

(b) Cumulative distribution of the anonymity size.

Fig. 6. Anonymity size statistics over the population of top$-N$ mobility networks for increasing network size.

networks, there is already a cluster with only 3 members, while for all $N > 4$ there always exist isolated isomorphic clusters. $k-anonymity$ decreases to 1 even for $N = 3$ when considering directionality. Moreover, the *identifiability set* dramatically increases with the size of network: approximately 60% of the users are uniquely identifiable from their top-10 location network. This percentage increases to 93% in directed networks. For the entire population of the 1 500 users, we find that 15 and 19 locations suffice to form uniquely identifiable directed and undirected networks respectively.

The difference between our empirical findings and our theoretical analysis suggests that large parts of the top$-N$ networks are common to many people. This can be attributed to patterns that are widely shared (e.g. the trip from work to home, and from home to work).

Figure 6 shows some additional statistics of the anonymous isomorphic clusters formed for varying net-work sizes. Median anonymity becomes one for network sizes of five and eight in directed and undirected networks respectively; see Figure 6a. In Figure 6b we observe that the population arranges into clusters with small anonymity even for very small network sizes: around 5% of the users have at most 10-anonymity when considering only five locations in their network, while this percentage increases to 80% and 100% for networks with 10 and 15 locations. This result confirms that anonymity is even harder when the directionality of edges are provided, since the space of directed networks is much larger than the space of the undirected networks with the same number of nodes.

The above empirical results indicate that the diversity of individuals mobility is reflected in the network representations we use, thus we can meaningfully proceed to discriminative tasks on the population of mobility networks.

5 Evaluation of Privacy Loss in Longitudinal Mobility Traces

In this section we empirically quantify the privacy leakage implied by the information of longitudinal mobility networks for the population of users in the Device Analyzer dataset. For this purpose we undertake experiments in graph matching using different kernel functions, and assume an adversary has access to a variety of mobility network information.

5.1 Experimental Setup

For our experiments we split the *cid* sequences of each user into two sets: the *training* sequences where user identities are disclosed to the adversary, and the *test* sequences where user identities are undisclosed to the adversary but are used to quantify the success of the adversarial attack. Therefore each user has two mobility networks: one derived from the training sequences, and one derived from the test sequences. The objective of the adversary is to successfully match every test mobility network with the training mobility network representing the same underlying user. To do so, the adversary computes the pairwise distances between training mobility networks and test mobility networks. We partitioned *cid* sequences of each user by time, placing all *cid*s before the partition point in the training set, and all *cid*s after into the test set. We choose the partition point separately for each user as a random number from the uniform distribution with range 0.3 to 0.7.

5.2 Mobility Networks & Kernels

We computed the pairwise distances between training and test mobility networks using kernels from the categories described in Section 3. Node attributes are supported in the computation of Weisfeiler-Lehman and Shortest-Path kernel. Thus we augmented the individual mobility networks with categorical features to add some information about the different roles of nodes in users mobility routine. Such attributes are computed independently for each user on the basis of the topological information of each network. After experimenting with several schemes, we obtained the best performance on the kernels when dividing locations into three categories with respect to the frequency in which each node is visited by the user. Concretely, we computed the distribu-

tion of users' visits to locations and added the following values to the nodes:

$$
a_{c=3}\left(v_i^u\right) = \begin{cases} 3, & \text{if } v_i^u \in \text{top}-20\% \text{ locations of } u \\ 2, & \text{if } v_i^u \notin \text{top}-20\% \text{ locations of } u \\ & \text{and } v_i^u \in \text{top}-80\% \text{ locations} \\ 1 & \text{otherwise.} \end{cases}
$$

This scheme allowed a coarse, yet informative, characterisation of locations in users networks, which was robust to the variance in the frequency of visits between the two observation periods. In addition, we removed 40% of edges with the smallest edge weights and retained only the largest connected component for each user.

Due to its linear complexity, computation of the Weisfeiler-Lehman kernel could scale over entire mobility networks. However, we had to reduce the network size in order to apply the Shortest-Path kernel. This was done using top$-N$ networks for varying size N.

