Fast estimation of privacy risk in human mobility data

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

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

Abstract. Mobility data are an important proxy to understand the patterns of human movements, develop analytical services and design models for simulation and prediction of human dynamics. Unfortunately mobility data are also very sensitive, since they may contain personal information about the individuals involved. Existing frameworks for privacy risk assessment enable the data providers to quantify and mitigate privacy risks, but they suffer two main limitations: (i) they have a high computational complexity; (ii) the privacy risk must be re-computed for each new set of individuals, geographic areas or time windows. In this paper we explore a fast and flexible solution to estimate privacy risk in human mobility data, using predictive models to capture the relation between an individual's mobility patterns and her privacy risk. We show the effectiveness of our approach by experimentation on a real-world GPS dataset and provide a comparison with traditional methods.

1 Introduction

In the last years human mobility analysis has attracted a growing interest due to its importance in several applications such as urban planning, transportation engineering and public health (10). The availability of these data has offered the opportunity to observe human movements at large scales and in great detail, leading to the discovery of quantitative patterns (8), the mathematical modeling of human mobility (9; 14) etc. Unfortunately mobility data are sensitive because they may reveal personal information or allow the re-identification of individuals, creating serious privacy risks if they are analyzed with malicious intent (12). Driven by these sensitive issues, researchers have developed methodologies and frameworks to mitigate the individual privacy risks associated to the study of GPS trajectories and Big Data in general (1). These tools aim at preserving both the right to individual's privacy and the effectiveness of the analytical results, trying to find a reasonable trade-off between privacy protection and data quality. They allow the definition of infrastructures for supporting privacy and of technical requirements for data protection, enforcing cross-relations between privacy-preserving solutions and legal regulations, since assessing privacy risk is required by the new EU General Data Protection Regulation. To this aim, Pratesi et al. (11) propose a framework for the privacy risk assessment of individuals in mobility datasets. Although frameworks like the one presented in (11) are effective in many scenarios, they suffer a drawback: the privacy risk assessment has a high computational complexity (non-polynomial in time) because it computes the maximum privacy risk given an external knowledge that a malicious adversary may have, i.e., it considers all the possible ways the adversary can try to re-identify an individual. Moreover, 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 paper we propose a data mining approach for privacy risk assessment that overcomes the computational limitations of existing frameworks. We first introduce some possible re-identification attacks on mobility data, and then we use linear regression to predict the privacy risk of an individual based on her mobility patterns, and we compute the individual privacy risk level according to the re-identification attacks. We then train a regressor on such data to estimate in polynomial time the privacy risk level of *previously unseen* vehicles based just on their individual mobility patterns. In a scenario where a Data Analyst asks a Data Provider for mobility data to deploy an analytical service, the Data Provider (e.g., a mobile phone carrier) can use the regressor to immediately identify individuals with a high privacy risk. Then, the Data Provider can select the most suitable privacy-preserving technique (e.g., k-anonymity, differential privacy) to mitigate their privacy risk and release only safe data to the Data Analyst. Our experiments on GPS data shows that our approach is fairly accurate in predicting the privacy risk of unseen individuals in an urban area.

The rest of the paper 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. Section 6 presents the main works related to our paper and finally Section 7 concludes the paper proposing some lines of new research.

2 Data Definitions

The approach we present in this paper is tailored for human mobility data, i.e., data describing the movements of a set of individuals. This type of data is generally collected in an automatic way through electronic devices (e.g., mobile phones, GPS devices) in form of raw trajectory data. Every record has the following fields: the identifier of the individual, a geographic location expressed in coordinates (generally latitude and longitude), a timestamp indicating when the individual stopped in or went through that location. Depending on the specific application, a trajectory can be aggregated into different data structures:

Definition 1 (Trajectory). *The trajectory T_u of an individual u is a temporally ordered sequence of tuples $T_u = \langle (l_1, t_1), (l_2, t_2), \ldots, (l_n, t_n) \rangle$, 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 (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.*

We denote with D a mobility dataset, which we assume is a set of a one of the above data types (trajectory or frequency vector).

