A Data Mining Approach to Assess Privacy Risk in Human Mobility Data

ROBERTO PELLUNGRINI, Department of Computer Science, University of Pisa, Italy
LUCA PAPPALARDO and FRANCESCA PRATESI, Department of Computer Science, University of Pisa, Italy – ISTI-CNR, Pisa, Italy
ANNA MONREALE, Department of Computer Science, University of Pisa, Italy

Human mobility data are an important proxy to understand human mobility dynamics, develop analytical services, and design mathematical models for simulation and what-if analysis. Unfortunately mobility data are very sensitive since they may enable the re-identification of individuals in a database. Existing frameworks for privacy risk assessment provide data providers with tools to control and mitigate privacy risks, but they suffer two main shortcomings: (i) they have a high computational complexity; (ii) the privacy risk must be recomputed every time new data records become available and for every selection of individuals, geographic areas, or time windows. In this article, we propose a fast and flexible approach to estimate privacy risk in human mobility data. The idea is to train classifiers to capture the relation between individual mobility patterns and the level of privacy risk of individuals. We show the effectiveness of our approach by an extensive experiment on real-world GPS data in two urban areas and investigate the relations between human mobility patterns and the privacy risk of individuals.

CCS Concepts: • **Security and privacy** → **Pseudonymity, anonymity and untraceability**; **Usability in security and privacy**; • **Computing methodologies** → *Classification and regression trees*; Transfer learning;

Additional Key Words and Phrases: Human mobility, data mining, privacy

ACM Reference format:
Roberto Pellungrini, Luca Pappalardo, Francesca Pratesi, and Anna Monreale. 2017. A Data Mining Approach to Assess Privacy Risk in Human Mobility Data. *ACM Trans. Intell. Syst. Technol.* 9, 3, Article 31 (December 2017), 27 pages.
https://doi.org/10.1145/3106774

1 INTRODUCTION

Human mobility analysis has attracted in the past decade a growing interest from different disciplines due to its importance in a wide range of applications, ranging from urban planning and transportation engineering (Wang et al. 2012; Pappalardo et al. 2015; Marchetti et al. 2015; Pappalardo et al. 2016) to public health (Colizza et al. 2007; Tizzoni et al. 2014). The availability of massive collections of mobility data and the development of sophisticated techniques for their analysis and mining (Zheng and Zhou 2011; Zheng 2015) have offered the unprecedented

This work has been partially funded by the European project SoBigData RI (Grant Agreement 654024).

Authors' addresses: R. Pellungrini, L. Pappalardo, and A. Monreale, Computer Science Department, University of Pisa, Italy; emails: roberto.pellungrini@di.unipi.it, luca.pappalardo@isti.cnr.it, anna.monreale@unipi.it; F. Pratesi, KDDLab - ISTI - CNR, Via G. Moruzzi 1, Pisa, Italy; email: francesca.pratesi@isti.cnr.it.

ACM 2157-6904/2017/12-ART31 $15.00
https://doi.org/10.1145/3106774

opportunity to observe human mobility at large scales and in great detail, leading to the discovery of the fundamental quantitative patterns of human mobility (Gonzalez et al. 2008; Pappalardo et al. 2013; Song et al. 2010b; Pappalardo et al. 2015), accurate predictions of future human whereabouts (Gambs et al. 2012; Lu et al. 2013), and the mathematical modeling of the main aspects of human mobility dynamics (Jiang et al. 2016; Pappalardo et al. 2016; Pappalardo and Simini 2016; Song et al. 2010a). These analyses are generally conducted on large datasets storing detailed information about the spatio-temporal points visited by individuals in a territory, like GPS tracks (Bazzani et al. 2010; Giannotti et al. 2011; Pappalardo et al. 2013, 2015) or mobile phone data (Gonzalez et al. 2008; Song et al. 2010b; Simini et al. 2012; Pappalardo et al. 2015). It goes without saying that mobility data are sensitive because people's whereabouts might reveal intimate personal information or allow the reidentification of individuals in a database, creating serious privacy risks (Rubinstein 2013). For example it has been shown that just four spatio-temporal points can be enough to uniquely identify 95% of individuals in a mobility dataset (de Montjoye et al. 2013). Therefore, if mobility data are analyzed with malicious intent, there can be a serious violation of the privacy rights of the individuals involved.

Driven by these sensitive issues, in recent years, researchers from different disciplines have developed algorithms, methodologies, and frameworks to mitigate the individual privacy risks associated with the analysis of GPS trajectories, mobile phone data, and Big Data in general (Abul et al. 2008a; Monreale et al. 2014b; Wong et al. 2007). These tools aim at preserving both the right to privacy of individuals and the quality of the analytical results. However, to enable a practical application of the privacy-preserving techniques proposed in the literature, it is necessary to find a trade-off between privacy protection and data quality. To this aim Pratesi et al. (2016) proposes a framework for the privacy risk assessment of individuals in a mobility dataset. This framework is compliant with the new EU General Data Protection Regulation, which explicitly imposes on data controllers an assessment of the impact of data protection for the most risky processes.[1]

Although frameworks like the one presented in Pratesi et al. (2016) are proved to be effective in many mobility scenarios, they suffer a major limitation: The privacy risk assessment has a high computational complexity because it requires a computation of the maximum risk of reidentification (or privacy risk) given the external knowledge that a malicious adversary might use in conducting an attack. The generation of the external knowledge is nonpolynomial in time since it considers all the possible ways the adversary can try to reidentify an individual in a mobility dataset. The computational complexity is a severe limitation because the privacy risks must be recomputed every time new data become available and for every selection of individuals, geographic areas, and periods of time.

In this article, we propose a data mining approach for privacy risk assessment that overcomes the computational shortcomings of existing frameworks. We first introduce a repertoire of reidentification attacks on mobility data and then use a data mining classifier to predict the level of privacy risk for an individual based solely on her mobility patterns. We evaluate our approach on real-world mobility data with an extensive experiment. Starting from a dataset of around 1 million GPS tracks produced by 12,000 private vehicles traveling in two urban areas in Italy during one month, we extract individual mobility patterns and compute the privacy risk level associated with vehicles according to the repertoire of reidentification attacks. We then train data mining classifiers and use them to determine (in polynomial time) the privacy risk level of *previously unseen* vehicles whose data were *not* used in the learning phase, based just on their individual mobility patterns. In a scenario where a Data Analyst requests a Data Provider for mobility data to develop an analytical service, the Data Provider (e.g., a mobile phone carrier) can use the classifiers

[1] The EU General Data Protection Regulation can be found at http://bit.ly/1TlgbjI.

to immediately identify risky individuals (i.e., individuals with a high level of privacy risk). Then, the Data Provider can select the most suitable privacy-preserving technique (e.g., k-anonymity, differential privacy) to reduce their privacy risk and release safe data to the Data Analyst. Although our approach is constrained to a fixed set of re-identification attacks, it can be easily extended to any type of attack defined on human mobility data.

Our experiments on GPS data show two main results. First, the classifiers are accurate in classifying the privacy risk level of unseen individuals in the two urban areas. The classifiers' predictions are particularly accurate in classifying the lowest and the highest levels of privacy risk, allowing an immediate distinction between safe individuals and risky individuals. In particular, we observe a high recall (99%) on the class of maximum privacy risk, meaning that the probability of misclassifying a high-risk individual as a low-risk individual is negligible. The second remarkable result is that the classifiers built on one urban area are effective when used to determine the privacy risk level of individuals in the *other* urban area. This suggests that the predictive models are able to infer rather general relationships between mobility patterns and privacy risk, which are independent of the number of individuals, the width of the geographic area, and the length of the period of observation. This means that the Data Provider can reuse the same classifiers for every selection of the dataset without the need to redo the training process every time. Finally, we quantify the impact of every individual mobility measure on the classifiers, observing that it changes with the type of re-identification attack considered (i.e., different attacks are based on gathering information about different mobility patterns). Based on the results, we think our work provides two main contributions. First, we show that we can effectively use data mining to estimate the privacy risk of individuals in a fast, accurate, and precise way, overcoming the computational issues related to existing frameworks. Second, we shed light on the relationships between individual human mobility patterns and risk of re-identification, which was not clearly investigated in the literature.

The article is organized as follows. In Section 2, we define the data structures to describe human mobility data according to different data aggregations. In Section 3, we introduce the framework used for the privacy risk assessment, while Section 4 describes the data mining approach we propose. In Section 5, we show the results of our experiments, and we discuss them in Section 6. Section 7 presents the main works related to our article, and, finally, Section 8 concludes the article by proposing some lines of new research.

2 DATA DEFINITIONS

The approach we present in this article is tailored for human mobility data: data describing the movements of a set of individuals during a period of observation. This type of data is generally collected in an automatic way through electronic devices (e.g., mobile phones, GPS devices) in the 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 (Zheng and Zhou 2011; Zheng 2015). Every record has the following fields: the identifier of the individual, a geographic location expressed in coordinates (generally latitude and longitude), and 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:

Definition 2.1 (Trajectory). The trajectory T_u of an individual u is a temporally ordered sequence of tuples $T_u = \langle (l_1, t_1), (l_2, t_2), \ldots, (l_n, t_n) \rangle$, 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 2.2 (Frequency vector). The frequency vector W_u of an individual u is a sequence of tuples $W_u = \langle (l_1, w_1), (l_2, w_2), \ldots, (l_n, w_n) \rangle$ 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 2.3 (Probability vector). The probability vector P_u of an individual u is a sequence of tuples $P_u = \langle (l_1, p_1), (l_2, p_2), \ldots, (l_n, p_n) \rangle$, 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.

Definition 2.4 (Mobility Dataset). A mobility dataset is a set of mobility data structures $D = \{S_1, S_2, \ldots, S_n\}$ where S_u $(1 \leq u \leq n)$ is the mobility data structure of individual u. For example, a mobility dataset can be a set of trajectories $\{T_1, \ldots, T_n\}$, a set of frequency vectors $\{W_1, \ldots, W_n\}$, or a set of probability vectors $\{P_1, \ldots, P_n\}$. Note that the three sets have the same size n.

In the following, using the terms *visit* or *point*, we refer indifferently to a tuple in a trajectory, a tuple in a frequency vector, or a tuple in a probability vector. In other words, a visit v_i indicates a pair consisting of a location l_i and a supplementary information (e.g., the timestamp t_i, the frequency w_i, or the probability p_i of the location). Moreover, we denote by $U_{set} = \{u_1, \ldots, u_n\}$ the set of distinct individuals and by $L_{set} = \{l_1, \ldots, l_m\}$ the set of distinct locations in a mobility dataset D. In this article, we assume that mobility data are represented with one of the data structures just described.

3 PRIVACY RISK ASSESSMENT FRAMEWORK

Several methodologies have been proposed in the literature for privacy risk assessment. In this article, we consider the framework proposed in Pratesi et al. (2016), which allows for the assessment of the privacy risk inherent to human mobility data. The framework considers a scenario where a Data Analyst requests a Data Provider human mobility data in order to develop an analytical service. For its part, the Data Provider has to guarantee the right to privacy of the individuals whose data are recorded. As a first step, the Data Analyst communicates to the Data Provider the data requirements for the analytical service. Assuming that the Data Provider stores a database $\mathcal{D}$, it aggregates, selects, and filters the dataset $\mathcal{D}$ to meet the requirements of the Data Analyst and produces a set of mobility datasets $\{D_1, \ldots, D_z\}$ each with a different data structure and/or aggregation of the data. The Data Provider then reiterates a four-step procedure until it considers the data delivery safe:

Procedure 3.1: DATA DELIVERY PROCEDURE by the Data Provider

(1) *Identification of Attacks*: identify a set of possible attacks that a malicious adversary might conduct in order to re-identify the individuals in the mobility datasets $\{D_1, \ldots, D_z\}$;

(2) ***Privacy Risk Computation***: simulate the attacks and compute the set of privacy risk values for every individual in the mobility datasets $\{D_1, \ldots, D_z\}$;

(3) *Dataset Selection*: select a mobility dataset $D \in \{D_1, \ldots, D_z\}$ with the best trade-off between the privacy risks of the individuals and the quality of the data, given a certain level of tolerated privacy risk and the data requirements by the Data Analyst;

(4) *Risk Mitigation and Data delivery*: apply a privacy-preserving transformation (e.g., generalization, randomization, etc.) on the chosen mobility dataset D to eliminate the residual privacy risk, producing a filtered mobility dataset D_{filt}. Deliver the mobility dataset D_{filt} to the Data Analyst when the D_{filt} is adequately safe.