5.3 Evaluation & Discussion

We evaluated graph kernels functions from the following categories:

– DSP_N: Deep Shortest-Path kernel on top$-N$ network
– DWL_N: Deep Weisfeiler-Lehman kernel on top$-N$ network
– DD: Degree Distribution kernel through Gaussian RBF
– WD: Weighted Degree distribution through Gaussian RBF

The Cumulative Density Functions (*CDF*s) of the true label rank for the best performing kernel of each category are presented in Figure 7.

If mobility networks are unique, an *ideal retrieval mechanism* would correspond to a curve that reaches 1 at rank one, indicating a system able to correctly deanonymize all traces by matching the closest training graph. This would be the case when users training and test networks are identical, thus the knowledge of the latter implies maximum privacy loss.

Our baseline, *random*, is a strategy which reflects the policy of an adversary with *zero knowledge* about the mobility networks of the users, who simply returns uniformly random orderings of the labels. The *CDF* of true labels rank for *random* lies on the diagonal line. We observe that atomic structure based kernels significantly
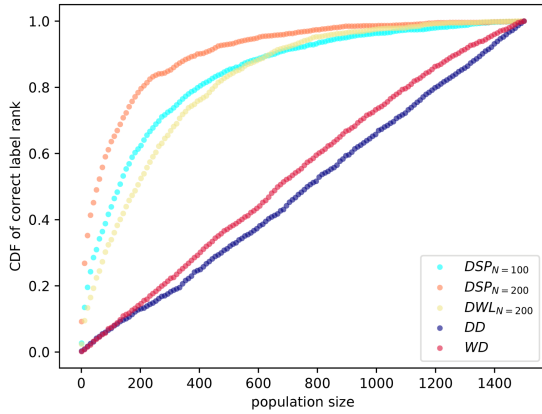
Fig. 7. *CDF* of true rank over the population according to different kernels.

outperform the random baseline performance by defining a meaningful similarity ranking across the mobility networks.

The best overall performance is achieved by the *DSP* kernel on graphs pruned to 200 nodes. In particular, this kernel places the true identity among the closest 10 networks for 10% of the individuals, and among the closest 200 networks for 80% of the population. The Shortest-Path kernel has an intuitive interpretation in the case of mobility networks, since its atomic substructures take into account the hop distances among the locations in a user's mobility network and the popularity categories of the departing and arrival location. The deep variant can also account for variation in the level of such substructures, which are more realistic when considering the stochasticity in the mobility patterns inherent to our dataset.

The best performance of the Weisfeiler-Lehman kernel is achieved by its deep variant for $h = 2$ iterations of the *WL* test for a mobility network pruned to 200 nodes. This phenomenon is explainable via the statistical properties of the mobility networks. As we saw in Section 4.3, the networks display power law degree distribution and small diameters. Taking into account the steps of the *WL* test, it is clear that these topological properties will lead the node relabeling scheme to cover the entire network after a very small number of iterations. Thus local structural patterns will be described by few features produced in the first iterations of the test. Furthermore, the feature space of the kernel increases very quickly as a function of h, which leads to sparsity and low levels of similarity over the population of networks.

Histograms of length 1 000 were also computed for the unweighted and weighted degree distributions and passed through a Gaussian RBF kernel. We can see that the degree distribution gives almost a random ranking, as it is heavily dependent on the network size. When including the normalized edge weights, the *WD* kernel only barely outperforms a random ranking. Repetitions on pruned versions did not improve the performance and are not presented for brevity.

Based on the insights obtained from our experiment, we can make the following observations with respect to attributes of individual mobility and their impact on the identifiability of networks:

- **Location pruning**: Reducing the number of nodes (locations) in a mobility network does not necessarily make it more privacy-preserving. On the contrary, if location pruning is done by keeping the most frequently visited locations, it can enhance reidentification. In our experiments we obtain similar, or even enhanced, performance for graph kernels when applying them on increasingly pruned networks with size down to 100 locations.
- **Transition pruning**: Including very rare transitions in longitudinal mobility does not add discriminative information. We consistently obtained better results when truncating the long tail of edge weight distribution, which led us to analyze versions of the networks where 40% of the weakest edges were removed.
- **Frequency information of locations**: The frequency of visits to nodes in the mobility network allows better ranking by kernels which support node attributes, e.g. Weisfeiler-Lehman and Shortest-Path kernel. This information should follow a coarse scheme, in order to compensate for the temporal variation of location popularity in mobility networks.
- **Directionality of transitions**: Directionality generally enhances the identifiability of networks and guides the similarity computation when using Shortest-Path kernels.

