Analyzing Privacy Risk in Human Mobility Data

Roberto Pellungrini[1], Luca Pappalardo[2], Francesca Pratesi[1,2], and Anna Monreale[1]

[1] Department of Computer Science, University of Pisa, Italy
[2] ISTI-CNR, Pisa, Italy

Abstract. Mobility data are of fundamental importance for understanding the patterns of human movements, developing analytical services and modeling human dynamics. Unfortunately, mobility data also contain individual sensitive information, making it necessary an accurate privacy risk assessment for the individuals involved. In this paper, we propose a methodology for assessing privacy risk in human mobility data. Given a set of individual and collective mobility features, we define the minimum data format necessary for the computation of each feature and we define a set of possible attacks on these data formats. We perform experiments computing the empirical risk in a real-world mobility dataset, and show how the distributions of the considered mobility features are affected by the removal of individuals with different levels of privacy risk.

1 Introduction

In the last years, human mobility analysis has attracted a growing interest due to its importance in a wide range of applications, from urban management and public health [13], to the discovery of quantitative patterns [12] and the prediction of human future whereabouts [8]. The worrying side of this story is that human mobility data are sensitive, because they may allow the re-identification of individuals and lead to severe privacy issues if analyzed with malicious intent [18]. In order to prevent these problems, researchers have developed methodologies, frameworks and algorithms to reduce the individual privacy risk associated to the analysis of human mobility data [1]. Tools like the one presented in [15] try to balance both the individuals' privacy protection and the effectiveness of the analytical results.[3]Starting from [15], we study the empirical trade-off between individual privacy risk and data quality w.r.t. a set of state-of-the-art individual and collective mobility measures. We first introduce a set of mobility data structures, each with a different level of detail on an individual's mobility history, and then present a set of re-identification attacks based on these structures. In a scenario where a data owner wants to share human mobility data with an external entity (e.g., a data analyst), it can simulate the re-identification attacks to assess the privacy risk of every individual in the dataset. Having this information, the data owner can simply delete the individuals beyond a certain

[3] In compliance with the new EU General Data Protection Regulation

threshold of privacy risk or select the most suitable privacy-preserving technique (e.g., based on k-anonymity, differential privacy) to mitigate individual privacy risk. We use a real-world human mobility dataset to compute the distribution of privacy risk for every re-identification attack. We then compare the distributions of the considered mobility features computed on the original data and on data obtained removing high risk individuals. We show how these distributions vary much less when computed on more aggregated structures.

2 Individual Mobility Features

The approach we present in this paper is tailored for human mobility data, i.e., data describing the movements of a set of individuals during a period of observation. The mobility dynamics of an individual can be described by a set of measures widely used in literature. Some measures describe specific aspects of an individual's mobility; other measures describe an individual's mobility in relation to collective mobility. The Maximum Distance is defined as the length of the longest trip of an individual during the period of observation [24]. The Sum Of Distances is the sum of all the trip lengths traveled by the individual during the period of observation [24]. The Radius of Gyration is the characteristic distance traveled by an individual during the period of observation, formally defined in [12]; this measure represents one of the major components useful for describing human mobility. The Mobility Entropy is a measure of the predictability of an individual's trajectory; formally, it is defined as the Shannon entropy of an individual's movements [7]. We can also define some measures related to locations instead of individuals, like the Location Entropy, i.e., the predictability of who visits the location. We also use Location Density, a measure of how many individuals have that location as their most visited location, and the Flow of a location defined as the number of trips that have that location as origin or destination.

3 Data Definitions

Human mobility data is generally collected in an automatic way through electronic devices (e.g., mobile phones, GPS devices) in form of raw trajectory data. A raw trajectory of an individual is a sequence of records identifying the movements of that individual during the period of observation [26]. Every record has the following fields: the identifier of the individual, a geographic location expressed in coordinates (generally latitude and longitude), a timestamp indicating when the individual stopped in or went through that location. Depending on the specific application, a raw trajectory can be aggregated into different mobility data structures introduced in the following.

Definition 1 (Trajectory). *The trajectory T_u of an individual u is a temporally ordered sequence of tuples $T_u = \langle(l_1, t_1), (l_2, t_2), \dots, (l_n, t_n)\rangle$, where $l_i = (x_i, y_i)$ is a location, x_i and y_i are the coordinates of the geographic location, and t_i is the corresponding timestamp, $t_i < t_j$ if $i < j$.*

Definition 2 (Frequency vector). *The frequency vector W_u of an individual u is a sequence of tuples $W_u = \langle (l_1, w_1), (l_2, w_2), \dots, (l_n, w_n) \rangle$ where $l_i = (x_i, y_i)$ is a location, w_i is the frequency of the location, i.e., how many times location l_i appears in the individual's trajectory T_u, and $w_i > w_j$ if $i < j$. A frequency vector W_u is hence an aggregation of a trajectory T_u.*

Definition 3 (Probability vector). *The probability vector P_u of an individual u is a sequence of tuples $P_u = \langle (l_1, p_1), (l_2, p_2), \dots, (l_n, p_n) \rangle$, where $l_i = (x_i, y_i)$ is a location, p_i is the probability that location l_i appears in W_u, i.e., $p_i = \frac{w_i}{\sum_{l_i \in W_u} w_i}$, and $p_i > p_j$ if $i < j$. A probability vector P_u is hence an aggregation of a frequency vector W_u.*