In this article, we focus on improving Step (2) of the Data Delivery Procedure (i.e., Privacy Risk Computation), which is the most critical one from a computational point of view. Computing the privacy risk of an individual means simulating several possible attacks that a malicious adversary can perform and computing the privacy risks associated with each attack. The privacy risk of an individual is related to her probability of re-identification in a mobility dataset with respect to a set of re-identification attacks. A re-identification attack assumes that an adversary gains access to a mobility dataset. 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 reidentify all the records in the dataset regarding the individual under attack. In this article, we use the definition of privacy risk (or re-identification risk) introduced in Samarati and Sweeney (1998a), Samarati (2001), and Sweeney (2002) and widely used in the literature. There can be many background knowledge categories, every category may have several background knowledge configurations, and every configuration may have 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 (Section 2). Two examples of background knowledge categories are a subset of the locations visited by an individual (spatial dimension) and the specific times an individual visited those locations (spatial and temporal dimensions). The number k of the elements of a category known by the adversary is called the *background knowledge configuration*. An example of background knowledge configuration is the knowledge by the adversary of $k = 3$ locations of an individual. Finally, an instance of background knowledge is the specific information known by the adversary, such as a visit in a specific location. We formalize these concepts as follows:

Definition 3.1 (Background knowledge configuration). Given a background knowledge category $\mathcal{B}$, we denote with $B_k \in \mathcal{B} = \{B_1, B_2, \ldots, B_n\}$ 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.

Example 3.2. Suppose a trajectory $T_u = \langle v_1, v_2, v_3, v_4 \rangle$ of an individual u is present in the Data Provider's dataset D, where $v_i = (l_i, t_i)$ is a visit, l_i is a location, and t_i the time when u visited location l_i, with $i = 1, \ldots, 4$ and $t_i < t_j$ if $i < j$. Based on T_u, the Data Provider can generate all the possible instances of a background knowledge configuration that an adversary might use the re-identify the whole trajectory T_u. Considering the knowledge by the adversary of ordered subsequences of locations and $k = 2$, we obtain the background knowledge configuration $B_2 = \{(v_1, v_2), (v_1, v_3), (v_1, v_4), (v_2, v_3), (v_2, v_4), (v_3, v_4)\}$. The adversary, for example, might know instance $b = (v_1, v_4) \in B_2$ and aim at detecting all the records in D regarding individual u in order to reconstruct the whole trajectory T_u.

Let $\mathcal{D}$ be a database, D a mobility dataset extracted from $\mathcal{D}$ as an aggregation of the data on specific dimensions (e.g., an aggregated data structure and/or a filtering on time and/or space), and D_u the set of records representing individual u in D; we define the probability of re-identification as follows:

Definition 3.3 (Probability of re-identification). Given an attack, a function *matching(d, b)* indicating whether or not a record $d \in D$ matches the instance of background knowledge configuration $b \in B_k$, and a function $M(D, b) = \{d{\in}D | matching(d, b) = True\}$, we define the *probability of re-identification* of an individual u in dataset D as:

$$PR_D(d = u|b) = \frac{1}{|M(D,b)|,}$$

that is the probability to associate a record $d \in D$ to an individual u, given instance $b \in B_k$.

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 3.4 (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 $u \notin D$.

To clarify the concepts of probability of re-identification and privacy risk, we provide the following example that, given a mobility dataset D of trajectories, shows how we can compute the two measures for a specific attack.

Example 3.5. Consider a set of individuals $U_{set}{=}\{u_1, u_2, u_3, u_4, u_5, u_6\}$ and the corresponding dataset D of trajectories:

$D = \{$

$T_{u_1} = \langle (2011/02/03, Lucca), (2011/02/03, Leghorn), (2011/02/03, Pisa), (2011/02/04, Florence) \rangle$

$T_{u_2} = \langle (2011/02/03, Lucca), (2011/02/03, Pisa), (2011/02/04, Lucca), (2011/02/04, Leghorn) \rangle$

$T_{u_3} = \langle (2011/02/03, Leghorn), (2011/02/03, Pisa), (2011/02/04, Lucca), (2011/02/04, Florence) \rangle$

$T_{u_4} = \langle (2011/02/04, Pisa), (2011/02/04, Leghorn), (2011/02/04, Florence) \rangle$

$T_{u_5} = \langle (2011/02/04, Pisa), (2011/02/04, Florence), (2011/02/05, Lucca) \rangle$

$T_{u_6} = \langle (2011/02/04, Lucca), (2011/02/04, Leghorn) \rangle$

$\}$

Assume an adversary wants to perform an attack on individual u_1 knowing only the locations she visited (without any information about the time), with background knowledge configuration B_2 (i.e., the adversary knows two of the locations visited by individual u_1). We compute the risk of re-identification of individual u_1, given the dataset D of trajectories and the knowledge of the adversary, in two steps:

(1) We compute the probability of re-identification for every possible instance $b{\in}B_2$. Instance $b{=}\{Lucca, Leghorn\}$ has a probability of re-identification $PR_D(d{=}u_1|\{Lucca, Leghorn\}){=}\frac{1}{4}$ because the pair $\{Lucca, Leghorn\}$ appears in trajectories T_{u_1}, T_{u_2}, T_{u_3}, and T_{u_6} (i.e., in a total of four trajectories). Instance $\{Lucca, Pisa\}$ has a probability of re-identification $PR_D(d{=}u_1|\{Lucca, Pisa\}){=}\frac{1}{4}$ because the pair appears in four trajectories T_{u_1}, T_{u_2}, T_{u_3}, and T_{u_5}. Instance $\{Lucca, Florence\}$ has a probability of re-identification $PR_D(d{=}u_1|\{Lucca, Florence\}){=}\frac{1}{3}$ because the pair appears in three trajectories T_{u_1}, T_{u_3}, and T_{u_5}. Analogously, we compute the probability of re-identification for the other three possible instances: $PR_D(d{=}u_1|\{Leghorn, Pisa\}){=}\frac{1}{4}$, $PR_D(d{=}u_1|\{Leghorn, Florence\}){=}\frac{1}{3}$, $PR_D(d{=}u_1|\{Pisa, Florence\}){=}\frac{1}{4}$;

(2) We compute the risk of re-identification of individual u_1 as the maximum of the probabilities of re-identification among all instances in B_2: $Risk(u_1){=}max(\frac{1}{4}, \frac{1}{4}, \frac{1}{3}, \frac{1}{4}, \frac{1}{3}, \frac{1}{4}) = \frac{1}{3}$.

We remark that the Data Provider does *not* know in advance the instance associated with the highest probability of re-identification of individual u_1 (i.e., the "best" combination of points from the perspective of the malicious adversary). The Data Provider can use the preceding computation in a preventive manner to identify the instance yielding the highest probability of re-identification, which is, for individual u_1, instance $\{Leghorn, Florence\}$. Due to the definition of risk, which depends on both an attacked individual's structure and the structures of all the other individuals in the dataset, identifying *a priori* an attack where the adversary has access to the best k-combination of points is difficult for the Data Provider. A particular case where the Data Provider can immediately recognize the best k-combination of points is a scenario where the adversary knows a location visited only by the individual under attack. Since the Data Provider has a view of the entire dataset, she can simulate such an attack by selecting the locations visited by just one individual (i.e., with number of visits equal to 1). In such a case, computing the privacy risk for the individuals visiting those locations does not require any combinatorial computation because the privacy risk is 1 for any value of k.

An individual is hence associated with several privacy risks, each for every background knowledge configuration of an attack. Every privacy risk of an individual can be computed using the following procedure (see also Section 1 in Supplementary Material):

Procedure 3.2: Privacy Risk Computation

(1) given an individual, define an attack based on a specific background knowledge category;

(2) consider a set of m background knowledge configurations $\{B_1, \ldots, B_m\}$;

(3) for every configuration $B_k \in \{B_1, \ldots, B_m\}$ compute all the possible instances $b \in B_k$ and the corresponding probability of re-identification;

(4) select the privacy risk of the individual for a configuration B_k as the maximum probability of re-identification across all the instances $b \in B_k$.

3.1 Computational Complexity of Privacy Risk Computation

The procedure of privacy risk computation has a high computational complexity. We assume that the adversary uses all the information available to her when conducting a re-identification attack on an individual. Since it is unlikely that an adversary knows the complete movements of an individual (i.e., all the points), we introduced the concept of background knowledge configuration B_k, which indicates the portion of points k known by the adversary when performing an attack on an individual. The higher the k, the higher is the number of points known by the adversary about the individual's movement. The maximum possible value of k is len, the length of the data structure of an individual.

The best k-combination of points is the one leading to the highest probability of re-identification of the individual under attack. However, we do not know such a best combination in advance. For this reason, given k, when we simulate an attack, we compute all the possible k-combinations of points an adversary could know. Given a combination of k points, we assume that the adversary uses *all* these k points to conduct the attack. This leads to a high overall computational complexity $O(\binom{len}{k} \times N)$, since the framework generates $\binom{len}{k}$ background knowledge configuration instances and, for each instance, it executes N matching operations by applying function *matching*.

In the extreme case where the adversary knows the complete movement of an individual (i.e., she knows all the points), we have $k = len$ and the computational complexity is $O(N)$. In general, in the range $k \in [1, \frac{len}{2}]$ of the computational complexity of the attack simulation increases with k, while for $k \in [\frac{len}{2}, n]$ the computational complexity decreases with k. While all the $\binom{len}{k}$ possible instances must be necessarily considered since, as already stated, we cannot exclude any of them a priori, we can reduce the number N of matching operations between a single instance and the data structures in the dataset by eliminating unnecessary comparisons. To clarify the point, consider an attack where we try to match an instance $b = (l_1, t_1), (l_2, t_2)$ against a trajectory T starting with the visit (l_3, t_3), with $t_1 < t_2 < t_3$. Since T is temporally ordered (see Definition 2.1), we can immediately exclude that b can be found in T. Although the overall worst-case complexity of the attack remains $\binom{len}{k}$, in practice, this optimization speeds up the execution by skipping unnecessary comparisons during the matching between an instance and a trajectory. However, as we will show in Section 5, in practice, the matching optimizations do not eliminate the computational problem, and the simulation of the attacks can take up to 2 weeks to compute the privacy risks of individuals in our datasets.

Example 3.6. Consider the following scenario where an adversary knows 5 locations of an individual with a trajectory length $len = 50$. Computing the privacy risk of an individual with respect to the background knowledge configuration B_5 requires the generation of the $\binom{50}{5} = 2{,}118{,}760$ background knowledge instances. In a dataset of $N = 100{,}000$ individuals, each with $len = 50$, the overall simulation of the attack would take around 210 billions of matching operations.

4 A DATA MINING APPROACH FOR PRIVACY RISK ASSESSMENT

Given its computational complexity, Procedure 3.2 (Privacy Risk Computation) becomes unfeasible as the size of the dataset increases since it requires enormous time and computational costs. This drawback is even more serious if we consider that the privacy risks must be necessarily recomputed every time the mobility dataset is updated with new data records and for every selection of individuals, geographic areas, and periods of time. To overcome these problems, we propose a fast and flexible data mining approach. The idea is to train a predictive model to predict the privacy risk of an individual based solely on her individual mobility patterns. The predictive model can be either a regression model, if we want to predict the actual value of privacy risk, or a classification model, if we want to predict the level of privacy risk. The training of the predictive model uses a training dataset where every example refers to a single individual and consists of *(i)* a vector of the individual's mobility features and *(ii)* the privacy risk value or the privacy risk level of the individual, depending on whether we perform a regression or a classification task, respectively. Formally, we define a regression training dataset as a tuple $\text{TR} = (F, R)$ where F is the set of the individual's mobility feature vectors and R is the vector of the individual's privacy risk. Similarly, we define a classification training dataset as a tuple $\text{TC} = (F, C)$ where C is the vector of the individual's privacy risk level (e.g., from level 1 indicating no risk to level 10 indicating maximum privacy risk). We define a possible set F of mobility features in Section 4.1, and we introduce a repertoire of attacks on mobility data that can be used to assess privacy risks in Section 4.2. We describe how to construct the regression training dataset and the classification training dataset in Section 4.3. In Section 4.4, we describe how a Data Provider can use our approach in practice to determine the privacy risk of individuals in her database. We make our approach parametric with respect to the predictive algorithm: In our experiments, we use a Random Forest regressor and a Random Forest classifier for the regression and classification experiments, respectively (Section 5), but every algorithm available in the literature can be used for the predictive tasks. Note that our approach is constrained to the fixed well-defined set of attacks introduced in Section 4.2, which is a representative set of nine sufficiently diverse attacks tailored for the data structures required to compute standard individual human mobility measures. Our approach can be easily extended to any type of attack defined on human mobility data by using the privacy framework proposed by Pratesi et al. (2016).