5.4 Quantification of Privay Loss

The Deep Shortest-Path kernel on top−200 networks offers the best ranking of identities for the test networks. As observed in Figure 8, the mean of the true rank has been shifted from 750 to 140 for our population. In addition, the variance is much smaller: approximately 218, instead of 423 for the random ordering.

The obtained ordering implies a significant decrease in user privacy, since the ranking can be leveraged by an adversary to determine the most likely matches between a training mobility network and a test mobility network. The adversary can estimate the true identity of a given test network G', as suggested in Section 3.4.2, applying some simple probabilistic policy that uses pairwise similarity information. For example, let us examine the privacy loss implied by update rule in (5) for function f:

$$f\big(K_{\mathrm{DSP}}(G_i, G')\big) = \frac{1}{rank\big(K_{DSP}(G_i, G')\big)}. \quad (7)$$

This means that the adversary updates her probability estimate for the identity corresponding to a test network, by assigning to each possible identity a probability that is inversely proportional to the rank of the similarity between the test network and the training network corresponding to the identity.

From equation (6), we can compute the induced privacy loss for each test network, and the statistics of privacy loss over the networks of the Device Analyzer population. Figure 9 demonstrates considerable privacy loss with a median of 2.52. This means that the informed adversary can achieve a median deanonymization probability 3.52 times higher than an uninformed adversary. Moreover, the positive mean of privacy loss (≈ 27) means that the probabilities of the true identities of the test networks have, on average, much higher values in the adversarial estimate compared to the uninformed random strategy. Hence, revealing the kernel values makes an adversarial attack easier.
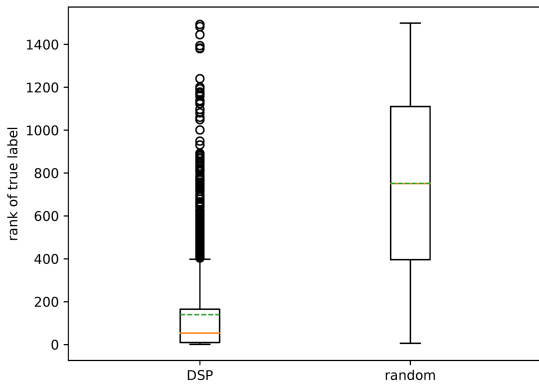
5.5 Defense Mechanisms

The demonstrated privacy leakage motivates the quest for defense mechanisms against this category of attacks. There are a variety of techniques which we could apply in order to *reduce the recurring patterns of an individual's mobility network over time* and *decrease the diversity of mobility networks across a population*, and therefore enhance the privacy inherent in these graphs. Examples include noise injection on network structure via several strategies: randomization of node attributes, perturbations of network edges or node removal. It is currently unclear how effective such techniques will be, and what trade-off can be achieved between utility in mobility networks and the privacy guarantees offered to individuals whose data the graphs represent. Moreover, it seems appropriate to devise kernel-agnostic techniques, suitable for generic defense mechanisms. For example, it is of interest to assess the resistance of our best similarity metric to noise, as the main purpose of deep graph kernels is to be robust to small dissimilarities at the substructure level.

We think this study is important for one further reason: kernel-based methods allow us to apply a rich toolbox of learning algorithms without accessing the original datapoints, or their feature vectors, but instead by using their kernel matrix. Thus studying the anonymity associated with kernels is valuable for ensuring that such learning systems do not leak privacy of the original data. We leave this direction to future work.
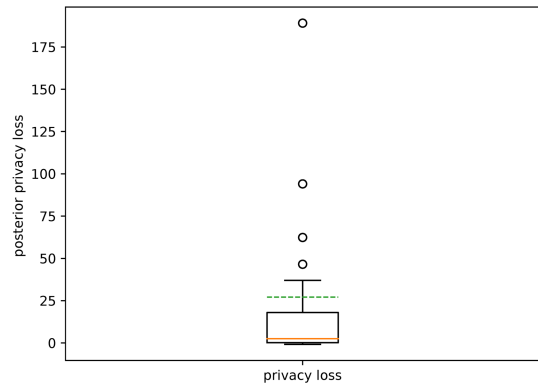
Fig. 8. Boxplot of rank for the true labels of the population according to a Deep Shortest-Path kernel and to a random ordering.

