A Privacy Risk Model for Trajectory Data

Anirban Basu[1], Anna Monreale[2], Juan Camilo Corena[1], Fosca Giannotti[3], Dino Pedreschi[2], Shinsaku Kiyomoto[1], Yutaka Miyake[1], Tadashi Yanagihara[4], and Roberto Trasarti[3]

[1] KDDI R&D Laboratories, Japan
{basu,corena,kiyomoto,miyake}@kddilabs.jp
[2] University of Pisa, Italy
{annam,dino}@di.unipi.it
[3] ISTI-CNR, Italy
{fosca.giannotti,roberto.trasarti}@isti.cnr.it
[4] Toyota ITC, Japan
ta-yanagihara@jp.toyota-itc.com

Abstract. Time sequence data relating to users, such as medical histories and mobility data, are good candidates for data mining, but often contain highly sensitive information. Different methods in privacy-preserving data publishing are utilised to release such private data so that individual records in the released data cannot be re-linked to specific users with a high degree of certainty. These methods provide theoretical worst-case privacy risks as measures of the privacy protection that they offer. However, often with many real-world data the worst-case scenario is too pessimistic and does not provide a realistic view of the privacy risks: the real probability of re-identification is often much lower than the theoretical worst-case risk. In this paper we propose a novel empirical risk model for privacy which, in relation to the cost of privacy attacks, demonstrates better the practical risks associated with a privacy preserving data release. We show detailed evaluation of the proposed risk model by using k-anonymised real-world mobility data.

Keywords: privacy, risk, utility, model, anonymisation, sequential data.

1 Introduction

The big data originating from the digital breadcrumbs of human activities, sensed as a by-product of the ICT systems, record different dimensions of human social life. These data describing human activities are valuable assets for data mining and big data analytics and their availability enables a new generation of personalised intelligent services. Most of these data are of sequential nature, such as time-stamped transactions, users' medical histories and trajectories. They describe sequences of events or users' actions where the timestamps make the temporal sequentiality of the events powerful sources of information. Unfortunately, such information often contain sensitive information that are protected under the legal frameworks for user data protection. Thus, when such data has to be released to any third party for analysis, privacy-preserving mechanisms are utilised

J. Zhou et al. (Eds.): IFIPTM 2014, IFIP AICT 430, pp. 125–140, 2014.

to de-link individual records from their associated users. Privacy-preserving data publishing (PPDP) aims at preserving statistical properties of the data while removing the details that can help the re-identification of users. Any PPDP method provides a worst-case probabilistic risk of user re-identification as a measure for how safe the anonymised data is.

One such well-known anonymisation model typically used for PPDP is the k-anonymity model [1, 2]. It states that in the worst case, there are at least k (and no less) users that can be re-identified given a k-anonymised dataset. Thus, the re-identification probability for any single user, in the worst case, is equal to $1/k$. The higher the value of k, the lower the probability of any attack succeeding. However, at the same time the higher the value of k, the lower the utility of the data where the utility relates how well the anonymised data represents the original one. This worst case scenario hardly gives us the view of the realistic re-identification probabilities, which are often much lower than $1/k$. We envisage that the worst case guarantee, by itself, is not sufficient to help the user understand the risks; and it is also not enough to communicate in a legal language the risks associated with any of these anonymisation methods.

In this paper, we propose an empirical risk model for privacy based on k-anonymous data release. We also discuss the relation of risk to the cost of any attack on privacy as well as the utility of the data. We validate our model against experimental car trajectory data gathered in the Italian cities of Pisa and Florence.

The rest of the paper is organised as follows. In §2, §3 and §4, we propose our empirical risk model with a running example based on k-anonymous sequence data the inadequacy of worst-case risk evaluation. We validate our empirical model by tests on real world trajectory data in §5 followed by the state-of-the-art related to the information privacy and its measurements in §6 before concluding the paper in §7.

2 From Theoretical Guarantees to an Empirical Risk Model

2.1 Preliminaries: Trajectory Data

A trajectory dataset is a collection of trajectories $\mathcal{D}_T = \{t_1, t_2, \ldots, t_m\}$. A trajectory $t = \langle x_1, y_1, ts_1\rangle, \ldots, \langle x_n, y_n, ts_n\rangle$, is a sequence of spatio-temporal points, i.e., triples $\langle x_i, y_i, ts_i\rangle$, where (x_i, y_i) are points in $\mathbf{R}^2$, i.e., spatial coordinates, and ts_i $(i = 1 \ldots n)$ denotes a timestamp such that $\forall 1 < i < n\ ts_i < ts_{i+1}$. Intuitively, each triple $\langle x_i, y_i, ts_i\rangle$ indicates that the object is in the position (x_i, y_i) at time ts_i. A trajectory $t' = \langle x'_1, y'_1, ts'_1\rangle, \ldots, \langle x'_m, y'_m, ts'_m\rangle$ is a *sub-trajectory* of t $(t' \preceq t)$ if there exist integers $1 < i_1 < \ldots < i_m \leq n$ such that $\forall 1 \leq j \leq m\ \langle x'_j, y'_j, ts'_j\rangle = \langle x_{i_j}, y_{i_j}, ts_{i_j}\rangle$. We refer to the number of trajectories in $\mathcal{D}_T$ containing a sub-trajectory t' as *support of* t' and denote it by $N_{\mathcal{D}_T}(t') = |\{t \in \mathcal{D}_T | t' \preceq t\}|$.