4.1 Individual Mobility Features

The mobility dynamics of an individual can be described by a set of measures widely used in the literature. Some measures describe specific aspects of an individual's mobility; other measures describe an individual's mobility in relation to collective mobility.

A subset of these measures can be simply obtained as aggregation of an individual's trajectory or frequency vector. The number of visits V of an individual is the length of her trajectory (i.e., the sum of all the visits she did in any location during the period of observation (Gonzalez et al. 2008; Pappalardo et al. 2015)). By dividing this quantity by the number of days in the period of observation, we obtain the average number of daily visits $\overline{V}$, which is a measure of the erratic behavior of an individual during the day (Pappalardo and Simini 2016). The length $Locs$ of the frequency vector of an individual indicates the number of distinct places visited by the individual during the period of observation (Gonzalez et al. 2008; Song et al. 2010a). Dividing $Locs$ by the number of available locations in the considered territory, we obtain $Locs_{ratio}$, which indicates the fraction of territory exploited by an individual in her mobility behavior. The maximum distance D_{max} traveled by an individual is defined as the length of the longest trip of the individual during the period of observation (Williams et al. 2015), while D_{max}^{trip} is defined as the ratio between D_{max} and the maximum possible distance between the locations in the area of observation. The sum of all the trip lengths traveled by the individual during the period of observation is defined as D_{sum} (Williams et al. 2015). It can be also averaged over the days in the period of observation, thus obtaining $\overline{D}_{sum}$.

In addition to these simple quantities, more complex measures can be computed based on an individual's mobility data, such as the radius of gyration (Gonzalez et al. 2008; Pappalardo et al. 2013) and the mobility entropy (Eagle and Pentland 2009; Song et al. 2010b). The radius of gyration r_g is the characteristic distance traveled by an individual during the period of observation, formally defined by Gonzalez et al. (2008) and Pappalardo et al. (2013, 2015) as:

$$r_g = \sqrt{\frac{1}{V} \sum_{i \in L} w_i (r_i - r_{cm})^2},$$

where w_i is the individual's visitation frequency of location i, V is the total number of visits of the individual, r_i is a bi-dimensional vector describing the geographical coordinates of location i, and $r_{cm} = \frac{1}{V} \sum_{i \in L} r_i$ is the center of mass of the individual (Gonzalez et al. 2008; Pappalardo et al. 2013). The mobility entropy E is a measure of the predictability of an individual's trajectory. Formally, it is defined as the Shannon entropy of an individual's movements (Eagle and Pentland 2009; Song et al. 2010b):

$$E = -\sum_{i \in L} p_i \log_2 p_i,$$

where p_i is the probability of location i in an individual's probability vector.

Also, for each individual, we keep track of the characteristics of three different locations: the most visited location, the second most visited location, and the least visited location. The frequency w_i of a location i is the number of times an individual visited location i during the period of observation, while the average frequency $\overline{w}_i$ is the daily average frequency of location i. We also define w_i^{pop} as the frequency of a location divided by the popularity of that location in the whole dataset. The quantity U_i^{ratio} is the number of distinct individuals who visited a location i divided by the total number $|U_{set}|$ of individuals in the dataset, while U_i is the number of distinct individuals

Table 1. The Individual Mobility Measures Used in Our Work

Symbol	Name	Structures	Attacks
V	visits	trajectory	Location Location Sequence Visit
$\overline{V}$	daily visits		
D_{max}	max distance		
D_{sum}	sum distances		
$\overline{D}_{sum}$	D_{sum} per day		
D_{max}^{trip}	D_{max} over area	trajectory location set	
$Locs$	distinct locations	frequency vector	Frequent Location Frequent Loc. Sequence
$Locs_{ratio}$	$Locs$ over area	frequency vector location set	
R_g	radius of gyration	probability vector	Probability
E	mobility entropy		
E_i	location entropy	probability vector probability vector dataset	
U_i	individuals per location	frequency vector, frequency vector dataset	Frequency Proportion Home and Work
U_i^{ratio}	U_i over individuals		
w_i	location frequency		
w_i^{pop}	w_i over overall frequency		
$\overline{w}_i$	daily location frequency		

For every mobility measure, we indicate the minimal data structures (among those presented in Section 2) needed to compute it and the attacks that can be performed on the corresponding data structures.

who visited location i during the period of observation. Finally, the location entropy E_i is the predictability of location i, defined as:

$$E_i = -\sum_{u \in U_i} p_u \log_2 p_u,$$

where p_u is the probability that individual u visits location i.

Table 1 indicates, for every mobility measure, the minimal data structures (among those presented in Section 2) required for its computation and the possible re-identification attacks that can be conducted on these structures. Every individual u in the dataset is described by a mobility vector $\overline{m}_u$ of the 16 mobility features described earlier. The vectors of all the mobility vectors of individual $u_1, \ldots, u_n$ is the mobility matrix $F = (m_{u_1}, \ldots, m_{u_n})$. It is worth noting that all the measures can be computed in linear time on the size of the corresponding data structure.

4.2 Privacy Attacks on Mobility Data

In this section, we describe the attacks we use in this article: the Proportion and Probability attacks are a novel contribution, while the others are attacks already existing in the literature. In Section 2 of the Supplementary Material, we provide the pseudocode to reproduce the attacks and some toy examples that illustrate how the attacks work.[2]

[2] We also provide the Python code we use for the simulation of the attacks at https://github.com/pellungrobe/privacy-lib.

4.2.1 Location Attack. 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. Since an individual might visit the same location multiple times in a trajectory, the adversary's knowledge is a multiset that may contain more occurrences of the same location. This is similar to considering the locations as items of transactions. Similar attacks on transactional databases are used in Terrovitis et al. (2008), Xu et al. (2008a), and Xu et al. (2008b) with the difference that a transaction is a set of items and not a multiset. 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 4.1 (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 $L(T_s)$.

Since each instance $b \in B_k$ is a subset of locations $X_s \subseteq L(T_s)$ of length k, given a record $d \in D$ and the corresponding individual u, we define the matching function as:

$$matching(d,b) = \begin{cases} true & b \subseteq L(T_u) \\ false & otherwise \end{cases} \tag{1}$$

4.2.2 Location Sequence Attack. In a Location Sequence attack, introduced in Mohammed et al. (2009) and Monreale et al. (2014a), the adversary knows a subset of the locations visited by the individual 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 defined as follows:

Definition 4.2 (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 $L(T_s)$.

We indicate with $a \preceq b$ that a is a subsequence of b. Each instance $b \in B_k$ is a subsequence of location $X_s \preceq L(T_s)$ of length k. Given a record $d \in D$ and the corresponding individual u, we define the matching function as:

$$matching(d,b) = \begin{cases} true & b \preceq L(T_u) \\ false & otherwise \end{cases} \tag{2}$$

4.2.3 Visit Attack. In a Visit attack, introduced in Abul et al. (2008b), Yarovoy et al. (2009), Monreale et al. (2010a), and de Montjoye et al. (2013), an adversary knows a subset of the locations visited by the individual and the time the individual visited these locations. The background knowledge category of a Visit attack is defined as:

Definition 4.3 (Visit based 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.

Each instance $b \in B_k$ is a spatio-temporal subsequence X_s of length k. The subsequence X_s has a positive match with a specific trajectory if the latter supports b in terms of both spatial and temporal dimensions. Thus, given a record $d \in D$, we define the matching function as:

$$matching(d,b) = \begin{cases} true & \forall (l_i, t_i) \in b, \exists (l_i^d, t_i^d) \in d \mid l_i = l_i^d \wedge t_i = t_i^d \\ false & otherwise \end{cases} \tag{3}$$

4.2.4 Frequent Location and Sequence Attack. We also introduce two attacks based on the knowledge of the location frequency. In the Frequent Location attack, the adversary knows a number of *frequent* locations visited by an individual, while in the Frequent Location Sequence attack, the adversary knows a subset of the locations visited by an individual and the relative ordering with respect to the frequencies (from most frequent to least frequent). The Frequent Location attack is similar to the Location attack with the difference that, in frequency vectors, a location can appear only once. As a consequence, this attack follows the same principle of Terrovitis et al. (2008) and Xu et al. (2008a, 2008b). The Frequent Location Sequence attack is similar to the Location Sequence attack, with two differences: first, a location can appear only once in the vector; second, locations in a frequency vector are ordered by descending frequency and not by time. Thus, the locations/sequence X_s of length k cannot contain repetitions of locations. We omit the definition of the matching functions because they are similar to those of the attacks conducted on trajectories: They must only consider the absence of location repetitions.

4.2.5 Frequency Attack. 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 to the locations. Thus, 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 with visit $v \in W$ the pair composed by the frequent location and its frequency. We also recall that we denote with W_s the frequency vector of individual s. The background knowledge category of a Frequency attack is defined as follows:

Definition 4.4 (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.

Each instance $b \in B_k$ is a frequency vector, and, given a record $d \in D$, we define the matching function as:

$$matching(d, b) = \begin{cases} true & \forall (l_i, w_i) \in b, \exists (l_i^d, w_i^d) \in W \mid l_i = l_i^d \land w_i \leq w_i^d \\ false & otherwise \end{cases} \tag{4}$$

4.2.6 Home And Work Attack. In the Home and Work attack introduced in Zang and Bolot (2011), the adversary knows the two most frequent locations of an individual and their frequencies. It essentially assumes the same background knowledge of the Frequency attack but related only to two locations. This is the only attack where the background knowledge configuration is composed of just a single 2-combination for each individual. Mechanically, the matching function for this type of attack is identical to the matching function of the Frequency attack.

4.2.7 Proportion Attack. We introduce an attack assuming that an adversary knows a subset of locations visited by an individual and also the relative proportion between the number of visits to these locations. In particular, the adversary knows the proportion between the frequency of the most frequent known location and the frequency of the other known locations. This means that the candidate set of possible matches consists of all the set of locations with similar proportions. Given a set of visits $X \subset W$, we denote with $l1$ the most frequent location of X and with w_1 its frequency. We also denote with pr_i the proportion between w_i and w_1 for each $v_i \neq v_1 \in X$. We then denote with 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 4.5 (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.

Each adversary's knowledge $b \in B_k$ is a LR structure, as previously defined. Given a record $d \in D$, we define the matching function as:

$$matching(d,b) = \begin{cases} true & \forall (l_i, pr_i) \in b, \exists (l_i^d, pr_i^d) \in LR^d \mid l_i = l_i^d \wedge pr_i \in [pr_i^d - \delta, pr_i^d + \delta] \\ false & otherwise \end{cases} \quad (5)$$

In the equation, δ is a tolerance factor for the matching of proportions. In our experiments, $\delta = 0.1$.

4.2.8 Probability Attack. In a Probability attack an adversary knows the locations visited by an individual and the probability of that individual to visit each location. This attack is similar to the one introduced by Unnikrishnan and Naini (2013), where the goal is to match m users with m public statistics, like empirical frequencies. However, there are some differences between the two attacks: The attack proposed in Unnikrishnan and Naini (2013) works on two sets of data, called *strings*. One of the sets represents the published aggregated data of individuals, the other represents the auxiliary information known by the adversary about the individuals in the data. The two sets are equal in size, and also all the strings in the two sets have the same length. Given these assumptions, Unnikrishnan and Naini (2013) propose an attack based on the minimum weight bipartite matching. Conversely, in our Probability attack, we try to match a single background knowledge instance with the set of probability vectors. Therefore, we cannot rely on matching algorithms on a bipartite graph because we can not make assumptions regarding the length of the sets or the length of the data: In general, 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 with visit $v \in P$ the pair composed of the frequent location and its probability. We also recall that we denote with P_s the probability vector of individual s. The background knowledge category for this attack is defined as follows:

Definition 4.6 (Probability background knowledge). Let k be the number of visits v of the probability vector of individual s known by the adversary. The Probability-based 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.