In the following, with the terms *visit* we refer indifferently to a tuple in a trajectory or in a frequency or probability vector. In other words, a visit indicates a pair consisting of a location and a supplementary information, e.g., the timestamp or the frequency. We denote with D a mobility dataset, i.e., a set of a one of the above data types (trajectory, frequency or probability vectors). Each data structure allows the computation of some of the mobility features presented in Section 2: with the trajectory, the most detailed of the three structures, we can compute all the mobility features presented. With the vector structures we can compute only Radius of Gyration, User Entropy, Location Entropy and Location Density. Lowering the detail of the structure we can compute less features but we expose less information about the individuals represented.

4 Privacy Risk Assessment Model

Several methodologies have been proposed in literature for privacy risk assessment. In this paper we start from the framework proposed in [15], which allows for the assessment of the privacy risk inherent to human mobility data. At the core of this framework, there is the identification of the minimum data structure, the definition of a set of possible attacks that a malicious adversary might conduct in order to re-identify her target and the simulation of the attacks. The privacy risk of an individual is related to her probability of re-identification in a mobility dataset w.r.t. a set of re-identification attacks. A re-identification attack assumes that an adversary gains access to a mobility dataset, then, on the basis of some background knowledge about an individual, i.e., the knowledge of a subset of her mobility data, the adversary tries to re-identify all the records in the dataset regarding the individual under attack. In this paper we use the definition of privacy risk (or re-identification risk) introduced in [19].

There can be many background knowledge categories, every category may have several background knowledge configurations, every configurations has many instances. A background knowledge category is a kind of information known by the adversary about a specific set of dimensions of an individual's mobility data. Typical dimensions in mobility data are space, time, frequency of visiting a location and probability of visiting a location. Examples of background knowledge categories are a subset of the locations visited by an individual and specific times an individual visited those locations. The number k of the elements of a category known by the adversary is called background knowledge configuration: an example is the knowledge by the adversary of $k = 3$ locations of an individual. Finally, an instance of background knowledge is the specific knowledge of the adversary, such as a visit in a specific location. We formalize these concepts as follows.

Definition 4 (Background knowledge configuration). *Given a background knowledge category $\mathcal{B}$, we denote with $B_k \in \mathcal{B} = \{B_1, B_2, \dots, B_n\}$ a specific background knowledge configuration, where k represents the number of elements in $\mathcal{B}$ known by the adversary. We define an element $b \in B_k$ as an* instance *of background knowledge configuration.*

Let $\mathcal{D}$ be a database, D a mobility dataset extracted from $\mathcal{D}$ (e.g., a data structure as defined in Section 3), and D_u the set of records representing individual u in D, we define the probability of re-identification as follows.

Definition 5 (Probability of re-identification). *The probability of re-identification $PR_D(d = u|b)$ of an individual u in a mobility dataset D is the probability to associate a record $d \in \mathcal{D}$ to an individual u, given an instance of background knowledge configuration $b \in B_k$.*

Note that $PR_D(d{=}u|b) = 0$ if the individual u is not represented in D. Since each instance $b \in B_k$ has its own probability of re-identification, we define the risk of re-identification of an individual as the maximum probability of re-identification over the set of instances of a background knowledge configuration.

Definition 6 (Risk of re-identification or Privacy risk). The risk of re-identification (or privacy risk) of an individual u given a background knowledge configuration B_k is her maximum probability of re-identification $Risk(u, D) = \max PR_D(d{=}u|b)$ for $b \in B_k$. The risk of re-identification has the lower bound $\frac{|D_u|}{|D|}$ (a random choice in D), and $Risk(u, D) = 0$ if $u \notin D$.

4.1 Privacy attacks on mobility data

In this section we describe the attacks we use in this paper.

Location In a Location attack the adversary knows a certain number of locations visited by the individual but she does not know the temporal order of the visits.This is similar to considering the locations as items of transactions [22] with the difference that a transaction is a set of items and not a multiset (an individual might visit the same location multiple times). Given an individual s, we denote by $L(T_s)$ the multiset of locations $l_i \in T_s$ visited by s. The background knowledge category of a Location attack is defined as follows.

Definition 7 (Location background knowledge). *Let k be the number of locations l_i of an individual s known by the adversary. The Location background* *knowledge is a set of configurations based on k locations, defined as $B_k = L(T_s)^{[k]}$. Here $L(T_s)^{[k]}$ denotes the set of all the possible k-combinations of the elements in set $L(T_s)$.*

Given $b \in B_k$, we can give the definition for the set of users matching the Location background knowledge, and consequently, the probability of re-identification.

Definition 8 (Location attack). *Let $b \in B_k$ be the adversary Location background knowledge. We define by $R = \{u \in U | b \subseteq L(T_u)\}$ the candidate set of users whose trajectory contains the instance b. The probability of re-identification of the user u is $\frac{1}{|R|}$.*

Location Sequence In a Location Sequence attack [9] the adversary knows a subset of the locations and the temporal ordering of the visits. Given an individual s, we denote by $L(T_s)$ the sequence of locations $l_i \in T_s$ visited by s. The background knowledge category of a Location Sequence attack is the following.