Fig. 9. Privacy loss over the test data of our population for an adversary adopting the informed policy of (7). Median privacy loss is 2.52.

6 Conclusions & Future Work

In this paper we have shown that the mobility networks of individuals exhibit significant diversity and the topology of the mobility network itself, without labels, may be unique and therefore uniquely identifying.

An individual's mobility network is dynamic over time. Therefore, an adversary with access to mobility data of a person from one time period cannot simply test for graph isomorphism to find the same user in a dataset recorded at a later point in time. Hence we proposed graph kernel methods to detect structural similarities between two mobility networks, and thus provide the adversary with information on the likelihood that two mobility networks represent the same individual. While graph kernel methods are imperfect predictors, they perform significantly better than a random strategy and therefore our approach induces significant privacy loss. Our approach does not make use of geographic information or fine-grained temporal information and therefore it is immune to commonly adopted privacy-preserving techniques of geographic information removal and temporal cloaking, and thus our method may lead to new mobility deanonymization attacks.

Moreover, we find that reducing the number of nodes (locations) or edges (transitions between locations) in a mobility network does not necessarily make the network more privacy-preserving. Conversely, the frequency of node visits and the direction of transition in a mobility network does enhance the identifiablility of a mobility network for some graph kernel methods. We provide empirical evidence that neighborhood relations in the high-dimensional spaces generated by deep graph kernels remain meaningful for our networks [3]. Further work is needed to shed more light on the geometry of those spaces in order to derive the optimal substructures and dimensionality required to support best graph matching. More work is also required to understand the sensitivity of our approach to the time period over which mobility networks are constructed. There is also an opportunity to explore better ways of exploiting pairwise distance information.

Apart from emphasizing the vulnerability of popular anonymization techniques based on user-specific location pseudonymization, our work provides insights into network features that can facilitate the identifiability of location traces. Our framework also opens the door to new anonymization techniques that can apply structural similarity methods to individual traces in order to cluster people with similar mobility behaviour.

This approach may then support statistically faithful population mobility studies on mobility networks with $k-$anonymity guarantees to participants.

Apart from graph kernel similarity metrics, tools for network deanonymization can also be sought in the direction of graph mining: applying heavy subgraph mining techniques [4] or searching for persistent cashcades [20]. Frequent substructure pattern mining (gSpan [40]) and discriminative frequent subgraph mining (CORK [33]) techniques can also be considered.

Our methodology is, in principle, applicable to all types of data where individuals transition between a set of discrete states. Therefore, one of our immediate goals is to evaluate the performance of such retrieval strategies on different categories of datasets, such as web browsing histories or smartphone application usage sequences.

A drawback of our current approach is that it cannot be directly used to mimic individual or group mobility by synthesizing traces. Fitting a generative model on mobility traces and then defining a kernel on this model may provide better anonymity, and therefore privacy, and it would also support the generation of artificial traces which mimic the mobility of users.

Ethics Statement

Acknowledgments

References

[1] Charu C. Aggarwal and Philip S. Yu. 2008. A General Survey of Privacy-Preserving Data Mining Models and Algorithms. In *Privacy-Preserving Data Mining*, Charu C. Aggarwal, Philip S. Yu, and Ahmed K. Elmagarmid (Eds.). The Kluwer International Series on Advances in Database Systems, Vol. 34. Springer US, 11–52. DOI: http://dx.doi.org/10.1007/978-0-387-70992-5_2

[2] Berker Agir, Kévin Huguenin, Urs Hengartner, and Jean-Pierre Hubaux. 2016. On the Privacy Implications of Location Semantics. *PoPETs* 2016, 4 (2016), 165–183. DOI: http://dx.doi.org/10.1515/popets-2016-0034

[3] Kevin S. Beyer, Jonathan Goldstein, Raghu Ramakrishnan, and Uri Shaft. 1999. When Is "Nearest Neighbor" Meaningful?. In *Proceedings of the 7th International Conference on Database Theory (ICDT '99)*. Springer-Verlag, London, UK, UK, 217–235. http://dl.acm.org/citation.cfm?id=645503.656271