Each adversary's knowledge $b \in B_k$ is a probability vector, and, given a record $d \in D$, we define the matching function as:

$$matching(d,b) = \begin{cases} true & \forall (l_i, p_i) \in b, \exists (l_i^d, p_i^d) \in d \mid l_i = l_i^d \wedge p_i \in [p_i^d - \delta, p_i^d + \delta] \\ false & otherwise \end{cases} \quad (6)$$

In the equation, δ is a tolerance factor for the matching of probabilities. In our experiments, $\delta = 0.1$.

4.3 Construction of Training Dataset

Given an attack i based on a specific background knowledge configuration B_j^i, the regression training dataset TR_j^i and the classification training dataset TC_j^i can be constructed by the following three-step procedure:

(1) Given a mobility dataset D, for every individual u we compute the set of individual mobility features described in Section 4.1 based on her mobility data. Every individual u is

hence described by a mobility feature vector $\overline{m}_u$. All the individuals' mobility feature vectors compose mobility matrix $F=(\overline{m}_1, \ldots, \overline{m}_n)$, where n is the number of individuals in D;

(2) For every individual, we simulate the attack with background knowledge configuration B_j^i on D in order to compute a privacy risk value for every individual. We obtain a privacy risk vector $R_j^i = (r_1, \ldots, r_n)$. The regression training set is hence $\mathrm{TR}_j^i = (F, R_j^i)$;

(3) We transform the regression training set TR_j^i into a classification training set TC_j^i by discretizing vector R_j^i on the intervals $[0.0]$, $(0, 0.1]$, $(0.1, 0.2]$, $(0.2, 0.3]$, $(0.3, 0.5]$, $(0.5, 1.0]$. We obtain in this way a privacy risk level vector $C_j^i = (c_1, \ldots, c_n)$. The classification training set is hence $\mathrm{TC}_j^i = (F, C_j^i)$.

Every regression classification dataset TR_j^i or classification training dataset TC_j^i is used to train a predictive model M_j^i. The predictive model will be used by the Data Provider to immediately estimate the privacy risk value or the privacy risk level of *previously unseen* individuals whose data were not used in the learning process, with respect to attack i, background knowledge configuration B_j^i, and dataset D.

Example 4.7 (Construction of classification training set). Consider a mobility dataset of trajectories $D=\{T_{u_1}, T_{u_2}, T_{u_3}, T_{u_4}, T_{u_5}\}$ corresponding to five individuals u_1, u_2, u_3, u_4, and u_5. Given an attack i, a background knowledge configuration B_j^i, and dataset D, we construct the classification training set TC_j^i as follows:

(1) For every individual u_i, we compute the 16 individual mobility measures based on her trajectory T_{u_i}. Every individual u_i is hence described by a mobility feature vector of length 16 $\overline{m}_{u_i} = (m_1^{(u_i)}, \ldots, m_{16}^{(u_i)})$. All the mobility feature vectors compose mobility matrix $F=(\overline{m}_{u_1}, \overline{m}_{u_2}, \overline{m}_{u_3}, \overline{m}_{u_4}, \overline{m}_{u_5})$;

(2) We simulate the attack with configuration B_j^i on dataset D and obtain a vector of five privacy risk values $R_j^i = (r_{u_1}, r_{u_2}, r_{u_3}, r_{u_4}, r_{u_5})$, each for every individual;

(3) Suppose that the actual privacy risks resulting from simulation are $R_j^i=(1.0, 0.5, 1.0, 0.25, 0.03)$. We discretize the values of the privacy risk vector R_j^i on the intervals $[0.0]$, $[0, 0.1]$, $(0.1, 0.2]$, $(0.2, 0.3]$, $(0.3, 0.5]$, $(0.5, 1.0]$. We hence obtain a privacy risk level vector $C_j^i = ((0.5, 1.0], (0.3, 0.5], (0.5, 1.0], (0.2, 0.3], [0, 0.1])$ and the classification training dataset $\mathrm{TC}_j^i = (F, C_j^i)$.

4.4 Usage of the Data Mining Approach

The Data Provider can use a classifier M_j^i to determine the level of privacy risk with respect to an attack i and background knowledge configuration B_j^i for: *(i) previously unseen* individuals whose data were *not* used in the learning process or *(ii)* a selection of individuals in the database already used in the learning process. It is worth noting that, with existing methods, the privacy risk of individuals in scenario *(ii)* must be recomputed by simulating attack (i) from scratch. In contrast, the usage of classifier M_j^i allows us to obtain the privacy risk of the selected individuals immediately. The computation of the mobility measures and the classification of privacy risk level can be done in polynomial time as a one-off procedure.

To clarify this point, consider the following scenario. A Data Analyst requests the Data Provider for updated mobility data about a new set of individuals with the purpose of studying their characteristic traveled distance (radius of gyration r_g) and the predictability of their movements (mobility entropy E). Since both measures can be computed by using a probability vector (see Table 1), the

Data Provider can release just the probability vectors of the individuals requested. Before that, however, the Data Provider wants to determine the level of privacy risk to the individuals with respect to the Probability attack (P) and several background knowledge configurations B_j^P. The Data Provider uses classifier M_j^P previously trained to obtain the privacy risk level of the individuals. On the basis of privacy risks obtained from M_j^P, the Data Provider can immediately identify risky individuals (i.e., individuals with a high level of privacy risk). She then can decide to either filter out the risky individuals or to select suitable privacy-preserving techniques (e.g., k-anonimity or differential privacy) and transform their mobility data in such a way that their privacy is preserved. In the next section, we present an extensive evaluation of our methodology on real-world mobility data and show the effectiveness of the proposed data mining approach.

5 EXPERIMENTS

For all the attacks defined except the Home and Work attack, we consider four background knowledge configurations B_k with $k = 2, 3, 4, 5$, where configuration B_k corresponds to an attack where the adversary knows k locations visited by the individual. For the Home and Work attack, we have just one possible background knowledge configuration, where the adversary knows the most frequent location and the second most frequent location of an individual.

We use a dataset provided by Octo Telematics[3] storing the GPS tracks of private vehicles traveling in two Italian urban areas, Florence and Pisa, from May 1, 2011, to May 31, 2011. In particular, we have 9,715 private vehicles in the Florence dataset and 2,280 vehicles in the Pisa dataset. The GPS device embedded in a vehicle automatically turns on when the vehicle starts, and the sequence of GPS points that the device produces every 30 seconds forms the global GPS track of a vehicle. When the vehicle stops, no points are logged or sent. We exploit these stops to split the global GPS track of a vehicle into several subtracks, corresponding to the trips performed by the vehicle. To ignore small stops like traffic lights and gas stations, we follow the strategy commonly used in the literature (Pappalardo et al. 2013, 2015) and choose a stop duration threshold of at least 20 minutes: If the time interval between two consecutive GPS points of the vehicle is larger than 20 minutes, the first point is considered as the end of a trip and the second one as the start of another trip.[4] We assign each origin and destination point of the obtained subtracks to the corresponding census cell according to the information provided by the Italian National Statistics Bureau (ISTAT) in order to assign every origin and destination point to a location (Pappalardo et al. 2015). This allows us to describe the mobility of every vehicle in the Florence or the Pisa datasets in terms of a trajectory, in compliance with the definition introduced in Section 2. Since our purpose is to provide a tool to immediately discriminate between individuals with low risk and individuals with high risk, in this section, we show the results of classification experiments. We also perform regression experiments where we predict the exact value of privacy risk and show the corresponding results in Section 3.4 of the Supplementary Material.

We construct a classification training dataset TC_j^i for every distinct background knowledge configuration B_j^i of the attacks described in Section 4.2. This means that, in our experiments, we build a total of 33 distinct classification training datasets for 33 distinct classification experiments. This is because we consider four background knowledge configurations (k=2, 3, 4, 5) for eight attacks (Visit, Frequency, Location, Frequent Location Sequence, Frequent Location, Probability, Proportion, Sequence), and just one background knowledge configuration for the Home and Work attack. So we construct a total of $(8 \times 4) + 1 = 33$ distinct classification training datasets.

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

[4] We also performed the extraction of the trips using different stop duration thresholds (5, 10, 15, 20, 30, 40 minutes), without finding significant differences in the sample of short trips and in the statistical analysis we present in this article.

Every classification dataset TC_j^i is used to train a classifier M_j^i using Random Forest (Hastie et al. 2009).[5] We evaluate the overall performance of a classifier by two metrics (Tan et al. 2005): *(i)* the accuracy of classification $\text{ACC} = \frac{|\hat{f}(x_i)=f(x_i)|}{n}$, where $f(x_i)$ is the actual label of individual i, $\hat{f}(x_i)$ is the predicted label, and n is the number of individuals in the training dataset; and *(ii)* the weighted average F-measure, defined as $F = \sum_{c \in C} |c| \frac{2\text{TP}}{2\text{TP}+\text{FP}+\text{FN}}$, where TP, FP, FN stand for the numbers of true positives, false positives, and false negatives resulting from classification; C is the set of labels; and $|c|$ is the support of a label. All the experiments are performed using a k-fold cross validation procedure with k=10. We also perform a holdout-validation finding similar results in terms of accuracy and the F-measure with respect to the cross-validation method (see Supplementary Material, Section 3.3).[6]

Table 2 (columns Florence and Pisa) summarizes the results of the 33 classification tasks for both the Florence and Pisa datasets. We compare the performance of a classifier M_j^i with the performance of a baseline classifier which generates predictions by respecting the distribution of privacy risk labels in C_j^i.[5] In Table 2, we observe a significant gain in both accuracy and F-measure of the classifiers over the baseline. For example, in predicting the Probability privacy risk levels, the classifier reaches maximum performance values of ACC = 0.95 and F-measure = 0.95 (configuration k=4, Florence), a significant improvement with respect to the baseline model with ACC = 0.56 and F = 0.56. The Home and Work variable has the weakest relation with the individual mobility features, reaching the lowest performance values: ACC = 0.62 and F = 0.59 (where the baseline has ACC = 0.37 and F = 0.37). Finally, the classification results for Florence and Pisa are comparable, with slightly better performances for the Florence dataset (see Table 2). It is worth noting that, for some attacks such as the Visit attack, we have very similar performances in terms of both accuracy and F-measure for any k. This is due to the fact that the privacy risk distributions resulting from simulating the attack are similar for any $k \geq 2$ (Figure 1(c) and 4(c) in the Supplementary Material). In contrast, for the Location Sequence attack, we observe that the distribution of privacy risk for k=2 differs from the distributions of privacy risk for $k \geq 3$ (Figures 1(b) and 4(b) in the Supplementary Material). In our classification results, this results in a difference between k=2 and $k \geq 3$: The classification performances become stable for $k \geq 3$. Since the classifiers are accurate especially for the class of maximum risk $(0.5, 1]$, and since for $k \geq 3$ the number of individuals with maximum privacy risk increases, as a consequence, the performance of classifiers improve.

It is important to highlight that classifying a high-risk individual as a low-risk individual can be a major issue. For our application, the *recall* is important to evaluate the performance of a classifier: A high recall on the highest risk class $(0.5, 1.0]$ indicates that a very low number of high-risk individuals are misclassified as low-risk individuals. To be usable in practice, classifiers need to have a high recall on the highest risk class. Figure 1(a)-(b) show a matrix representing the classification error for every label of background knowledge configuration $k = 4$ of the Probability attack, for Florence (a) and Pisa (b). An element i, j in the matrix indicates the fraction of instances for which the actual label j is classified as label i by the classifier. The diagonal of the matrix, hence, indicates the classifier's recall for every label. We observe that the recall of the highest risk class $(0.5, 1.0]$ is 99% for Florence and 98% for Pisa. In particular, we observe that all the misclassifications of the classifiers for the highest risk class are made predicting class $(0.3, 0.5]$ (i.e., the second highest class of risk). So there is a zero probability of misclassifying high-risk individuals

[5] We use the implementation provided by the scikit-learn package in Python (Pedregosa et al. 2011).