Definition 9 (Location Sequence background knowledge). *Let k be the number of locations l_i of a individual s known by the adversary. The Location Sequence background knowledge is a set of configurations based on k locations, defined as $B_k = L(T_s)^{[k]}$, where $L(T_s)^{[k]}$ denotes the set of all the possible k-subsequences of the elements in set $L(T_s)$.*

The set of users matching this background knowledge is defined in the following where we denote by $a \preceq b$ that a is a subsequence of b. .

Definition 10 (Location Sequence attack). *Let $b \in B_k$ be the Location Sequence background knowledge. We define by $R = \{u \in U | b \preceq L(T_u)\}$ the candidate set of users whose trajectory contains the combination b. The probability of re-identification of the user u is $\frac{1}{|R|}$.*

Visit In a Visit attack [25] an adversary knows a subset of the locations visited by the individual and the time the individual visited these locations.

Definition 11 (Visit background knowledge). *Let k be the number of visits v of a individual s known by the adversary. The Visit background knowledge is a set of configurations based on k visits, defined as $B_k = T_s^{[k]}$ where $T_s^{[k]}$ denotes the set of all the possible k-subsequences of the elements in trajectory T_s.*

We recall that in the case of trajectories we denote by visit $v \in T$ the pair (l_i, t_i) composed by the location l_i and its timestamp t_i. Formally, the set of all trajectories supporting b from both a spatial and a temporal point of view is:

Definition 12 (Visit attack). *Let $b \in B_k$ be the Visit background knowledge. We define by $R = \{u \in U \mid \forall\ (l_i, t_i)\ \in b,\ \exists\ (l_i^u, t_i^u)\ \in T_u\ .\ l_i = l_i^u \wedge t_i \leq t_i^u\}$ the candidate set of users whose trajectories contain b. The probability of re-identification of the user u is $\frac{1}{|R|}$.*

Frequent Location, Frequent Location Sequence We also introduce two attacks based on the knowledge of the location applied to vectors. The Frequent Location attack is similar to the Location attack but here a location can appear only once, so it follows the same principle of [22]. In the Frequent Location Sequence attack the adversary knows a subset of the locations visited by an individual and the relative ordering w.r.t. the frequencies (from most frequent to least frequent). This attack is similar to the Location Sequence attack, with two differences: a location can appear only once and locations are ordered by descending frequency. We omit the definitions of the background knowledge and attacks because they are similar to the ones defined on trajectories.

Frequency We introduce an attack where an adversary knows the locations visited by the individual, their reciprocal ordering of frequency, and the minimum number of visits of the individual in the locations. This means that, when searching for specific subsequences, the adversary must consider also subsequences containing the known locations with a greater frequency. We recall that in the case of frequency vectors we denote by visit $v \in W$ the pair (l_i, w_i) composed by the frequent location l_i and its frequency w_i. The background knowledge category of a Frequency attack is defined as follows.

Definition 13 (Frequency background knowledge). *Let k be the number of visits v of the frequency vector of individual s known by the adversary. The Frequency background knowledge is a set of configurations based on k visits, defined as $B_k = W_s^{[k]}$ where $W_s^{[k]}$ denotes the set of all possible k-combinations of frequency vector W_s.*

The set of users matching a single $b \in B_k$ is defined as follows.

Definition 14 (Frequency attack). *Let $b \in B_k$ be the Frequency background knowledge. We define by $R = \{u \in U \mid \forall\ (l_i, w_i)\ \in b,\ \exists\ (l_i^u, w_i^u)\ \in W_u\ .\ l_i = l_i^u \wedge w_i \leq w_i^u\}$ the candidate set of users whose frequency vectors contain the instance b. The probability of re-identification of the user u is $\frac{1}{|R|}$.*

Home & Work In the Home & Work attack [27], the adversary knows the two most frequent locations of an individual and their frequencies. This is the only attack where the background knowledge configuration is just a single 2-combination. Mechanically, this attack is identical to the Frequency attack.

Probability In a Probability attack an adversary knows the locations visited by an individual and the probability for that individual to visit each location. This attack is similar to the one introduced by [28], but we cannot rely on matching algorithms on bipartite graph because the length of the probability vectors is not the same among the individuals and is greater than the length of the background knowledge configuration instances. We recall that in the case of probability vectors we denote by visit $v \in P$ the pair (l_i, p_i) composed by the frequent location l_i and its probability p_i. The background knowledge category for this attack is defined as follows.

Definition 15 (Probability background knowledge). *Let k be the number of visits v of the probability vector of individual s known by the adversary. The Probability background knowledge is a set of configurations based on k visits, defined as $B_k = P_s^{[k]}$ where $P_s^{[k]}$ denotes the set of all possible k-combinations of probability vector P_s.*

Again, the set of users matching a single $b \in B_k$ can be defined as follows.