3 Privacy Risk Assessment Framework

In this paper we consider the work proposed in (11), which allows for the privacy risk assessment of human mobility data. This framework considers a scenario where a Data Analyst asks a Data Provider for data to develop an analytical service. The Data Provider must guarantee the right to privacy of the individuals whose data are recorded. First, the Data Analyst transmits to the Data Provider the data requirements for the service. With these specifications, the Data Provider queries its dataset $\mathcal{D}$, producing a set of datasets $\{D_1, \ldots, D_z\}$, each with different data structures and data aggregations. The Data Provider then reiterates a procedure until it considers the data delivery safe:

(1) *Identification of Attacks*: identify a set of possible attacks that an adversary might conduct in order to re-identify individuals in the 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 individuals and the data quality, given a certain level of tolerated privacy risk and the Data Analyst's requirements;
(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 D_{filt} to the Data Analyst when the D_{filt} is adequately safe.

In this paper we focus on improving step (2), 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 a malicious adversary can perform and computing the privacy risks associated to each attack. The privacy risk of an individual is related to her probability of re-identification in a dataset w.r.t. to a set of re-identification attacks. A re-identification attack assumes that an adversary gains access to a 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 re-identify all the records in the dataset regarding the individual under attack. In this paper we use the definition of privacy risk (or re-identification risk) introduced in (13).

A background knowledge represents both the kind and quantity of information known by an adversary. Two examples of kinds of background knowledge 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). We denote with k the number of the elements known by the adversary. So for example a specific background knowledge is the knowledge of three specific locations visited by the individual under attack. We denote a set of background knowledge of size k with B_k and a specific background knowledge with b.

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 (Probability of re-identification). *Given an attack, a function* matching(d, b) *indicating whether or not a record $d \in D$ matches the background knowledge b, 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 record $d \in D$ to individual u, given background knowledge b.*

Note that $PR_D(d{=}u|b) = 0$ if the individual u is not represented in D. Since each background knowledge b 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 possible background knowledge:

Definition 4 (Risk of re-identification or Privacy risk). The risk of re-identification (or privacy risk) of an individual u given a set of background knowledge 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 $u \notin D$.

An individual is hence associated to several privacy risks, each for every background knowledge of an attack. Every privacy risk of an individual can be computed using the following procedure: *(i)* define an attack based on a specific background knowledge, *(ii)* given an individual and fixing k, compute all the possible $b \in B_k$ and the corresponding probability of re-identification, and *(iii)* select the privacy risk of the individual for a set B_k as the maximum probability of re-identification across all $b \in B_k$.

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. The maximum possible value of k is len, the length of the data structure of an individual. Since it is unlikely that an adversary knows the complete movement of an individual (i.e., all the points), we have to reason about different and reasonable values of k. To compute all $b \in B_k$ we have to compute a k-combination of elements from the original data structure. We need all b to correctly compute the risk of re-identification, since we have to know all the possible probabilities of re-identification. This leads to a high overall computational complexity $\mathcal{O}(\binom{len}{k} \times N)$, since the framework generates $\binom{len}{k}$ background knowledge b and, for each b, it executes N matching operations by applying function *matching*. While some optimizations can be made depending on the kind of attack simulated, the overall complexity of the procedure is dominated by the $\binom{len}{k}$ term.

4 Fast Privacy Risk Assessment with Data Mining

Given its computational complexity, the privacy risk computation becomes unfeasible as the size of the dataset increases. This drawback is even more serious if we consider that the privacy risks must be necessarily re-computed every time the mobility dataset is updated and for every selection of individuals, geographic areas and periods of time. In order to overcome these problems, we propose a fast and flexible data mining approach. The idea is to train a regression model to predict the privacy risk of an individual based solely on her individual mobility patterns. The training of the predictive model is made by using a dataset where every record refers to an individual and consists of *(i)* a vector of the individual's mobility features and *(ii)* the privacy risk value of the individual. We make our approach parametric with respect to the predictive algorithm: in our experiments we use a Random Forest regressor, but every algorithm available in 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 (11).