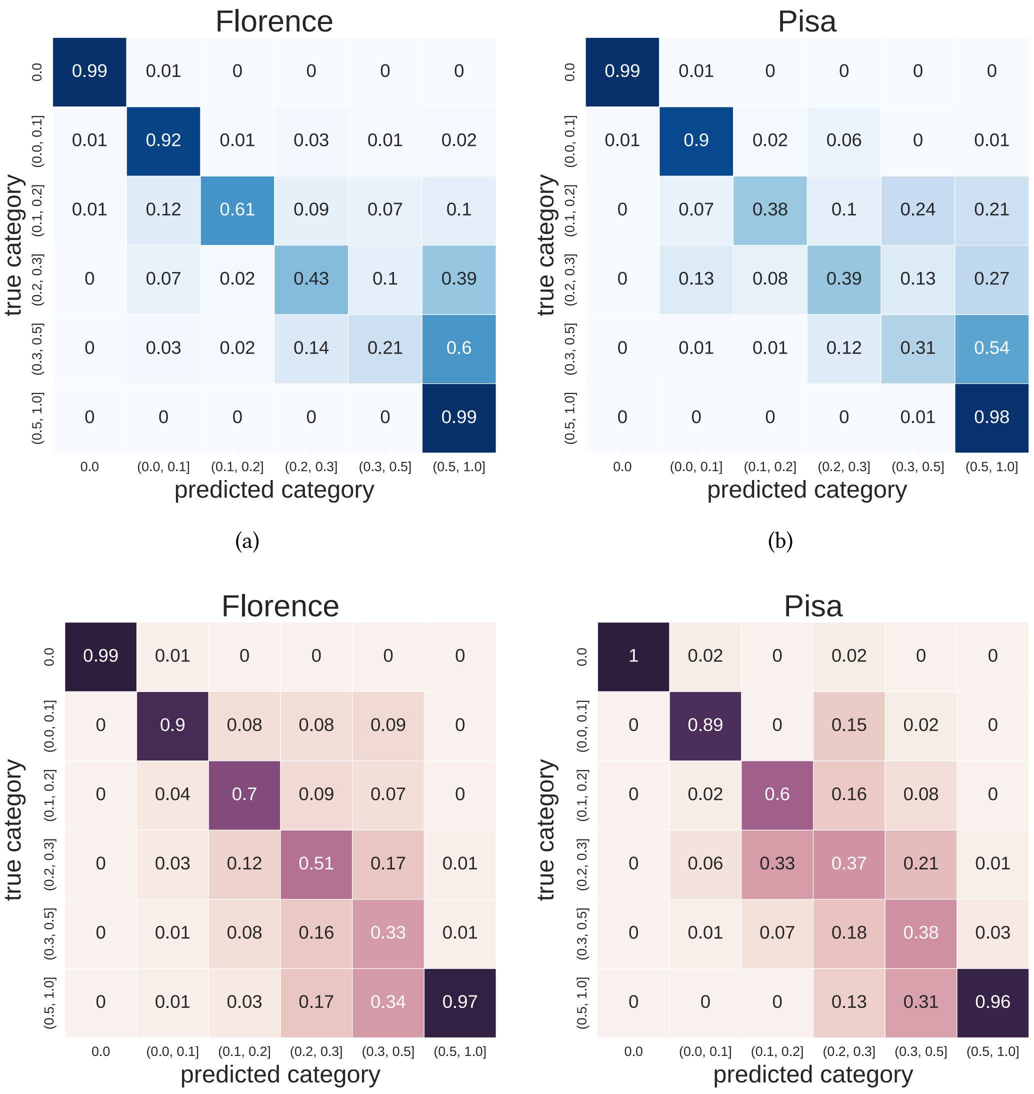
[6] The Python code for attacks simulation and classification tasks is available at https://github.com/pellungrobe/privacy-mobility-lib.

Table 2. Results of the 33 Classification Experiments for the Florence and the Pisa Datasets

	configuration		Florence		Pisa		FI → PI		PI → FI	
			ACC	F	ACC	F	ACC	F	ACC	F
Visit	locations with timestamps	$k = 2$	0.94	0.94	0.93	0.93	0.93	0.92	0.93	0.93
		$k = 3$	0.94	0.94	0.93	0.93	0.93	0.93	0.93	0.93
		$k = 4$	0.94	0.94	0.93	0.93	0.93	0.93	0.92	0.92
		$k = 5$	0.94	0.94	0.92	0.92	0.93	0.93	0.91	0.92
	avg baseline		0.82	0.81	0.81	0.80				
Frequency	locations with frequencies	$k = 2$	0.90	0.89	0.83	0.82	0.79	0.79	0.76	0.70
		$k = 3$	0.94	0.93	0.89	0.89	0.84	0.86	0.83	0.79
		$k = 4$	0.92	0.93	0.89	0.89	0.85	0.86	0.85	0.85
		$k = 5$	0.93	0.93	0.89	0.89	0.71	0.73	0.85	0.82
	avg baseline		0.53	0.53	0.41	0.41				
HW	two most frequent locations		0.62	0.59	0.57	0.54	0.57	0.55	0.51	0.49
	avg baseline		0.37	0.37	0.28	0.29				
Location	locations without sequence	$k = 2$	0.93	0.92	0.86	0.86	0.87	0.87	0.85	0.81
		$k = 3$	0.95	0.95	0.91	0.91	0.87	0.87	0.87	0.82
		$k = 4$	0.95	0.95	0.91	0.91	0.89	0.89	0.89	0.86
		$k = 5$	0.95	0.95	0.91	0.91	0.89	0.90	0.87	0.85
	avg baseline		0.57	0.56	0.44	0.44				
Freq.Loc. Sequence	locations with sequence	$k = 2$	0.93	0.92	0.88	0.87	0.88	0.87	0.86	0.83
		$k = 3$	0.94	0.94	0.88	0.89	0.90	0.89	0.73	0.66
		$k = 4$	0.94	0.94	0.89	0.89	0.85	0.87	0.86	0.82
		$k = 5$	0.93	0.94	0.89	0.89	0.90	0.90	0.86	0.83
	avg baseline		0.58	0.57	0.46	0.45				
Frequent Location	locations without sequence	$k = 2$	0.81	0.79	0.71	0.69	0.73	0.74	0.65	0.62
		$k = 3$	0.86	0.85	0.8	0.78	0.81	0.81	0.75	0.72
		$k = 4$	0.87	0.86	0.81	0.79	0.83	0.83	0.79	0.75
		$k = 5$	0.87	0.87	0.81	0.8	0.82	0.83	0.78	0.75
	avg baseline		0.65	0.65	0.56	0.55				
Probability	locations with probability	$k = 2$	0.93	0.92	0.86	0.86	0.86	0.85	0.82	0.80
		$k = 3$	0.95	0.95	0.92	0.92	0.89	0.89	0.86	0.83
		$k = 4$	0.95	0.95	0.91	0.91	0.91	0.90	0.85	0.81
		$k = 5$	0.95	0.95	0.92	0.92	0.92	0.92	0.87	0.83
	avg baseline		0.56	0.56	0.45	0.44				
Proportion	locations with proportion	$k = 2$	0.90	0.89	0.83	0.81	0.79	0.79	0.79	0.76
		$k = 3$	0.94	0.93	0.89	0.89	0.89	0.89	0.83	0.78
		$k = 4$	0.93	0.93	0.89	0.89	0.85	0.86	0.84	0.81
		$k = 5$	0.93	0.93	0.89	0.89	0.83	0.84	0.83	0.77
	avg baseline		0.54	0.54	0.42	0.40				
Location Sequence	locations with sequence	$k = 2$	0.88	0.86	0.79	0.77	0.83	0.82	0.78	0.74
		$k = 3$	0.92	0.92	0.87	0.86	0.88	0.88	0.86	0.83
		$k = 4$	0.92	0.92	0.88	0.87	0.88	0.88	0.87	0.85
		$k = 5$	0.93	0.93	0.88	0.87	0.91	0.90	0.87	0.84
	avg baseline		0.64	0.64	0.55	0.54				

The classification performance is evaluated by the overall accuracy (ACC) and the weighted F-measure (F) by using a k-fold cross validation with k=10. In columns FI → PI and PI → FI, where FI indicates Florence and PI indicates Pisa, we show the results of classification where we train the classifiers on the first urban area and try to predict the privacy risks of individuals in the second urban area.

as low-risk individuals (i.e., classes [0.0] and (0.0, 0.1]). Similarly, in Figure 1(c)-(d), an element i, j in the matrix indicates the fraction of instances for which the predicted label j is actually label i in the dataset. The diagonal matrix indicates in this case the classifier's precision for every label. We observe that the classifier is very precise for the two lowest (risk $\in$ [0.0] and risk $\in$ (0.0, 0.1]) and the highest (risk $\in$ (0.5, 1.0]) privacy risk labels: Both the recall and the precision of these labels



Florence

	0.0	(0.0, 0.1]	(0.1, 0.2]	(0.2, 0.3]	(0.3, 0.5]	(0.5, 1.0]
0.0	0.99	0.01	0	0	0	0
(0.0, 0.1]	0.01	0.92	0.01	0.03	0.01	0.02
(0.1, 0.2]	0.01	0.12	0.61	0.09	0.07	0.1
(0.2, 0.3]	0	0.07	0.02	0.43	0.1	0.39
(0.3, 0.5]	0	0.03	0.02	0.14	0.21	0.6
(0.5, 1.0]	0	0	0	0	0	0.99

(a)

Pisa

	0.0	(0.0, 0.1]	(0.1, 0.2]	(0.2, 0.3]	(0.3, 0.5]	(0.5, 1.0]
0.0	0.99	0.01	0	0	0	0
(0.0, 0.1]	0.01	0.9	0.02	0.06	0	0.01
(0.1, 0.2]	0	0.07	0.38	0.1	0.24	0.21
(0.2, 0.3]	0	0.13	0.08	0.39	0.13	0.27
(0.3, 0.5]	0	0.01	0.01	0.12	0.31	0.54
(0.5, 1.0]	0	0	0	0	0.01	0.98

(b)

Florence

	0.0	(0.0, 0.1]	(0.1, 0.2]	(0.2, 0.3]	(0.3, 0.5]	(0.5, 1.0]
0.0	0.99	0.01	0	0	0	0
(0.0, 0.1]	0	0.9	0.08	0.08	0.09	0
(0.1, 0.2]	0	0.04	0.7	0.09	0.07	0
(0.2, 0.3]	0	0.03	0.12	0.51	0.17	0.01
(0.3, 0.5]	0	0.01	0.08	0.16	0.33	0.01
(0.5, 1.0]	0	0.01	0.03	0.17	0.34	0.97

(c)

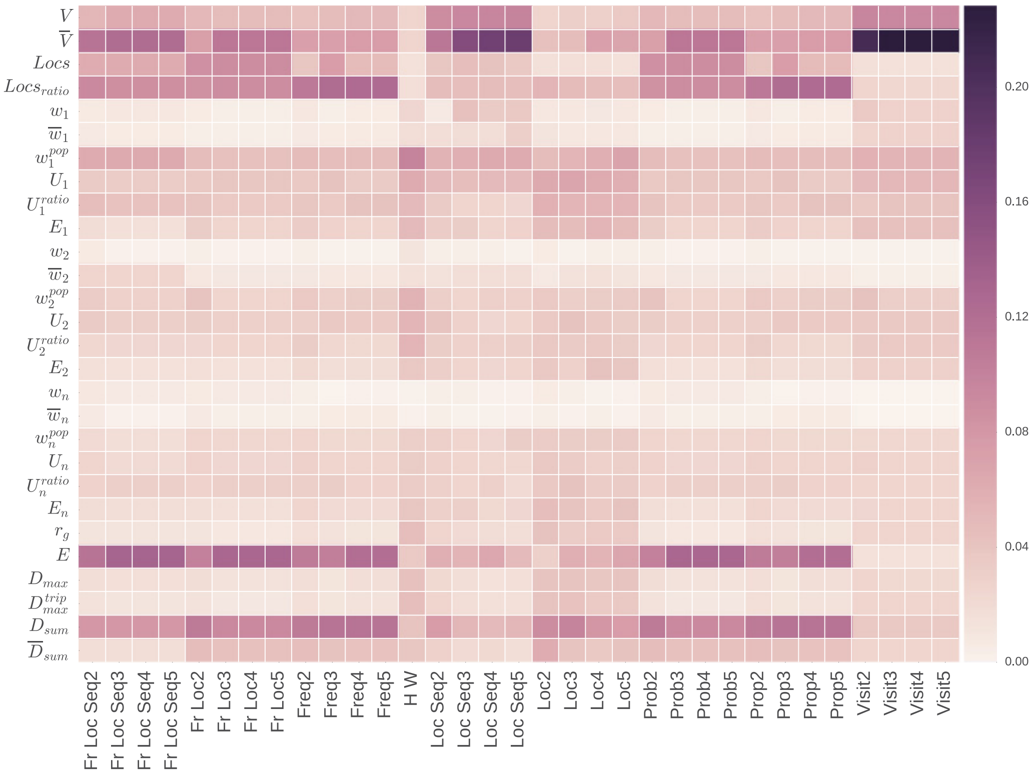
Pisa

	0.0	(0.0, 0.1]	(0.1, 0.2]	(0.2, 0.3]	(0.3, 0.5]	(0.5, 1.0]
0.0	1	0.02	0	0.02	0	0
(0.0, 0.1]	0	0.89	0	0.15	0.02	0
(0.1, 0.2]	0	0.02	0.6	0.16	0.08	0
(0.2, 0.3]	0	0.06	0.33	0.37	0.21	0.01
(0.3, 0.5]	0	0.01	0.07	0.18	0.38	0.03
(0.5, 1.0]	0	0	0	0.13	0.31	0.96