Definition 16 (Probability attack). *Let $b \in B_k$ be the Probability background knowledge. We define by $R = \{u \in U \mid \forall\ (l_i, p_i)\ \in b,\ \exists\ (l_i^u, p_i^u)\ \in P_u\ .\ l_i = l_i^u\ \wedge\ p_i \in [p_i^u - \delta, p_i^u + \delta]\}$ the candidate set of users who in their frequency vectors contain the instance b tolerating for the probability match a tolerance δ. The probability of re-identification of the user u is $\frac{1}{|R|}$.*

Proportion We introduce an attack assuming that an adversary knows a subset of locations and the relative proportion between the number of visits to these locations, i.e., between the frequency of the most frequent known location and the frequency of the other known locations. Given a set of visits $X \subset W$ we denote by $l1$ the most frequent location of X and with w_1 its frequency. We also denote by pr_i the proportion between w_i and w_1 for each $v_i \neq v_1 \in X$, and denote by LR a set of frequent locations l_i with their respective pr_i. The background knowledge category for this attack is defined as follows.

Definition 17 (Proportion background knowledge). *Let k be the number of locations l_i of an individual s known by the adversary. The Proportion background knowledge is a set of configurations based on k locations, defined as $B_k = LR_s^{[k]}$ where $LR_s^{[k]}$ denotes the set of all possible k-combinations of the frequent locations l_i with associated pr_i.*

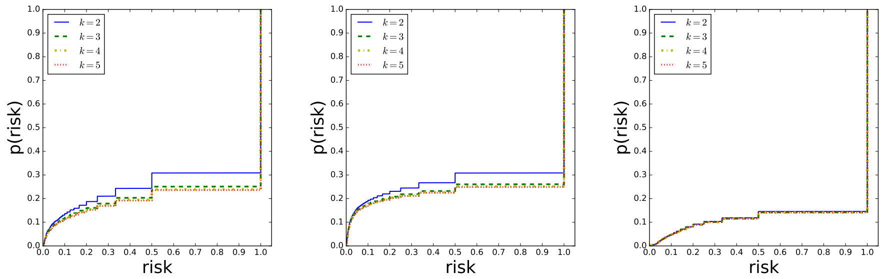
The set of users matching a single $b \in B_k$ is defined as follows.

Definition 18 (Proportion attack). *Let $b \in B_k$ be the Proportion background knowledge. We define by $R = \{u \in U \mid \forall\ (l_i, pr_i)\ \in b,\ \exists\ (l_i^u, pr_i^u) \in LR^u\ .\ l_i = l_i^u\ \wedge\ pr_i \in [pr_i^u - \delta, pr_i^u + \delta]\}$ the candidate set of users who in their frequency vectors compatible with b. Note that δ is a tolerance factor for the matching of proportions. The probability of re-identification of the user u is $\frac{1}{|R|}$.*

Note that each attack is associated with a specific data structure: Location, Location Sequence and Visit require the trajectory data structure; Frequent Location, Frequent Location Sequence and Frequency require the frequency vector; Home & Work, Proportion and Probability require the probability vector.

5 Experiments

For all the attacks defined except the Home & Work attack we consider four sets of background knowledge configuration B_k with $k = 2, 3, 4, 5$, while for the Home & Work attack we have just one possible background knowledge configuration,



(a) Location (b) Location sequence (c) Visit (day)

Fig. 1: Cumulative distributions for trajectory attacks.

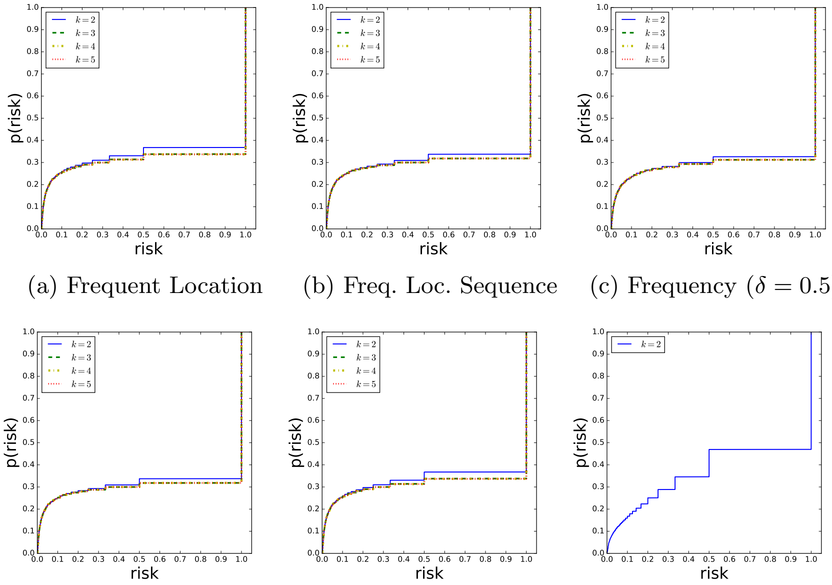
where the adversary knows the two most frequent locations of an individual. Note that for the Visit attack we considered only the day as time frame for the granularity of the attack. We use a dataset provided by Octo Telematics[4] storing the GPS tracks of 9,715 private vehicles traveling in Florence from 1st May to 31st May 2011, corresponding to 179,318 trajectories. We assign each origin and destination point of trajectories to the corresponding census cells [12] provided by the Italian National Statistics Bureau. This allows us to describe the mobility of every vehicle in terms of a trajectory as defined in Section 3. We performed a simulation of the attacks computing the privacy risk values for all individuals in the dataset and for all B_k.[5] We then show the distribution of the mobility features presented in Section 2 at varying levels of risk: we compare the distribution of the features computed on the original dataset, i.e., the dataset with the complete set of trajectories, with the distributions obtained using only trajectories belonging to individuals below certain thresholds of privacy risk.