4.1 Individual Mobility Features

The mobility dynamics of an individual can be described by a set of measures widely used in literature. 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 (8). 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 (9). The length *Locs* of the frequency vector indicates the number of distinct places visited by an individual during the period of observation (14). Dividing *Locs* by the number of available locations on 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 is defined as the length of the longest trip of an individual (19), while D_{max}^{trip} is defined as the ratio between D_{max} and the maximum possible distance between the locations in the area. The sum of all the trip lengths is defined as D_{sum} (19). It can be also averaged over the days in the period of observation obtaining $\overline{D}_{sum}$. The radius of gyration r_g is the characteristic distance traveled by an individual during the period of observation (8). The mobility entropy E is a measure of the predictability of an individual's trajectory (6). Also, for each individual we keep track of the characteristics of three different locations: the most, the second most and the least visited location. The frequency w_i of a location i is the number of times an individual visited i during the period of observation, while the average frequency $\overline{w}_i$ is the daily average frequency of i. We also define w_i^{pop} as the frequency of location i divided by the popularity of i in the whole dataset. The quantity U_i^{ratio} is the number of distinct individuals that visited location i divided by the total number $|U_{set}|$ of individuals in the dataset, while U_i is the number of distinct individuals that visited i during the period of observation. Finally, the location entropy E_i is the predictability of i, defined as a variation of the Shannon entropy.

Every individual u in the dataset is described by a mobility vector $\overline{m}_u$ of the 16 mobility features described above. 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 paper:

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.

Location Sequence Attack. Here, the adversary knows a subset of the locations visited by the individual and the temporal ordering of the visits.

Visit Attack. In a Visit attack the adversary knows a subset of the locations visited by the individual and the time the individual visited these locations.

Frequent Location and Sequence Attack. We introduce two attacks based on location knowledge applied to frequency vectors. 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 but in frequency vectors a location can appear only once. The Frequent Location Sequence attack is similar to the Location Sequence attack, but a location can appear only once in the vector and locations are ordered by descending frequency and not by time.

Frequency Attack. We introduce an attack where the adversary knows the locations visited by the individual, their reciprocal ordering of frequency, and the minimum number of visits of the individual. This means that, when searching for specific subsequences, the adversary must consider also subsequences containing the known locations with a greater frequency.

Home And Work Attack. In the Home and Work attack the adversary knows the two most frequent locations of an individual and their frequencies. It assumes the same background knowledge of Frequency attack but related only to two locations. This is the only attack where the set of background knowledge is fixed and composed of just a single 2-combination for each individual.

4.3 Construction of training dataset

Given an attack i based on a specific set of background knowledge B^i_j, the regression training dataset TR^i_j can be constructed by the following procedure: first, given a mobility dataset D, for every individual u we compute the set of 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' feature vectors compose mobility matrix $F{=}(\overline{m}_1, \ldots, \overline{m}_n)$, where n is the number of individuals in D. Second, for every individual we simulate the attack with B^i_j on D, in order to compute a privacy risk value for every individual. We obtain a privacy risk vector $R^i_j = (r_1, \ldots, r_n)$. The regression training set is hence $\mathrm{TR}^i_j = (F, R^i_j)$;

Every regression dataset TR^i_j is used to train a predictive model M^i_j. If $0 \leq i \leq I$ where I is the number of different kinds of attack and $0 \leq j \leq J$ where J is the number of different sets of possible background knowledge, we have a total of $J \times I$ models. For example, if we consider sets of background knowledge ranging in size from $j = 1$ to $j = 5$ for 7 different attacks, we would have $I = 7$ and $J = 5$. The predictive model will be used by the Data Provider to immediately estimate the privacy risk value of *previously unseen* individuals, whose data were not used in the learning process, with respect to attack i, set of background knowledge B^i_j and dataset D.

Example 1 (Construction of regression training set). Let us 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 set of background knowledge B^i_j and dataset D, we construct the regression training set TC^i_j as follows: first, for every individual u_i we compute the 21 individual mobility measures based on her trajectory T_{u_i}. Every individual u_i is hence described by a mobility feature vector of length 21 $\overline{m}_{u_i} = (m^{(u_i)}_1, \ldots, m^{(u_i)}_{21})$. All the mobility feature vectors compose mobility matrix $F{=}(\overline{m}_{u_1}, \overline{m}_{u_2}, \overline{m}_{u_3}, \overline{m}_{u_4}, \overline{m}_{u_5})$; second, we simulate the attack with B^i_j on dataset D and obtain a vector of five privacy risk values $R^i_j = (r_{u_1}, r_{u_2}, r_{u_3}, r_{u_4}, r_{u_5})$, each for every individual.