(d)

Fig. 1. Classification error per class for classifier M_4^P Probability attack P and background knowledge configuration B_4^P, for Florence (a, c) and Pisa (b, d). An element i, j in the matrices (a) and (b) indicates the fraction of instances for which the actual class j is classified as class i. The diagonal of the matrices (a) and (b), hence, indicates the classifier's recall for every class. An element i, j in the matrices (c) and (d) indicates the fraction of instances for which the predicted class j is actually class i in the dataset. The diagonal of matrices (c) and (d) indicates in this case the classifier's precision for every class. We observe that the classifier has both high recall and high precision on the first two classes (low risk) and the last class (maximum risk). We provide the matrices for all the other classifiers in Section 3.6 of the Supplementary Material.

are close to 1. Even on the labels where recall and precision are lower (i.e., $(0.1, 0.2]$, $(0.2, 0.3]$, $(0.3, 0.5]$), the classifier is more prone to predict a higher level of risk than a lower level of risk. These conservative choices allow the Data Provider to limit the privacy violation of individuals: It is hence unlikely that a classifier assigns to an individual a privacy risk label that is lower than



(a)

Fig. 2. The distribution of average importance of the mobility features for all 33 classifiers (Florence dataset).

her actual privacy risk label. We report in Section 3.6 of the Supplementary Material the matrices corresponding to the classification results of all the other background knowledge configurations.

In Table 2 (columns FI → PI and PI → FI), we also show the results of other classification experiments where we train a classifier on the Florence dataset and use it to classify the privacy risk label of vehicles in the Pisa dataset, and vice versa. Even if the two datasets cover disjoint sets of vehicles, we observe good predictive performance, comparable to the performance of classifiers where the training set and the test set belong to the same original dataset.

Importance of Mobility Features. We quantify the importance of every mobility feature in a classifier M_j^i by taking its average importance in the decision trees of the resulting random forest. The importance of a feature in a decision tree is computed as the (normalized) total reduction of classification entropy brought by that feature in the tree (Hastie et al. 2009). Figure 2 shows a heatmap representing the average importance of every mobility feature to the 33 classifiers in Florence, where every column corresponds to a classifier and every row corresponds to a mobility feature. We report the same heatmap for Pisa in Section 3.5 of the Supplementary Material. We observe the following results. First, while classifiers corresponding to different configurations of the attack show similar distributions of importances, classifiers corresponding to configurations of different attacks produce different distributions. For example, in the classifiers corresponding to the four configurations of the Visit attack, the average number of visits $\overline{V}$ is, not surprisingly, the most important mobility feature (Figure 2). In contrast, in the classifiers corresponding to the

Table 3. The Average Importance of Every Mobility Feature Computed Over All 33 Classifiers for Florence and Pisa

	Florence		Pisa			Florence		Pisa	
	measure	**impo.**	**measure**	**impo.**		**measure**	**impo.**	**measure**	**impo.**
1	$\overline{V}$	3.66	$Locs_{ratio}$	3.24	15	U_2^{ratio}	0.96	U_2^{ratio}	0.92
2	E	2.92	D_{sum}	3.22	16	U_n	0.88	U_n	0.88
3	D_{sum}	2.75	$\overline{V}$	2.87	17	w_n^{pop}	0.83	r_g	0.87
4	$Locs_{ratio}$	2.51	E	2.62	18	E_n	0.79	E_n	0.79
5	V	1.91	V	1.69	19	E_2	0.74	E_2	0.75
6	w_1^{pop}	1.77	$Locs$	1.66	20	D_{max}	0.68	w_n^{pop}	0.73
7	$Locs$	1.67	w_1^{pop}	1.62	21	D_{max}^{trip}	0.63	D_{max}^{trip}	0.67
8	U_1	1.44	U_1	1.46	22	r_g	0.61	D_{max}	0.58
9	U_1^{ratio}	1.32	U_1^{ratio}	1.40	23	w_1	0.42	$\overline{w}_1$	0.48
10	$\overline{D}_{sum}$	1.19	U_2	1.16	24	$\overline{w}_2$	0.40	w_1	0.44
11	U_2	1.12	U_n^{ratio}	1.09	25	$\overline{w}_1$	0.36	$\overline{w}_2$	0.36
12	w_2^{pop}	1.07	w_2^{pop}	1.07	26	w_n	0.13	w_n	0.15
13	E_1	1.05	E_1	1.06	27	$\overline{w}_n$	0.12	w_2	0.13
14	U_n^{ratio}	0.99	$\overline{D}_{sum}$	0.98	28	w_2	0.10	$\overline{w}_n$	0.13

We observe a correlation $r = 0.96$ between the importance of the mobility features in Florence and Pisa.

Table 4. Comparison of Execution Times of Attack Simulations and Classification Tasks on Florence and Pisa

Variable ($\sum_2^5 k$)	**Florence**		**Pisa**	
	simulation	**classifier**	**simulation**	**classifier**
Home and Work	149s (2.5m)	7s	5s	3s
Frequency	645s (10m)	22s	20s	10s
Frequent Location Sequence	846s (14m)	22s	23s	10s
Proportion	900s (15m)	24s	30s	10s
Frequent Location	997s (10m)	22s	30s	10s
Probability	1,165s (20m)	22s	37s	10s
Visit	2,274s (38m)	16s	95s (1.5m)	9s
LocationSequence	>168h (1week)	22s	>168h (1week)	10s
Location	>168h (1week)	22s	>168h (1week)	10s
total	**>2weeks**	**172s**	**>2weeks**	**79s**

four configurations of the Proportion attack, $\overline{V}$ has a low importance while D_{sum}, E, and $Locs_{ratio}$ have the highest importance. A second result is that the distribution of the average importances for Florence and Pisa are similar: We observe a Pearson correlation $r = 0.96$ between the two importances of the same variables in the two urban areas. Table 3 shows a ranking of the average importance the mobility features have in the classifiers, for Florence and Pisa. Here, we observe that individual measures (e.g., E, V, $\overline{V}$) tend to be the most important ones, while location-based features (e.g., W_i, E_i) tend to be less important.

Execution Times. We show the computational improvement of our approach in terms of execution time by comparing in Table 4 the execution times of the attack simulations and the execution times of the classification tasks.[7] The execution time of a single classification task is the sum of three subtasks: *(i)* the execution time of training the classifier on the training set, *(ii)* the execution time of using the trained classifier to predict the classes on the test set, and *(iii)* the execution time of evaluating the performance of classification (i.e., computing accuracy and F-measure). Table 4 shows that the execution time of attack simulations is low for the Frequency, Frequent Location Sequence, Proportion, Frequent Location, Probability, and Visit attacks (a few seconds for Pisa and a few minutes for Florence). However, for Location Sequence and Location, the execution times are huge: more than 1 week each. In contrast, the classification tasks have constant execution times of around 10s for Pisa and 22s for Florence. In summary, our approach can compute the risk levels for all 33 attacks in both Florence and Pisa in 250 seconds (less than 5 minutes), while the attack simulations require more than 2 weeks of computation.

6 DISCUSSION

The implementation of our data mining approach on real mobility data produces three remarkable results. First, the classifiers provide precise estimates of individuals' privacy risk, especially for the lowest privacy risk level and the highest privacy risk levels (Table 2). Moreover, the classifiers built on a given dataset (e.g., Florence) can be effectively used to estimate the privacy risks in a *different* dataset (e.g., Pisa; Table 2). These outcomes suggest that the classifiers can be a valid and fast alternative to existing privacy risk assessment tools. Instead of recomputing all the privacy risks when new data records become available and for every selection of individuals and geographic areas or periods of time, which would result in high computational costs, a Data Provider can effectively use the classifiers to obtain immediate and reliable estimates for every individual.

Second, different types of attacks generate different distributions of importance of the mobility measures in the classifiers (Figure 2). In particular, while some mobility measures are irrelevant for determining the privacy risk of an individual regardless the type of the risk (e.g., w_n and $\overline{w_n}$), other mobility measures are very relevant to determine the privacy risk of an individual (e.g., $\overline{V}$ and E). In other words, while some mobility measures provide a high predictive power, others are irrelevant and cannot be used alone to determine the privacy risk level of an individual. This suggests that both the learning phase and the predictive task should be done by computing the extensive set of mobility measures by using the maximal data structure (trajectory), even when a more aggregated data structure (e.g., a frequency vector) is sufficient for the Data Analysts' needs. However, this is not a problem in terms of computational costs because all the measures can be computed in linear time of the size of the dataset. It is worth noting that our approach can easily deal with changes in the long-term mobility patterns of an individual due, for example, to migration or changes in home/workplace. Every time new mobility data for an individual become available, the Data Provider can recompute her mobility features. To take into account long-term changes in mobility patterns, the recomputation of mobility measures can be done at regular time intervals (e.g., every month) by considering a time window with the most recent data (e.g., the last 6 months of data). The regular updates and the time window allow the Data Provider to consider the recent mobility history of an individual and obtain up-to-date individual mobility patterns.

A third remarkable result is that on both datasets the mobility measures describing aspects related to the individual alone, such as the number of visits V in the individual's trajectory and the mobility entropy E, are the most important features with which to classify the privacy risk of individuals (Table 3), far larger than the location-based measures (e.g., w_1, w_2, w_n) and the ones comparing individual mobility to collective mobility patterns (e.g., D_{max}^{trip}, w_m^{pop}). This result

[7] For a given type of attack, we report the sum of the execution times of the attacks for configurations $k = 2, 3, 4, 5$. We perform the experiments on Ubuntu 16.04.1 LTS 64 bit, 32GB RAM, 3.30GHz Intel Core i7.

is important because, in contrast with existing privacy risk assessment frameworks, it allows for estimating the privacy risk of an individual based on a limited amount of information about the collectivity. Since every individual can obtain an estimate of her own privacy risk based just on her mobility data, this increases awareness about personal data and helps her in deciding whether or not to share mobility data with third parties. This is compliant with a user-centric ecosystem (Forum 2013) like the one implemented by the personal data store (de Montjoye et al. 2012), where each individual has the full control of her personal data life-cycle. For this reason, our data mining approach can be integrated into the personal data store as a further tool available to the data owner.

7 RELATED WORKS

This article focuses on the mobility data of individuals traveling by car. An overview on the problems, techniques, and methodologies related to urban mobility data and urban computing can be found in Zheng et al. (2014). Human mobility data contain personal, sensitive information and can reveal many facets of the private life of individuals, leading to the possibility of a serious privacy violation. Nevertheless, in the past years, many techniques for privacy-preserving analysis on human mobility data have been proposed in the literature (Giannotti et al. 2013) showing that it is possible to design analytical mobility services where the quality of results coexists with the protection of personal data. A widely used privacy-preserving model is k-anonymity (Samarati and Sweeney 1998a, 1998b), which requires that an individual should not be identifiable from a group of size smaller than k based on their *quasi-identifiers* (QIDs), a set of attributes that can be used to uniquely identify individuals. Abul et al. (2008b) propose the (k, δ)-anonymity model, which takes advantage of the inherent uncertainty of the moving object's whereabouts, where δ represents the location precision. Assuming that different adversaries own disjoint parts of an individual's trajectory, Terrovitis and Mamoulis (2008) reduce privacy risk by relying on the suppression of the dangerous observations from each individual's trajectory. Yarovoy et al. (2009) propose the attack-graphs method to defend against attacks, based on k-anonymity. Monreale et al. (2010b) illustrate a generalized approach to achieve k-anonymity.