[4] Petko Bogdanov, Misael Mongiovì, and Ambuj K. Singh. 2011. Mining Heavy Subgraphs in Time-Evolving Networks. In *Proceedings of the 2011 IEEE 11th International Conference on Data Mining (ICDM '11)*. IEEE Computer Society, Washington, DC, USA, 81–90. DOI: http://dx.doi.org/10.1109/ICDM.2011.101

[5] Karsten M. Borgwardt and Hans-Peter Kriegel. 2005. Shortest-Path Kernels on Graphs. In *Proceedings of the Fifth IEEE International Conference on Data Mining (ICDM 2005)*. IEEE Computer Society, Washington, DC, USA, 74–81. http://dx.doi.org/10.1109/ICDM.2005.132

[6] Yves-Alexandre de Montjoye, César A. Hidalgo, Michel Verleysen, and Vincent D. Blondel. 2013. Unique in the Crowd: The privacy bounds of human mobility. *Scientific reports* 3, 1 (dec 2013), 1376. DOI: http://dx.doi.org/10.1038/srep01376

[7] Yoni De Mulder, George Danezis, Lejla Batina, and Bart Preneel. 2008. 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*. 23–32. DOI: http://dx.doi.org/10.1145/1456403.1456409

[8] Cynthia Dwork, Frank McSherry, Kobbi Nissim, and Adam Smith. 2006. Calibrating noise to sensitivity in private data analysis. In *Theory of Cryptography Conference*. Springer, 265–284.

[9] Sébastien Gambs, Marc-Olivier Killijian, and Miguel Núñez Del Prado Cortez. 2014. De-anonymization Attack on Geolocated Data. *J. Comput. Syst. Sci.* 80, 8 (Dec. 2014), 1597–1614. DOI: http://dx.doi.org/10.1016/j.jcss.2014.04.024

[10] Philippe Golle and Kurt Partridge. 2009. *On the Anonymity of Home/Work Location Pairs*. Springer Berlin Heidelberg, Berlin, Heidelberg, 390–397. DOI: http://dx.doi.org/10.1007/978-3-642-01516-8_26

[11] Marco Gruteser and Dirk Grunwald. 2003. Anonymous Usage of Location-Based Services Through Spatial and Temporal Cloaking. In *Proceedings of the 1st International Conference on Mobile Systems, Applications and Services (MobiSys '03)*. ACM, New York, NY, USA, 31–42. DOI: http://dx.doi.org/10.1145/1066116.1189037

[12] David Haussler. 1999. *Convolution kernels on discrete structures*. Technical Report. Technical report, Department of Computer Science, University of California at Santa Cruz.

[13] Jong Hee Kang, William Welbourne, Benjamin Stewart, and Gaetano Borriello. 2005. Extracting places from traces of locations. *ACM SIGMOBILE Mobile Computing and Communications Review* 9, 3 (2005), 58. DOI: http://dx.doi.org/10.1145/1094549.1094558

[14] Juha K. Laurila, Jan Blom, Olivier Dousse, Daniel Gatica-Perez, Olivier Bornet, Julien Eberle, Imad Aad, and Markus Miettinen. The mobile data challenge: Big data for mobile computing research," in Proc. MDC Workshop, 2012.

[15] Ninghui Li, Tiancheng Li, and Suresh Venkatasubramanian. 2007. t-closeness: Privacy beyond k-anonymity and l-diversity. In *Data Engineering, 2007. ICDE 2007. IEEE 23rd International Conference on*. IEEE, 106–115.

[16] Miao Lin, Hong Cao, Vincent W. Zheng, Kevin Chen-Chuan Chang, and Shonali Krishnaswamy. 2015. Mobile user verification/identification using statistical mobility profile. In *2015 International Conference on Big Data and Smart Computing, BIGCOMP 2015, Jeju, South Korea, February 9-11, 2015*. 15–18. DOI: http://dx.doi.org/10.1109/35021BIGCOMP.2015.7072841

[17] Ashwin Machanavajjhala, Daniel Kifer, Johannes Gehrke, and Muthuramakrishnan Venkitasubramaniam. 2007. L-diversity: Privacy Beyond K-anonymity. *ACM Trans. Knowl. Discov. Data* 1, 1, Article 3 (March 2007). DOI: http://dx.doi.org/10.1145/1217299.1217302

[18] Brendan D. McKay and Adolfo Piperno. 2014. Practical graph isomorphism, II. *Journal of Symbolic Computation* 60, 0 (2014), 94 – 112. DOI: http://dx.doi.org/10.1016/j.jsc.2013.09.003

[19] Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781* (2013).