4.4 Usage of the regression approach

The Data Provider can use a regression model M_j^i to determine the value of privacy risk with respect to an attack i and a set of background knowledge B_j^i for: *(i) previously unseen* individuals, whose data were *not* used in the learning process; *(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 regression model M_j^i allows for obtaining the privacy risk of the selected individuals immediately. The computation of the mobility measures and the regression of privacy risk can be done in polynomial time as a one-off procedure. To clarify this point, let us 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 frequency vector, the Data Provider can release just the frequency vectors of the individuals requested. Before that, however, the Data Provider wants to determine the level of privacy risk of the individuals with respect to the Frequency attack (F) and several sets of background knowledge B_j^F. The Data Provider uses the regression model M_j^F previously trained to obtain the privacy risk of the individuals. So the Data Provider computes the mobility features for the individuals in the dataset and gives them in input to the regression model, obtaining an estimation of privacy risk. On the basis of privacy risks obtained from M_j^F, the Data Provider can identify risky individuals, i.e., individuals with a high privacy risk. She then can decide to either filter out the risky individuals or to select suitable privacy-preserving techniques (e.g., k-anonymity or differential privacy) and transform their mobility data in such a way that their privacy is preserved.

5 Experiments

For all the attacks defined except the Home and Work attack we consider four sets of background knowledge B_k with $k = 2, 3, 4, 5$, where each 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 set of background knowledge, where the adversary knows the two most frequent locations of an individual. We use a dataset provided by Octo Telematics [3] storing the GPS tracks of 9,715 private vehicles traveling in Florence, a very populous area of central Italy, from 1st May to 31st May 2011, corresponding to 179,318 trajectories. We assign each origin and destination point of the original raw trajectories to the corresponding census cell according to the information provided by the Italian National Statistics Bureau (8). We first performed a simulation of the attacks computing the privacy risk values for all individuals in the dataset and

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

for all B_k.[4] We then performed regression experiments using a Random Forest regressor.[5] Table 1 shows the average Mean Squared Error (mse) and the average coefficient of determination R^2 resulting from the regression experiments for all the attacks. The results are averaged over $k = 2, 3, 4, 5$, since the empirical distributions of privacy risk are fairly similar across different values of k. Also, mse and R^2 are almost identical for each kind of attack. The best results are obtained for the Frequent Location Sequence attack, with values of mse = 0.01 and $R^2 = 0.92$, while the weakest results are obtained for the Home and Work attack, with values of mse = 0.07 and $R^2 = 0.50$. Overall, the results show good predictive performance across all attacks, suggesting that regression could indeed be an accurate alternative to the direct computation of privacy risk.

Table 1. Results of regression experiments.

predicted variable	**mse**	**r2**
Frequent Location Sequence	0.01	0.92
Visit	0.01	0.89
Frequency	0.02	0.88
Location	0.02	0.90
Location Sequence	0.02	0.84
Frequent Location	0.03	0.73
Home and Work	0.07	0.50

Execution Times. We show the computational improvement of our approach in terms of execution time by comparing in Table 2 the execution times of the attack simulations and the execution times of the regression tasks.[6] The execution time of a single regression task is the sum of three subtasks: *(i)* the execution time of training the regressor on the training set; *(ii)* the execution time of using the trained regressor to predict the risk on the test set; *(iii)* the execution time of evaluating the performance of regression. Table 2 shows that the execution time of attack simulations is low for most of the attacks except for Location Sequence and Location, for which execution times are huge: more than 1 week each. In contrast the regression tasks have constant execution times of around 22s. In summary, our approach can compute the risk levels for all the 33 attacks in 179 seconds (less than 3 minutes), while the attack simulations require more than two weeks of computation.

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

[5] We use the Python package `scikit-learn` to perform the regression experiments.