5.1 Privacy Risk Simulations

We simulated attacks using $k = 2, 3, 4, 5$: the cumulative distribution functions for the trajectory attacks are depicted in Figure 1, where we can see that the privacy risk increase not only with increasing the amount of knowledge (from Figure 1 (a) to Figure 1 (c)), but also with increasing k. This is more evident for the Location attack and the Location Sequence attack (Figure 1(a) and (b) respectively). It is interesting to note that the greater gap is present, especially for the Location attack, varying k from 2 to 3, i.e., the greatest increasing of risk of re-identification occurs when the quantity of information known is lower. This implies that adding the same absolute amount of information, i.e., one single location, has less influence if the attacker already has a quite big knowledge. For the Visit attack (Figure 1 (c)), since here the background knowledge is already enough detailed, we can see that the increasing of k does not change so much

[4] https://www.octotelematics.com/

[5] The Python code for attacks simulation is available here: https://github.com/pellungrobe/privacy-mobility-lib



(a) Frequent Location (b) Freq. Loc. Sequence (c) Frequency ($\delta = 0.5$)

(d) Probability ($\delta = 0.1$) (e) Proportion ($\delta = 0.1$) (f) Home&work

Fig. 2: Cumulative distributions for frequency vector attacks.

the levels of privacy risk. The number of individuals with maximum risk of re-identification, i.e., equals to 1, ranges from 60% for the Location attack to more that 80% for the Visit attack, while we do have an increase in the number of individuals with risk of re-identification of 50% (or less) across the board.

Observing Figure 2, regarding attacks on vectors, the levels of risk decrease slightly from the attacks on trajectories. Moreover, it is clear how the the cumulative distribution function of the risk of re-identification is quite stable varying k or changing the category of knowledge. This can probably be due to the fact that, with vectors, we are dealing with distinct locations for each individual, thus, since many individuals have few distinct locations, the risk remains very similar when increasing k. With Home & Work attack (2 (f)) we have significantly lower risk. Indeed, we can observe much lower levels of risk in general, even if 50% of users still have maximum risk of re-identification.

5.2 Correlations Between Measures and Privacy Risk

In this section we want to show the correlation between the mobility measures introduced in Section 2 and the levels of risk calculated for each attack. The Pearson Correlation Coefficient is a measure of the linear dependence between two variables, in this case a mobility measure and the risk assessed for each attack. It ranges from -1 to +1 where -1 indicates total negative linear correlation, 0 indicates no linear correlation and +1 indicates total positive linear correlation. Since in Section 5.1 we saw that, varying k, privacy risk does not change

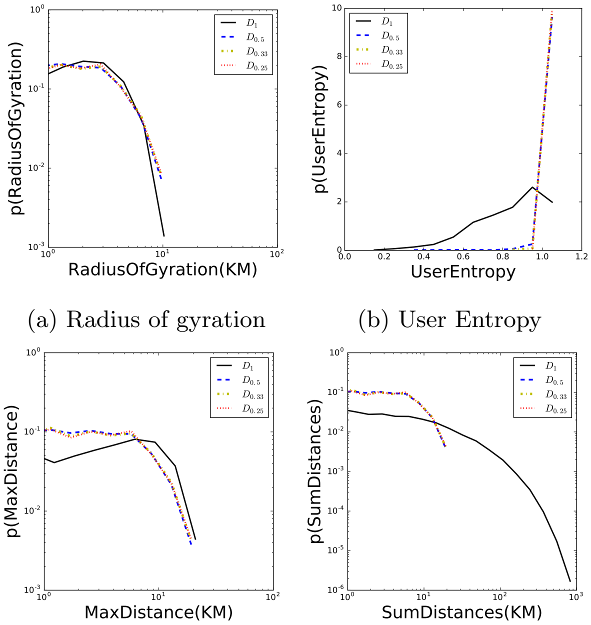
	RadiusGyration	**UserEntropy**	**MaxDistance**	**SumDistances**
Location	0.408326	0.654331	0.503459	0.352364
Location Sequence	0.333477	0.668218	0.463041	0.367661
Visit (Day)	0.219840	0.493934	0.320390	0.256473
Frequent Location	0.359895	0.749976	0.501241	0.426581
Freq.Loc. Sequence	0.352399	0.746065	0.490765	0.414132
Frequency	0.340739	0.733594	0.482271	0.410859
Probability	0.352399	0.746065	0.490765	0.414132
Proportion	0.359895	0.749976	0.501241	0.426581

Table 1: Correlation of Measures and Privacy Risk

too severely, we show the correlation only for a middle value, i.e. $k = 3$. We used only the features related to individuals and not the ones related to locations, because the privacy risk level is computed for each individual and does not have an association with locations. We show the results of correlation study in Table 1. Analyzing the attacks on trajectories, there is really no strong correlation. An interesting fact, which is compliant with the results showed in Section 5.1, is that the correlation tends to decrease as the levels of risk increase, thus, for the Visit attack, we observe a drop in the correlation coefficient. Another interesting result is that, especially for the attacks related to frequency and probability vectors the correlation between User Entropy and risk of re-identification is higher while no other strong correlation can be found among the various measures. So overall it seems that high levels of entropy correlate to high levels of risk.