[20] Steven Morse, Marta C. Gonzalez, and Natasha Markuzon. 2016. Persistent cascades: Measuring fundamental communication structure in social networks. In *2016 IEEE International Conference on Big Data, BigData 2016, Washington DC, USA, December 5-8, 2016*. 969–975. DOI: http://dx.doi.org/10.1109/BigData.2016.7840695

[21] Farid M Naini, Jayakrishnan Unnikrishnan, Patrick Thiran, and Martin Vetterli. 2016. Where You Are Is Who You Are: User Identification by Matching Statistics. *IEEE Transactions on Information Forensics and Security* 11, 2 (feb 2016), 358–372. DOI: http://dx.doi.org/10.1109/TIFS.2015.2498131

[22] Arvind Narayanan and Vitaly Shmatikov. 2008. Robust De-anonymization of Large Sparse Datasets. In *Proceedings of the 2008 IEEE Symposium on Security and Privacy (SP '08)*. IEEE Computer Society, Washington, DC, USA, 111–125. DOI: http://dx.doi.org/10.1109/SP.2008.33

[23] Arvind Narayanan and Vitaly Shmatikov. 2009. De-anonymizing social networks. In *Security and Privacy, 2009 30th IEEE Symposium on*. IEEE, 173–187.

[24] Lukasz Olejnik, Claude Castelluccia, and Artur Janc. 2014. On the uniqueness of Web browsing history patterns. *Annales des Télécommunications* 69, 1-2 (2014), 63–74. DOI: http://dx.doi.org/10.1007/s12243-013-0392-5

[25] Andreas Pfitzmann and Marit Hansen. 2010. A terminology for talking about privacy by data minimization: Anonymity, Unlinkability, Undetectability, Unobservability, Pseudonymity, and Identity Management. http://dud.inf.tu-dresden.de/literatur/Anon_Terminology_v0.34.pdf. (Aug. 2010). http://dud.inf.tu-dresden.de/literatur/Anon_Terminology_v0.34.pdf v0.34.

[26] Apostolos Pyrgelis, Carmela Troncoso, and Emiliano De Cristofaro. 2017. What Does The Crowd Say About

You? Evaluating Aggregation-based Location Privacy. *arXiv preprint arXiv:1703.00366* (2017).

[27] Luca Rossi, Matthew J. Williams, Christoph Stich, and Mirco Musolesi. 2015. Privacy and the City: User Identification and Location Semantics in Location-Based Social Networks. In *Proceedings of the Ninth International Conference on Web and Social Media, ICWSM 2015, University of Oxford, Oxford, UK, May 26-29, 2015*. 387–396. http://www.aaai.org/ocs/index.php/ICWSM/ICWSM15/paper/view/10498

[28] Ingo Scholtes. 2017. When is a Network a Network?: Multi-Order Graphical Model Selection in Pathways and Temporal Networks. In *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD '17)*. ACM, New York, NY, USA, 1037–1046. DOI:http://dx.doi.org/10.1145/3097983.3098145

[29] Kumar Sharad and George Danezis. 2014. An Automated Social Graph De-anonymization Technique. In *Proceedings of the 13th Workshop on Privacy in the Electronic Society (WPES '14)*. ACM, New York, NY, USA, 47–58. DOI:http://dx.doi.org/10.1145/2665943.2665960

[30] Nino Shervashidze, Pascal Schweitzer, Van Leeuwen, Erik Jan, Kurt Mehlhorn, and Karsten Borgwardt. 2011. Weisfeiler-Lehman graph kernels. *Journal of Machine Learning Research* 12 (2011), 2539–2561. DOI:http://dx.doi.org/10.1.1.232.1510

[31] Reza Shokri, Carmela Troncoso, Claudia Diaz, Julien Freudiger, and Jean-Pierre Hubaux. 2010. Unraveling an old cloak: k-anonymity for location privacy. In *Proceedings of the 9th annual ACM workshop on Privacy in the electronic society*. ACM, 115–118.