[6] 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, 32 GB RAM, 3.30GHz Intel Core i7.

Table 2. Execution times of attack simulations and regression tasks.

variable ($\sum_2^5 k$)	**simulation**	**regression**
Home and Work	149s (2.5m)	7s
Frequency	645s (10m)	22s
Frequent Location Sequence	846s (14m)	22s
Frequent Location	997s (10m)	22s
Visit	2,274s (38m)	16s
LocationSequence	> 168h (1week)	22s
Location	> 168h (1week)	22s
total	**> 2weeks**	**172s**

Discussion. The preliminary work presented above shows some promising results. The coefficient of determination and the execution times suggest that the regression can be a valid and fast alternative to existing privacy risk assessment tools. Instead of re-computing privacy risks when new data records become available, which would result in high computational costs, a Data Provider can effectively use the regressors to obtain immediate and reliable estimates for every individual. The mobility measures can be computed in linear time of the size of the dataset. Every time new mobility data of 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 six months of data).

6 Related Works

Human mobility data contains personal sensitive information and can reveal many facets of the private life of individuals, leading to potential privacy violation. To overcome the possibility of privacy leaks, many techniques have been proposed in literature. A widely used privacy-preserving model is k-anonymity (13), which requires that an individual should not be identifiable from a group of size smaller than k based on their quasi-identifiers (QIDs), i.e., a set of attributes that can be used to uniquely identify individuals. Assuming that adversaries own disjoint parts of a trajectory, (17) reduces privacy risk by relying on the suppression of the dangerous observations from each individual's trajectory. In (20), authors propose the attack-graphs method to defend against attacks, based on k-anonymity. Other works are based on the differential privacy model (5). (7) considers a privacy-preserving distributed aggregation framework for movement data. (3) proposes to publish a contingency table of trajectory data, where each cell contains the number of individuals commuting from a source to a destination. (24) defines several similarity metrics which can be combined in a unified framework to provide de-anonymization of mobility data and social network data. One of the most important work about privacy risk assessment is the Linddun methodology (4), a privacy-aware framework, useful for modeling privacy threats in software-based systems. In the last years, different techniques for risk management have been proposed, such as NIST's Special Publication 800-30 (16). Unfortunately, many of these works do not consider privacy risk assessment and simply include privacy considerations when assessing the impact of threats. In (18), authors elaborate an entropy-based method to evaluate the disclosure risk of personal data, trying to manage quantitatively privacy risks. The *unicity* measure proposed in (15) evaluates the privacy risk as the number of records/trajectories which are uniquely identified. (2) proposes a risk-aware framework for information disclosure which supports runtime risk assessment, using adaptive anonymization as risk-mitigation method. Unfortunately, this framework only works on relational datasets since it needs to discriminate between quasi-identifiers and sensitive attributes. In this paper we use the privacy risk assessment framework introduced by (11) to calculate the privacy risks of each individual in a mobility dataset.

7 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 can lead to a serious violation of the privacy of the individuals involved. In this paper we explored a fast and flexible solution for estimating the privacy risk in human mobility data, which overcomes the computational issues of existing privacy risk assessment frameworks. We showed through experimentations that our approach can achieve good estimations of privacy risks. As future work, it would be necessary to test our approach more extensively on different datasets and to evaluate the importance of mobility features with respect to the prediction of risk. Another possible extension of our method would be to apply more refined data mining techniques to assess the privacy risk of individuals. Moreover, 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. It would be interesting to perform an extensive experimentation to select the best techniques to reduce the privacy risk of individuals in mobility datasets and at same time ensuring high data quality for analytical services.

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

References

1. O. Abul, F. Bonchi, and M. Nanni. Never Walk Alone: Uncertainty for Anonymity in Moving Objects Databases. In *ICDE 2008*. 376–385.
2. A. Armando, M. Bezzi, N. Metoui, and A. Sabetta. Risk-Based Privacy-Aware Information Disclosure. *Int. J. Secur. Softw. Eng.* 6, 2 (April 2015), 70–89.