Other works are based on the so-called differential privacy model (Dwork et al. 2006). Monreale et al. (2013), for example, consider a privacy-preserving distributed aggregation framework for movement data, proposing the application of a ϵ-differential privacy model. Cormode et al. (2012) propose to publish a contingency table of trajectory data, where each cell in the table contains the number of individuals commuting from the given source location to the given destination location. Sebastien Gambs (2014) proposes a mobility model called Mobility Markov Chain, built upon mobility traces to re-identify an individual, while Ji et al. (2014) defines several similarity metrics which can be combined in a unified framework to provide de-anonymization of mobility and social network data.

One of the most important works on privacy risk assessment is the Linddun methodology (Deng et al. 2011), a privacy-aware threat analysis framework based on Microsoft's Stride methodology (Swiderski and Snyder 2004), useful for modeling privacy threats in software-based systems. In the past years, different techniques for risk management have been proposed, such as the OWASP's Risk Rating Methodology (OWASP 2016), NIST's Special Publication 800-30 (Stoneburner et al. 2002), SEI's OCTAVE (Alberts et al. 1999), and Microsoft's DREAD (Meier and Corporation 2003). Unfortunately, many of these works do not consider privacy risk assessment and simply include privacy considerations when assessing the impact of threats. Trabelsi et al. (2009) elaborate an entropy-based method to evaluate the disclosure risk of personal data, trying to manage quantitatively privacy risks. The *unicity* measure proposed in Song et al. (2014) and Achara et al. (2015) evaluates privacy risk as the number of records/trajectories which are uniquely identified.

Basu et al. (2014) propose an empirical risk model for the estimation of privacy risk for trajectory data and a framework to improve privacy risk estimation for mobility data, evaluating their model using k-anonymized data. Armando et al. (2015) propose a risk-aware framework for information disclosure which supports runtime risk assessment. In this framework, access-control decisions are based on the disclosure-risk associated with a data access request, and adaptive anonymization is used as a risk-mitigation method. Unfortunately, this framework only works on relational datasets since it needs to discriminate between QIDs and sensitive attributes.

Other works in the literature study the re-identification risk as a privacy measure in the context of network and social media data (Narayanan and Shmatikov 2009; Ramachandran et al. 2014) or combine network data and mobile phone data to re-identify people (Cecaj et al. 2016). The combination of multiple data sources for the attack and considering network data instead of mobility data in our methodology is one of the most interesting extensions that we intend to investigate as future work.

In this article, we use the privacy risk assessment framework introduced by Pratesi et al. (2016) (Section 3) to calculate the privacy risks of each individual in a mobility dataset. Our novel contribution is to overcome the inherent computational complexity of this framework by proposing a data mining approach that uses data mining classifiers to predict the privacy risk of an individual based solely on her mobility patterns.

8 CONCLUSION

Human mobility data are a precious proxy to improve our understanding of human dynamics, as well as to improve urban planning, transportation engineering, and epidemic modeling. Nevertheless, human mobility data contain sensitive information which, if analyzed with malicious intent, can lead to a serious violation of the privacy of the individuals involved. In this article, we proposed a fast and flexible data mining approach for estimating the privacy risk in human mobility data, one that overcomes the computational issues of existing privacy risk assessment frameworks. We validated our approach with an extensive experimentation on real-world GPS data, showing that we can achieve accurate estimations of privacy risks. In particular, the results showed that *(i)* the classifiers are accurate, especially on the highest and the lowest privacy risk classes; and *(ii)* the classifiers have a conservative behavior (i.e., misclassified individuals are assigned more likely to classes of higher risk than to classes of lower risk with respect to the actual class of privacy risk). Moreover, we observed that a classifier trained on data related to a specific urban area can be effectively used to predict the privacy risk of individuals in another urban area.

We want to highlight some limitations of the article that we plan to overcome in future works. First, we do not investigate Step (3) and Step (4) of the Data Delivery Procedure (Procedure 3.1), that is, the most suitable techniques to reduce the privacy risk of individuals in the dataset while still guaranteeing data quality for mobility analytics. Diverse techniques are proposed in the literature, such as k-anonymity (Samarati and Sweeney 1998b) or differential privacy (Dwork et al. 2006), ranging from removing a fraction of the records or individuals, to injecting artificial records to hide risky individuals, to modifying the data structures of the most risky individuals. Our approach provides a fast tool to immediately obtain the privacy risks of individuals, leaving to the Data Provider the choice of the most suitable privacy-preserving techniques to manage and mitigate the privacy risks of individuals. In future works, we plan to perform an extensive experimentation to select the best techniques to reduce the privacy risk of individuals in mobility datasets while at same time ensure high data quality for analytical services.

Our approach can be extended in several directions. First, we plan to apply our data mining approach to mobility datasets with different characteristics, such as mobile phone data which generally cover a larger geographic area (e.g., an entire country). This would allow us to deeply investigate the "portability" of our approach (i.e., at what extent the classifiers trained on a geographic zone can be used to predict the privacy risk of individuals in another geographic zone). Second, the repertoire of attacks can be extended by adding new attacks or by defining "multi-attacks" (i.e., combining multiple existing attacks). For example, a powerful "multi-attack" would be a combination of Proportion and Probability: An adversary would know a set of k locations and the corresponding probabilities and relative proportions. Many other multi-attacks can be designed, and we leave this interesting line of research for future work. Third, we plan to investigate whether our approach can be extended to contexts other than human mobility, such as the estimation of privacy risk in social networks. It would be indeed interesting to investigate at what extent data mining classifiers are able to infer the relations between social network metrics and individual risk of re-identification in social network data. Last, it would be interesting to repeat the experiments with a larger repertoire of machine learning algorithms and identify the best performer, or to combine them with boosting or bagging techniques to further improve the classification results. We leave these interesting tasks for future work.

REFERENCES

Osman Abul, Francesco Bonchi, and Mirco Nanni. 2008a. Never walk alone: Uncertainty for anonymity in moving objects databases. In *Proceedings of the 24th International Conference on Data Engineering (ICDE'08).* 376–385. DOI:https://doi.org/10.1109/ICDE.2008.4497446

Osman Abul, Francesco Bonchi, and Mirco Nanni. 2008b. Never walk alone: Uncertainty for anonymity in moving objects databases. In *ICDE'08.* 376–385.

Jagdish Prasad Achara, Gergely Ács, and Claude Castelluccia. 2015. On the unicity of smartphone applications. In *Proceedings of the 14th ACM Workshop on Privacy in the Electronic Society (WPES 2015), Denver, Colorado, USA, October 12, 2015.* 27–36. DOI:https://doi.org/10.1145/2808138.2808146

Christopher Alberts, Sandra Behrens, Richard Pethia, and William Wilson. 1999. *Operationally Critical Threat, Asset, and Vulnerability Evaluation (OCTAVE) Framework, Version 1.0.* Technical Report CMU/SEI-99-TR-017. Software Engineering Institute, Carnegie Mellon University, Pittsburgh, PA. http://resources.sei.cmu.edu/library/asset-view.cfm?AssetID=13473.

Alessandro Armando, Michele Bezzi, Nadia Metoui, and Antonino Sabetta. 2015. Risk-based privacy-aware information disclosure. *International Journal of Security Software Engineering* 6, 2 (April 2015), 70–89. DOI:https://doi.org/10.4018/IJSSE.2015040104

Anirban Basu, Anna Monreale, Juan Camilo Corena, Fosca Giannotti, Dino Pedreschi, Shinsaku Kiyomoto, Yutaka Miyake, Tadashi Yanagihara, and Roberto Trasarti. 2014. A privacy risk model for trajectory data. In *Trust Management VIII*, Jianying Zhou, Nurit Gal-Oz, Jie Zhang, and Ehud Gudes (Eds.). IFIP Advances in Information and Communication Technology, Vol. 430. Springer Berlin, 125–140. DOI:https://doi.org/10.1007/978-3-662-43813-8_9

Armando Bazzani, Bruno Giorgini, Sandro Rambaldi, Riccardo Gallotti, and Luca Giovannini. 2010. Statistical laws in urban mobility from microscopic GPS data in the area of Florence. *Journal of Statistical Mechanics: Theory and Experiment* 2010, 5 (2010), P05001. http://stacks.iop.org/1742-5468/2010/i=05/a=P05001

Alket Cecaj, Marco Mamei, and Franco Zambonelli. 2016. Re-identification and information fusion between anonymized CDR and social network data. *Journal of Ambient Intelligence and Humanized Computing* 7, 1 (2016), 83–96. DOI:https://doi.org/10.1007/s12652-015-0303-x

Vittoria Colizza, Alain Barrat, Marc Barthelemy, Alain-Jacques Valleron, and Alessandro Vespignani. 2007. Modeling the worldwide spread of Pandemic influenza: Baseline case and containment interventions. *PLOS Medicine* 4, 1 (Jan. 2007), 1–16. DOI:https://doi.org/10.1371/journal.pmed.0040013

Graham Cormode, Cecilia M. Procopiuc, Divesh Srivastava, and Thanh T. L. Tran. 2012. Differentially private summaries for sparse data. In *ICDT'12.* 299–311.

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 (March 2013), 1376. http://dx.doi.org/10.1038/srep01376.

Yves-Alexandre de Montjoye, Samuel S. Wang, and Alex Pentland. 2012. On the trusted use of large-scale personal data. *IEEE Data Engineering Bull.* 35, 4 (2012), 5–8.

Mina Deng, Kim Wuyts, Riccardo Scandariato, Bart Preneel, and Wouter Joosen. 2011. A privacy threat analysis framework: Supporting the elicitation and fulfillment of privacy requirements. *Requirements Engineering* 16, 1 (March 2011), 3–32. DOI:https://doi.org/10.1007/s00766-010-0115-7

Cynthia Dwork, Frank McSherry, Kobbi Nissim, and Adam Smith. 2006. Calibrating noise to sensitivity in private data analysis. In *TCC'06*. 265–284.

Nathan Eagle and Alex S. Pentland. 2009. Eigenbehaviors: Identifying structure in routine. *Behavioral Ecology and Sociobiology* 63, 7 (1 May 2009), 1057–1066. DOI: https://doi.org/10.1007/s00265-009-0739-0

Sébastien Gambs, Marc-Olivier Killijian, and Miguel Núñez del Prado Cortez. 2012. Next place prediction using mobility Markov chains. In *Proceedings of the 1st Workshop on Measurement, Privacy, and Mobility (MPM'12)*. ACM, New York, Article 3, 6 pages. DOI: https://doi.org/10.1145/2181196.2181199

Sebastien Gambs, Marc-Olivier Killijian, and Miguel Nuñez Del Prado Cortez. 2014. De-anonymization attack on geolocated data. *Journal of Computer System Science* 80 (2014), 1597–1614.

Fosca Giannotti, Anna Monreale, and Dino Pedreschi. 2013. Mobility data and privacy. In *Mobility Data Modeling, Management, and Understanding*, C. Renso, S. Spaccapietra, E. Zimanyi (Eds.). Springer, 174–193.

Fosca Giannotti, Mirco Nanni, Dino Pedreschi, Fabio Pinelli, Chiara Renso, Salvatore Rinzivillo, and Roberto Trasarti. 2011. Unveiling the complexity of human mobility by querying and mining massive trajectory data. *The VLDB Journal* 20, 5 (2011), 695. DOI: https://doi.org/10.1007/s00778-011-0244-8

Marta C. Gonzalez, Cesar A. Hidalgo, and Albert-Laszlo Barabasi. 2008. Understanding individual human mobility patterns. *Nature* 453, 7196 (June 2008), 779–782. DOI: https://doi.org/10.1038/nature06958

Trevor J. Hastie, Robert John Tibshirani, and Jerome H. Friedman. 2009. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. Springer, New York.

Shouling Ji, Weiqing Li, Mudhakar Srivatsa, Jing Selena He, and Raheem Beyah. 2014. *Structure Based Data De-Anonymization of Social Networks and Mobility Traces*. Springer International Publishing, Cham, 237–254. DOI: https://doi.org/10.1007/978-3-319-13257-0_14

Shan Jiang, Yingxiang Yang, Siddharth Gupta, Daniele Veneziano, Shounak Athavale, and Marta C. Gonzlez. 2016. The TimeGeo modeling framework for urban mobility without travel surveys. *Proceedings of the National Academy of Sciences* 113, 37 (2016), E5370–E5378. DOI: https://doi.org/10.1073/pnas.1524261113 arXiv:http://www.pnas.org/content/113/37/E5370.full.pdf.