5.3 Measure Distributions by Risk Levels

In this section we present an analysis on the distributions of mobility measures on the datasets used in the experiment, w.r.t. the changing levels of risk. We compare the distributions of the various measures and see how they vary with the levels of risk. We removed from the dataset individuals above a certain level of risk and then recomputed the measures. Thus, we obtained a set of distributions for each measure, one for each level of risk and attack. However, due to space limitations, we present the results only for two of them: the Visit and Frequency attacks. These are the two most representative of the differences between the attacks performed on different data structures, since they are two of the most powerful. For both attacks we show how each measure behaves with different levels of risk, comparing their distributions. For both datasets and for all possible attacks we selected four thresholds of risk. Then, we systematically eliminated from the original dataset users with a risk beyond the thresholds, obtaining four different derived datasets: the original dataset D_1 and $D_{0.5}$, $D_{0.33}$, $D_{0.25}$ obtained removing individuals with risk greater than 0.5, 0.33 and 0.25 respectively. Regarding the background knowledge configuration, we selected the



(a) Radius of gyration

(b) User Entropy

(c) Max Distance

(d) Sum of Distances

Fig. 3: Pdf of user related measures changing levels of risk (Visit attack (day))

risk calculated with $k = 2$. This for several reasons: it is a reasonable number of locations that an attacker might know, it is the level of risk that shows the most appreciable changes from one threshold of risk to the other in terms of users excluded/included, and it is also the k value that yields the lower levels of risk. In the following, we show the probability density functions (pdf) of the mobility features for the different datasets.

For the Visit attack with day precision for the time frame, Figure 3 reports results on users related measures. We observe some interesting results: User Entropy (Figure 3 (b)) becomes 1 for almost all remaining users in $D_{0.5}$, $D_{0.33}$ and $D_{0.25}$. Observing the Radius of gyration (Figure 3 (a)) we note that the shape of the distribution remains fairly similar but we find more individuals with high Radius of gyration proportionally to the total number of remaining individuals. For the Sum of Distances (Figure 3 (d)), we tend to lose the individuals who traveled the longest distances total. For Max Distance (Figure 3 (c)) the distribution remain substantially similar. Figure 4 shows that the distributions of location related measures for both datasets suffer heavy modifications. For Location Entropy (Figure 4-left) we observe a loss of the middle values: we have a significantly higher probability of locations with very low entropy (< 0.2) and a slight peak of locations with very high entropy, with no relevant values in between. This is also more evident the more we cut the data, i.e. for $D_{0.25}$. For Location Density and Flow (Figure 4-center & Figure 4-right) we observe a loss of the higher values but the overall shape of the distributions remains similar.

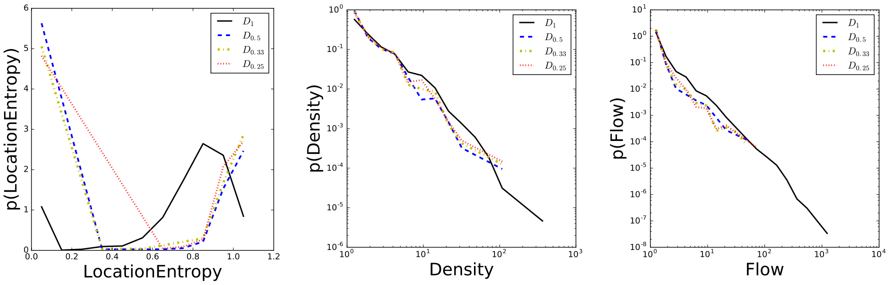


Fig. 4: Pdf of location related measures changing levels of risk (Visit attack (day))

Changing data structure from trajectory to frequency vector, we already observed in Section 5.1 generally lower levels of risk, thus we can maintain more individuals in the dataset cutting at the same thresholds. For this reason, we expect more similar distributions w.r.t. the original dataset. However, since we lose the information about the specific movements given by the trajectory structure, we cannot compute all the measures introduced in Section 2. The measures that we cannot compute are: Max Distance, Sum of Distances and Flow. For the frequency attack we show the results for individuals and locations related measures in Figure 5 and Figure 6 respectively. While User Entropy distribution (Figure 5-right) still exhibits some changes w.r.t. the original distribution at changing levels of risk, we observe less dramatic differences in comparison to the distributions presented in Figure 4-center regarding the Visit attack. For Location Entropy distribution (Figure 6-left) we still observe a peak of locations with very low entropy but the overall shape of the distributions is closer to the original one, maintaining similar peaks around higher values. Location Density (Figure 6-right) and Radius of gyration (Figure 5-left) distributions appear to remain almost identical for all thresholds of risk ($D_{0.5}$, $D_{0.33}$ and $D_{0.25}$). Summarizing, the distributions presented above give an empirical demonstration to the intuition that less detailed data structures, exposing less data about an individual, lead to generally lower levels of re-identification risk. Thus, for the considered features, choosing the minimum required data structure is fundamental to im-