3. G. Cormode, C. M. Procopiuc, D. Srivastava, and T. T. L. Tran. Differentially private summaries for sparse data. In *ICDT '12*. 299–311.
4. M. Deng, K. Wuyts, R. Scandariato, B. Preneel, and W. Joosen. A privacy threat analysis framework: supporting the elicitation and fulfillment of privacy requirements. *Requir. Eng.*16, 1 (2011), pp 3–32.
5. C. Dwork, F. McSherry, K. Nissim, and A. Smith. 2006. Calibrating Noise to Sensitivity in Private Data Analysis. In *TCC '06*. 265–284.
6. N. Eagle and A. S. Pentland. Eigenbehaviors: identifying structure in routine. *Behavioral Ecology and Sociobiology* 63, 7 (2009), 1057–1066.
7. A. Monreale, W. H. Wang, F. Pratesi, S. Rinzivillo, D. Pedreschi, G. Andrienko, and N. Andrienko. 2013. *Privacy-Preserving Distributed Movement Data Aggregation*. Springer International Publishing, 225–245.
8. L. Pappalardo, F. Simini, S. Rinzivillo, D. Pedreschi, F. Giannotti, and A.-L. Barabasi. Returners and explorers dichotomy in human mobility. *Nature Communications* 6 (2015).
9. L. Pappalardo and F. Simini. Modelling spatio-temporal routines in human mobility. *CoRR* abs/1607.05952 (2016).
10. L. Pappalardo, M. Vanhoof, L. Gabrielli, Z. Smoreda, D. Pedreschi, and F. Giannotti. An analytical framework to nowcast well-being using mobile phone data. *International Journal of Data Science and Analytics* 2, 1 (2016), 75–92.
11. F. Pratesi, A. Monreale, R. Trasarti, F. Giannotti, D. Pedreschi, and T. Yanagihara. *PRISQUIT: a System for Assessing Privacy Risk versus Quality in Data Sharing*. Technical Report 2016-TR-043. ISTI - CNR, Pisa, Italy. FriNov20162291.
12. I. S. Rubinstein. Big Data: The End of Privacy or a New Beginning? *International Data Privacy Law* (2013).
13. P. Samarati and L. Sweeney. 1998a. Generalizing Data to Provide Anonymity when Disclosing Information (Abstract). In *PODS*. 188.
14. C. Song, T. Koren, P. Wang, and A.-L. Barabasi. Modelling the scaling properties of human mobility. *Nat Phys* 6, 10 (2010), 818–823.
15. Y. Song, D. Dahlmeier, and S. Bressan. Not So Unique in the Crowd: a Simple and Effective Algorithm for Anonymizing Location Data. In *PIR@SIGIR 2014*. 19–24.
16. 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.
17. M. Terrovitis and N. Mamoulis. 2008. Privacy Preservation in the Publication of Trajectories. In *MDM*. 65–72.
18. S. Trabelsi, V. Salzgeber, M. Bezzi, and G. Montagnon. 2009. Data disclosure risk evaluation. In *CRiSIS '09*. 35–72.
19. N. E. Williams, T. A. Thomas, M. Dunbar, N. Eagle, and A. Dobra. Measures of Human Mobility Using Mobile Phone Records Enhanced with GIS Data. *PLoS ONE* 10, 7 (2015), 1–16.
20. R. Yarovoy, F. Bonchi, L. V. S. Lakshmanan, and W. H. Wang. 2009. Anonymizing moving objects: how to hide a MOB in a crowd?. In *EDBT*. 72–83.
21. N. Mohammed, B. C.M. Fung, and M. Debbabi. Walking in the Crowd: Anonymizing Trajectory Data for Pattern Analysis. In *CIKM 2009*. 1441–1444.
22. H. Zang and J. Bolot. Anonymization of Location Data Does Not Work: A Large-scale Measurement Study. In *MobiCom 2011*. 145–156.
23. J. Unnikrishnan and F. M. Naini. De-anonymizing private data by matching statistics. In *Allerton 2013*. 1616–1623.
24. S. Ji, Weiqing Li, M. Srivatsa, J. S. He, and R. Beyah. 2014. *Structure Based Data De-Anonymization of Social Networks and Mobility Traces*. 237–254.