[32] Latanya Sweeney. 2002. k-anonymity: A model for protecting privacy. *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems* 10, 05 (2002), 557–570.

[33] Marisa Thoma, Hong Cheng, Arthur Gretton, Jiawei Han, Hans Peter Kriegel, Alex Smola, Le Song, Philip S. Yu, Xifeng Yan, and Karsten M. Borgwardt. 2010. Discriminative frequent subgraph mining with optimality guarantees. *Statistical Analysis and Data Mining* 3, 5 (2010), 302–318. DOI:http://dx.doi.org/10.1002/sam.10084

[34] S.V.N. Vishwanathan, Nicol Schraudolph, Risi Kondor, and K.M. Borgwardt. 2010. Graph Kenrels. *Journal of Machine Learning Research* 11 (2010), 1201–1242. DOI:http://dx.doi.org/10.1142/9789812772435_0002

[35] Daniel T. Wagner, Andrew Rice, and Alastair R. Beresford. 2014. *Device Analyzer: Understanding Smartphone Usage*. Springer International Publishing, Cham, 195–208. DOI:http://dx.doi.org/10.1007/978-3-319-11569-6_16

[36] Boris Weisfeiler and AA Lehman. 1968. A reduction of a graph to a canonical form and an algebra arising during this reduction. *Nauchno-Technicheskaya Informatsia* 2, 9 (1968), 12–16.

[37] Pascal Welke, Ionut Andone, Konrad Blaszkiewicz, and Alexander Markowetz. 2016. Differentiating Smartphone Users by App Usage. In *Proceedings of the 2016 ACM International Joint Conference on Pervasive and Ubiquitous Computing (UbiComp '16)*. ACM, New York, NY, USA, 519–523. DOI:http://dx.doi.org/10.1145/2971648.2971707

[38] Fengli Xu, Zhen Tu, Yong Li, Pengyu Zhang, Xiaoming Fu, and Depeng Jin. 2017. Trajectory Recovery From Ash: User Privacy Is NOT Preserved in Aggregated Mobility Data. In *Proceedings of the 26th International Conference on World Wide Web*. International World Wide Web Conferences Steering Committee, 1241–1250.

[39] Jian Xu, Thanuka L. Wickramarathne, and Nitesh V. Chawla. 2016. Representing higher-order dependencies in networks. *Science Advances* 2, 5 (2016), e1600028–e1600028. DOI:http://dx.doi.org/10.1126/sciadv.1600028

[40] Xifeng Yan and Jiawei Han. 2002. gSpan: Graph-Based Substructure Pattern Mining. In *Proceedings of the 2002 IEEE International Conference on Data Mining (ICDM '02)*. IEEE Computer Society, Washington, DC, USA, 721–. http://dl.acm.org/citation.cfm?id=844380.844811

[41] Pinar Yanardag and S.V.N. Vishwanathan. 2015. Deep Graph Kernels. In *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD '15)*. ACM, New York, NY, USA, 1365–1374. DOI:http://dx.doi.org/10.1145/2783258.2783417

[42] Ting-Fang Yen, Yinglian Xie, Fang Yu, Roger (Peng) Yu, and Martin Abadi. 2012. Host Fingerprinting and Tracking on the Web:Privacy and Security Implications, In The 19th Annual Network and Distributed System Security Symposium (NDSS) 2012. (February 2012). https://www.microsoft.com/en-us/research/publication/host-fingerprinting-and-tracking-on-the-webprivacy-and-security-implications/

[43] Hui Zang and Jean Bolot. 2011. Anonymization of Location Data Does Not Work: A Large-scale Measurement Study. In *Proceedings of the 17th Annual International Conference on Mobile Computing and Networking (MobiCom '11)*. ACM, New York, NY, USA, 145–156. DOI:http://dx.doi.org/10.1145/2030613.2030630

[44] Elena Zheleva and Lise Getoor. 2009. To join or not to join: the illusion of privacy in social networks with mixed public and private user profiles. In *Proceedings of the 18th International Conference on World Wide Web, WWW 2009, Madrid, Spain, April 20-24, 2009*. 531–540.