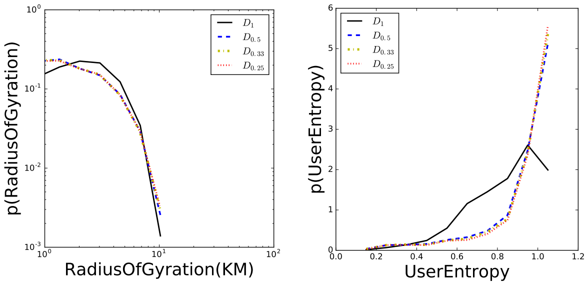


Fig. 5: Pdf of user related measures changing levels of risk (Frequency attack)

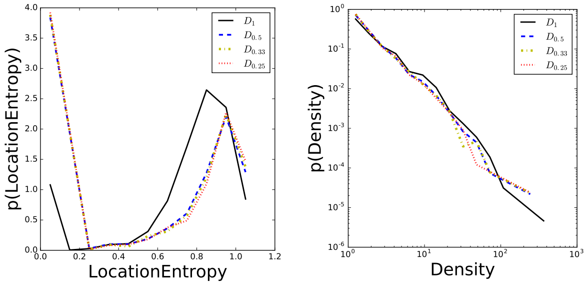


Fig. 6: Pdf of location related measures changing levels of risk (Frequency attack)

prove the quality of the distributions of the mobility features we want to study when computing them from sanitized datasets.

6 Related Work

To overcome privacy leaks, many techniques have been proposed in literature. A widely used privacy-preserving model is k-anonymity [19], which requires that an individual should not be identifiable from a group of size smaller than k based on their quasi-identifiers (QIDs), i.e., a set of attributes that can be used to uniquely identify individuals. Assuming that adversaries own disjoint parts of a trajectory, [22] reduces privacy risk by relying on the suppression of the dangerous observations from each individual's trajectory. In [25], authors propose the attack-graphs method to defend against attacks, based on k-anonymity. Other works are based on the differential privacy model [6]. [10] and [14] considers the problem of privacy on aggregations of movement data. [4] proposes to publish a contingency table of trajectory data, where each cell contains the number of individuals commuting from a source to a destination. One of the most important work about privacy risk assessment is the Linddun methodology [5], a privacy-aware framework, useful for modeling privacy threats in software-based systems. In the last years, different techniques for risk management have been proposed, such as NIST's Special Publication 800-30 [21] and SEI's OCTAVE [2]. Unfortunately, many of these works simply include privacy considerations when assessing the impact of threats. In [23], authors elaborate an entropy-based method to evaluate the disclosure risk of personal data, trying to manage quantitatively privacy risks. [11] studies the effect of co-location information on location privacy, considering an adversary such as a social network operator accessing to such information. The *unicity* measure proposed in [20] evaluates the privacy risk as the number of records which are uniquely identified. [3] proposes a risk-aware framework for information disclosure in tabular data supporting runtime risk assessment, using adaptive anonymization as risk-mitigation method. Lastly, in [15] authors introduced a privacy risk assessment framework specific for mobility data. Although this framework suffers from a high computational complexity, it is effective in many mobility scenarios. Other papers addressing the problem of measuring privacy risk in mobility data are [17, 16].

7 Conclusion

Human mobility data contain highly sensitive information that might lead to serious violations of individual privacy. In this paper we explored a repertoire of re-identification attacks that can be conducted on mobility data, analyzing the empirical privacy risk of thousands of individuals in a real-world mobility dataset. The considered attacks were designed for three common mobility data formats: trajectories, frequency vectors and probability vectors. Through experimentation on the real-world dataset, we observed on average high level of risk across the different types of re-identification attack. We then characterize how the distributions of state-of-the-art human mobility measures changes as individuals with high level of risk are deleted from the dataset, finding two main results: (1) higher privacy risk is related to a higher distortion of the distributions of mobility measures; (2) selecting the minimum required data structure can lead to significant improvements in the overall levels of privacy risk, while guaranteeing distributions of mobility features closer to the distributions derived from the original data. We observe that the methodology experimented in this paper may be applied, without changing the attacks definitions to any dataset of mobility and sequence data; clearly, in this last case instead of locations we would have events. As future work, we plan to investigate how distributions of mobility features can be further improved using privacy transformations more sophisticated than the simple suppression of individuals with high privacy risk.

Acknowledgment

Funded by the European project SoBigData (Grant Agreement 654024).