Xin Lu, Erik Wetter, Nita Bharti, Andrew J. Tatem, and Linus Bengtsson. 2013. Approaching the limit of predictability in human mobility. *Scientific Reports* 3, 1, 2923. http://dx.doi.org/10.1038/srep02923

Stefano Marchetti, Caterina Giusti, Monica Pratesi, Nicola Salvati, Fosca Giannotti, Dino Pedreschi, Salvatore Rinzivillo, Luca Pappalardo, and Lorenzo Gabrielli. 2015. Small area model-based estimators using big data sources. *Journal of Official Statistics* 31, 2 (2015), 263–281.

J. D. Meier and Microsoft Corporation. 2003. *Improving Web Application Security: Threats and Countermeasures*. Microsoft.

Noman Mohammed, Benjamin C. M. Fung, and Mourad Debbabi. 2009. Walking in the crowd: Anonymizing trajectory data for pattern analysis. In *Proceedings of the 18th ACM Conference on Information and Knowledge Management (CIKM'09)*. ACM, New York, 1441–1444.

Anna Monreale, Gennady Andrienko, Natalia Andrienko, Fosca Giannotti, Dino Pedreschi, Salvatore Rinzivillo, and Stefan Wrobel. 2010a. Movement data anonymity through generalization. *Transactions on Data Privacy* 3, 2 (Aug. 2010), 91–121.

Anna Monreale, Gennady L. Andrienko, Natalia V. Andrienko, Fosca Giannotti, Dino Pedreschi, Salvatore Rinzivillo, and Stefan Wrobel. 2010b. Movement data anonymity through generalization. *Transactions on Data Privacy* 3, 2 (2010), 91–121.

Anna Monreale, Dino Pedreschi, Ruggero G. Pensa, and Fabio Pinelli. 2014a. Anonymity preserving sequential pattern mining. *Artificial Intelligence and Law* 22, 2 (2014), 141–173. DOI: https://doi.org/10.1007/s10506-014-9154-6

Anna Monreale, Salvatore Rinzivillo, Francesca Pratesi, Fosca Giannotti, and Dino Pedreschi. 2014b. Privacy-by-design in big data analytics and social mining. *EPJ Data Science* 3, 1 (2014), 10. DOI: https://doi.org/10.1140/epjds/s13688-014-0010-4

Anna Monreale, Wendy Hui Wang, Francesca Pratesi, Salvatore Rinzivillo, Dino Pedreschi, Gennady Andrienko, and Natalia Andrienko. 2013. *Privacy-Preserving Distributed Movement Data Aggregation*. Springer International Publishing, 225–245. DOI: https://doi.org/10.1007/978-3-319-00615-4_13

Arvind Narayanan and Vitaly Shmatikov. 2009. De-anonymizing social networks. In *Proceedings of the 30th IEEE Symposium on Security and Privacy (S&P'09)*. 173–187. DOI: https://doi.org/10.1109/SP.2009.22

OWASP. 2016. Risk rating methodology. Retrieved from https://www.owasp.org/index.php/OWASP_Risk_Rating_Methodology.

Luca Pappalardo, Dino Pedreschi, Zbigniew Smoreda, and Fosca Giannotti. 2015. Using big data to study the link between human mobility and socio-economic development. In *Proceedings of the 2015 IEEE International Conference on Big Data (Big Data'15)*. 871–878. DOI: https://doi.org/10.1109/BigData.2015.7363835

Luca Pappalardo, Salvatore Rinzivillo, Zehui Qu, Dino Pedreschi, and Fosca Giannotti. 2013. Understanding the patterns of car travel. *The European Physical Journal Special Topics* 215, 1 (2013), 61–73. DOI: https://doi.org/10.1140/epjst%252fe2013-01715-5

Luca Pappalardo, Salvatore Rinzivillo, and Filippo Simini. 2016. Human mobility modelling: Exploration and preferential return meet the gravity model. *Procedia Computer Science* 83 (2016), 934–939. DOI:https://doi.org/10.1016/j.procs.2016.04.188 The 7th International Conference on Ambient Systems, Networks and Technologies (ANT 2016) / The 6th International Conference on Sustainable Energy Information Technology (SEIT-2016) / Affiliated Workshops.

Luca Pappalardo and Filippo Simini. 2016. Modelling spatio-temporal routines in human mobility. *CoRR* abs/1607.05952 (2016). http://arxiv.org/abs/1607.05952

Luca Pappalardo, Filippo Simini, Salvatore Rinzivillo, Dino Pedreschi, Fosca Giannotti, and Albert-Laszlo Barabasi. 2015. Returners and explorers dichotomy in human mobility. *Nature Communications* 6 (Sept. 2015). http://dx.doi.org/10.1038/ncomms9166

Luca Pappalardo, Maarten Vanhoof, Lorenzo Gabrielli, Zbigniew Smoreda, Dino Pedreschi, and Fosca Giannotti. 2016. An analytical framework to nowcast well-being using mobile phone data. *International Journal of Data Science and Analytics* 2, 1 (2016), 75–92. DOI:https://doi.org/10.1007/s41060-016-0013-2

F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. 2011. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research* 12 (2011), 2825–2830.

Francesca Pratesi, Anna Monreale, Roberto Trasarti, Fosca Giannotti, Dino Pedreschi, and Tadashi Yanagihara. 2016. *PRISQUIT: A System for Assessing Privacy Risk versus Quality in Data Sharing*. Technical Report 2016-TR-043. ISTI - CNR, Pisa, Italy.

Arthi Ramachandran, Yunsung Kim, and Augustin Chaintreau. 2014. "I knew they clicked when I saw them with their friends": Identifying your silent web visitors on social media. In *Proceedings of the 2nd ACM Conference on Online Social Networks (COSN'14)*. 239–246. DOI:https://doi.org/10.1145/2660460.2660461

Ira S. Rubinstein. 2013. Big data: The end of privacy or a new beginning? *International Data Privacy Law* (2013). DOI:https://doi.org/10.1093/idpl/ips036 arXiv:http://idpl.oxfordjournals.org/content/early/2013/01/24/idpl.ips03 6.full.pdf+html

Pierangela Samarati. 2001. Protecting respondents' identities in microdata release. *IEEE Transactions on Knowledge and Data Engineering* 13, 6 (2001), 1010–1027. DOI:https://doi.org/10.1109/69.971193

Pierangela Samarati and Latanya Sweeney. 1998a. Generalizing data to provide anonymity when disclosing information (abstract). In *PODS*. 188.

Pierangela Samarati and Latanya Sweeney. 1998b. Protecting privacy when disclosing information: K-anonymity and its enforcement through generalization and suppression. In *Proceedings of the IEEE Symposium on Research in Security and Privacy*. 384–393.

Filippo Simini, Marta C. Gonzalez, Amos Maritan, and Albert-Laszlo Barabasi. 2012. A universal model for mobility and migration patterns. *Nature* 484, 7392 (05 04 2012), 96–100. http://dx.doi.org/10.1038/nature10856

Chaoming Song, Tal Koren, Pu Wang, and Albert-Laszlo Barabasi. 2010a. Modelling the scaling properties of human mobility. *Nature and Physics* 6, 10 (10 2010), 818–823. http://dx.doi.org/10.1038/nphys1760

Chaoming Song, Zehui Qu, Nicholas Blumm, and Albert-Lszl Barabsi. 2010b. Limits of predictability in human mobility. *Science* 327, 5968 (2010), 1018–1021. DOI:https://doi.org/10.1126/science.1177170 arXiv:http://www.sciencemag.org/cgi/reprint/327/5968/1018.pdf

Yi Song, Daniel Dahlmeier, and Stéphane Bressan. 2014. Not so unique in the crowd: A simple and effective algorithm for anonymizing location data. In *Proceeding of the 1st International Workshop on Privacy-Preserving IR: When Information Retrieval Meets Privacy and Security co-located with 37th Annual International ACM SIGIR Conference (PIR@SIGIR'14)*. 19–24.

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. U.S. Department of Commerce, National Institute of Standards and Technology.

Latanya Sweeney. 2002. K-anonymity: A model for protecting privacy. *International Journal of Uncertainty and Fuzziness in Knowledge-Based Systems* 10, 5 (Oct. 2002), 557–570. DOI:https://doi.org/10.1142/S0218488502001648

Frank Swiderski and Window Snyder. 2004. *Threat Modeling*. O'Reilly Media.

Pang-Ning Tan, Michael Steinbach, and Vipin Kumar. 2005. *Introduction to Data Mining (1st Edition)*. Addison-Wesley Longman Publishing Co., Inc., Boston, MA.

Manolis Terrovitis and Nikos Mamoulis. 2008. Privacy preservation in the publication of trajectories. In *MDM*. 65–72.

Manolis Terrovitis, Nikos Mamoulis, and Panos Kalnis. 2008. Privacy-preserving anonymization of set-valued data. *Proceedings of the VLDB Endowment* 1, 1 (Aug. 2008), 115–125. DOI:https://doi.org/10.14778/1453856.1453874

Michele Tizzoni, Paolo Bajardi, Adeline Decuyper, Guillaume Kon Kam King, Christian M. Schneider, Vincent Blondel, Zbigniew Smoreda, Marta C. Gonzlez, and Vittoria Colizza. 2014. On the use of human mobility proxies for modeling epidemics. *PLOS Computational Biology* 10, 7 (Jul. 2014), 1–15. DOI:https://doi.org/10.1371/journal.pcbi.1003716

Slim Trabelsi, Vincent Salzgeber, Michele Bezzi, and Gilles Montagnon. 2009. Data disclosure risk evaluation. In *CRiSIS'09*. 35–72.

Jayakrishnan Unnikrishnan and Farid Movahedi Naini. 2013. De-anonymizing private data by matching statistics. In *Proceedings of the 51st Annual Allerton Conference on Communication, Control, and Computing (Allerton'13)*. 1616–1623. DOI:https://doi.org/10.1109/Allerton.2013.6736722

Pu Wang, Timothy Hunter, Alexandre M. Bayen, Katja Schechtner, and Marta C. González. 2012. Understanding road usage patterns in urban areas. *Scientific Reports* 2 (Dec. 2012), 1001 EP. http://dx.doi.org/10.1038/srep01001

Nathalie E. Williams, Timothy A. Thomas, Matthew Dunbar, Nathan Eagle, and Adrian Dobra. 2015. Measures of human mobility using mobile phone records enhanced with GIS data. *PLoS ONE* 10, 7 (07 2015), 1–16. DOI:https://doi.org/10.1371/journal.pone.0133630

W. K. Wong, David W. Cheung, Edward Hung, Ben Kao, and Nikos Mamoulis. 2007. Security in outsourcing of association rule mining. In *Proceedings of the 33rd International Conference on Very Large Data Bases (VLDB'07)*. VLDB Endowment, 111–122.

World Economic Forum. 2013. Unlocking the Value of Personal Data: From Collection to Usage. Retrieved from http://www3.weforum.org/docs/WEF_IT_UnlockingValuePersonalData_CollectionUsage_Report_2013.pdf.

Yabo Xu, Benjamin C. M. Fung, Ke Wang, Ada Wai-Chee Fu, and Jian Pei. 2008a. Publishing sensitive transactions for itemset utility. In *Proceedings of the 8th IEEE International Conference on Data Mining (ICDM'08)*. 1109–1114.

Yabo Xu, Ke Wang, Ada Wai-Chee Fu, and Philip S. Yu. 2008b. Anonymizing transaction databases for publication. In *Proceedings of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. 767–775.

Roman Yarovoy, Francesco Bonchi, Laks V. S. Lakshmanan, and Wendy Hui Wang. 2009. Anonymizing moving objects: How to hide a MOB in a crowd? In *EDBT*. 72–83.

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, 145–156. DOI:https://doi.org/10.1145/2030613.2030630

Yu Zheng. 2015. Trajectory data mining: An overview. *ACM Transactions on Intelligent Systems Technology* 6, 3 (2015), 29:1–29:41. DOI:https://doi.org/10.1145/2743025

Yu Zheng, Licia Capra, Ouri Wolfson, and Hai Yang. 2014. Urban computing: Concepts, methodologies, and applications. *ACM Transactions on Intelligent Systems Technology* 5, 3 (Sept. 2014), Article 38, 55 pages. DOI:https://doi.org/10.1145/2629592

Yu Zhengand Xiaofang Zhou (Eds.). 2011. *Computing with Spatial Trajectories*. Springer. DOI:https://doi.org/10.1007/978-1-4614-1629-6

Received December 2016; revised May 2017; accepted May 2017