References

1. O. Abul, F. Bonchi, and M. Nanni. Never Walk Alone: Uncertainty for Anonymity in Moving Objects Databases. In *ICDE 2008*. 376–385.
2. C. Alberts, S. Behrens, R. Pethia, and W. Wilson. 1999. *Operationally Critical Threat, Asset, and Vulnerability Evaluation (OCTAVE) Framework, Version 1.0.* CMU/SEI-99-TR-017. Software Engineering Institute, Carnegie Mellon University. http://resources.sei.cmu.edu/library/asset-view.cfm?AssetID=13473
3. A. Armando, M. Bezzi, N. Metoui, and A. Sabetta. Risk-Based Privacy-Aware Information Disclosure. *Int. J. Secur. Softw. Eng.* 6, 2 (April 2015), 70–89.
4. G. Cormode, C. M. Procopiuc, D. Srivastava, and T. T. L. Tran. 2012. Differentially private summaries for sparse data. In *ICDT '12*. 299–311.
5. M. Deng, K. Wuyts, R. Scandariato, B. Preneel, and W. Joosen. 2011. A privacy threat analysis framework: supporting the elicitation and fulfillment of privacy requirements. *Requir. Eng.* 16, 1, pp 3–32.
6. C. Dwork, F. McSherry, K. Nissim, and A. Smith. 2006. Calibrating Noise to Sensitivity in Private Data Analysis. In *TCC '06*. 265–284.
7. N. Eagle and A. S. Pentland. Eigenbehaviors: identifying structure in routine. *Behavioral Ecology and Sociobiology* 63, 7 (2009), 1057–1066.

8. S. Gambs, M. O. Killijian, and M. N. del Prado Cortez. Next Place Prediction Using Mobility Markov Chains. In *MPM 2012*. Art 4.
9. N. Mohammed, B. C.M. Fung, and M. Debbabi. Walking in the Crowd: Anonymizing Trajectory Data for Pattern Analysis. In *CIKM 2009*. 1441–1444.
10. A. Monreale, W. H. Wang, F. Pratesi, S. Rinzivillo, D. Pedreschi, G. Andrienko, and N. Andrienko. 2013. *Privacy-Preserving Distributed Movement Data Aggregation*. Springer International Publishing, 225–245.
11. A.M. Olteanu, K. Huguenin, R. Shokri, M. Humbert, J.P. Hubaux. *Quantifying interdependent privacy risks with location data.* IEEE Transactions on Mobile Computing, 16(3), pp.829-842, 2017
12. L. Pappalardo, F. Simini, S. Rinzivillo, D. Pedreschi, F. Giannotti, and A.-L. Barabasi. Returners and explorers dichotomy in human mobility. *Nature Communications* 6 (08 09 2015).
13. L. Pappalardo, M. Vanhoof, L. Gabrielli, Z. Smoreda, D. Pedreschi, and F. Giannotti. An analytical framework to nowcast well-being using mobile phone data. *International Journal of Data Science and Analytics* 2, 1 (2016), 75–92.
14. Pyrgelis, A., De Cristofaro, E. and Ross, G.J. *Privacy-friendly mobility analytics using aggregate location data.* In SIGSPATIAL international conference on advances in geographic information systems (p. 34), 2016
15. F. Pratesi, A. Monreale, R. Trasarti, F. Giannotti, D. Pedreschi, and T. Yanagihara. *PRUDEnce: a System for Assessing Privacy Risk vs Utility in Data Sharing Ecosystems.* To appear at Transactions on Data Privacy Journal.
16. L. Rossi, M. Musolesi. *It's the way you check-in: identifying users in location-based social networks.* newblock In ACM Conf. on Online social networks (pp. 215-226).
17. L. Rossi, J. Walker, M. Musolesi. *Spatio-temporal techniques for user identification by means of GPS mobility data.* EPJ Data Science, 4(1), p.11, 2015.
18. I. S. Rubinstein. Big Data: The End of Privacy or a New Beginning? *International Data Privacy Law* (2013).
19. P. Samarati and L. Sweeney. 1998a. Generalizing Data to Provide Anonymity when Disclosing Information (Abstract). In *PODS*. 188.
20. Y. Song, D. Dahlmeier, and S. Bressan. Not So Unique in the Crowd: a Simple and Effective Algorithm for Anonymizing Location Data. In *PIR@SIGIR 2014*. 19–24.
21. G. Stoneburner, A. Goguen, and A. Feringa. 2002. *Risk Management Guide for Information Technology Systems: Recommendations of the National Institute of Standards and Technology.* NIST special publication, Vol. 800.
22. M. Terrovitis and N. Mamoulis. 2008. Privacy Preservation in the Publication of Trajectories. In *MDM*. 65–72.
23. S. Trabelsi, V. Salzgeber, M. Bezzi, and G. Montagnon. 2009. Data disclosure risk evaluation. In *CRiSIS '09*. 35–72.
24. N. E. Williams, T. A. Thomas, M. Dunbar, N. Eagle, and A. Dobra. Measures of Human Mobility Using Mobile Phone Records Enhanced with GIS Data. *PLoS ONE* 10, 7 (2015), 1–16.
25. R. Yarovoy, F. Bonchi, L. V. S. Lakshmanan, and W. H. Wang. 2009. Anonymizing moving objects: how to hide a MOB in a crowd?. In *EDBT*. 72–83.
26. Y. Zheng. Trajectory Data Mining: An Overview. In *ACM TIST* 6, 3 (2015).
27. H. Zang and J. Bolot. Anonymization of Location Data Does Not Work: A Large-scale Measurement Study. In *MobiCom 2011*. 145–156.
28. J. Unnikrishnan and F. M. Naini. De-anonymizing private data by matching statistics. In *Allerton 2013*. 1616–1623.