2.2 The *k*-anonymity Framework for Trajectory Data

A well known method for anonymisation of data before release is k-anonymity [2]. The k-anonymity model was also studied in the context of trajectory data [3–5]. Given an input dataset $\mathcal{D}_T \subseteq T$ of trajectories, the objective of the data release is to transform $\mathcal{D}_T$ into some k-anonymised form $\mathcal{D}'_T$. Without this transformation, the publication of the original data can put at risk the privacy of individuals represented in the data. Indeed, an intruder who gains access to the anonymous dataset may possess some background knowledge allowing him/her to conduct attacks that may enable inferences on the dataset. We refer to any such intruders as an attacker. An attacker may know a sub-trajectory of the trajectory of some specific person and could use this information to infer the complete trajectory of the same person from the released dataset. Given the attacker's background knowledge of partial trajectories, a k-anonymous version has to guarantee that the re-identification probability of the whole trajectory within the released dataset has to be at most $\frac{1}{k}$. If we denote the probability of re-identification of the trajectories as $\Pr(re\_id|t')$ based on the trajectory t' known to the attacker then the theoretical k-anonymity framework implies that $\forall t' \in T, \Pr(re\_id|t') \leq \frac{1}{k}$. The parameter k is a given threshold that reflects the expected level of privacy.

Note that, given a trajectory dataset $\mathcal{D}_T$ and an anonymity threshold $k > 1$ we can have trajectories with a support lower than k ($N_{\mathcal{D}_T}(t') < k$) and trajectories that are frequent at least k times ($N_{\mathcal{D}_T}(t') \geq k$). The first type of trajectories are called k-harmful because their probabilities of re-identification are greater than $\frac{1}{k}$. In [5], the authors show that if a k-anonymisation method returns a dataset $\mathcal{D}'_T$ by guaranteeing that for each k-harmful trajectory t' in the original dataset, $t' \in \mathcal{D}_T$, either $N_{\mathcal{D}'_T}(t') = 0$ or $N_{\mathcal{D}'_T}(t') \geq k$, then we have the property that for any trajectory t known by an attacker (harmful or not), $\Pr(re\_id|t') \leq \frac{1}{k}$.

This fact is easy to verify. Indeed, given a k-anonymous version $\mathcal{D}'_T$ of a trajectory dataset $\mathcal{D}_T$ that satisfies the above condition, and a trajectory t known by the attacker two cases can arise:

- t **is k-harmful in $\mathcal{D}_T$:** In this case we can have either, $N_{\mathcal{D}'_T}(t) = 0$, which implies $\Pr(re\_id|t) = 0$, or $N_{\mathcal{D}'_T}(t') \geq k$, which implies $\Pr(re\_id|t) = \frac{1}{N_{\mathcal{D}'_T}(t)} \leq \frac{1}{k}$.
- t **is not k-harmful in $\mathcal{D}_T$:** In this case we have $N_{\mathcal{D}_T}(t) = F \geq k$ and t can have an arbitrary support in $\mathcal{D}'_T$. If $N_{\mathcal{D}'T}(t) = 0$ or $N_{\mathcal{D}'T}(t) \geq F$, then the same reasoning as in the previous case applies. If $0 < N_{\mathcal{D}'T}(t) < F$ then the probability to re-identify a user to the trajectory t is the probability that that user is present in $\mathcal{D}'T$ times the probability of picking that user in $\mathcal{D}'T$, i.e., $\frac{N_{\mathcal{D}'T}(t)}{F} \times \frac{1}{N_{\mathcal{D}'T}(t)} = \frac{1}{F} \leq \frac{1}{k}$.

The aforementioned mathematical condition that any k-anonymous dataset has to satisfy, is explained as follows. Given the attacker's knowledge of partial trajectories that are k-harmful, i.e., occurring only a few times in the dataset,

they can enable a few specific complete trajectories to be selected, and thus the probability that the sequence linking attack succeeds is very high. Therefore, there must be at least k trajectories in the anonymised dataset matching the attacker's knowledge. Alternatively, there can be no trajectories in the anonymised dataset matching the attacker's knowledge. If the attacker knows a sub-trajectory occurring many times (at least k times) then this means that it is compatible with too many subjects and this reduces the probability of a successful attack. If the partially observed trajectories lead to no match then it is equivalent to saying that the partially observed trajectories could be in any other dataset except from the one under attack, thus leading to an infinitely large search space. This is, somewhat, equivalent to $k \to \infty$. Thus, in this case, $\lim_{k\to\infty} \Pr(re\_id|t') = 0$.

This is the theoretical worst-case guarantee of the probability of re-identification of a k-anonymised dataset. However, we shall see in the following sub-section that this does not give us a complete picture of the probabilities of re-identification.

2.3 Why Is the Theoretical Worst-Case Guarantee Inadequate?

In order to explain the inadequacies of the theoretical worst-case guarantee, let us consider a toy example of trajectories. Let $\mathcal{D}_T$ be the example dataset. We can choose, as an example, a value of $k = 3$ and obtain the 3-anonymous dataset $\mathcal{D}'_T$, for which the theoretical worst-case guarantee is that $\forall t'$, $\Pr(re\_id|t') \leq \frac{1}{3}$.

$$\mathcal{D}_T = \begin{cases} t_1 & : A \to B \to C \to D \to E \to F \\ t_2 & : A \to B \to C \to D \to E \to F \\ t_3 & : A \to B \to C \to D \to E \to F \\ t_4 & : A \to D \to E \to F \\ t_5 & : A \to D \to E \to F \\ t_6 & : A \to D \to E \\ t_7 & : B \to K \to S \\ t_8 & : B \to K \\ t_9 & : B \to K \\ t_{10} & : D \to E \to J \to F \end{cases} \qquad \mathcal{D}'_T = \begin{cases} t'_1 & : A \to B \\ t'_2 & : A \to B \\ t'_3 & : A \to B \\ t'_4 & : A \to D \\ t'_5 & : A \to D \\ t'_6 & : A \to D \\ t'_7 & : A \to D \\ t'_8 & : B \to K \\ t'_9 & : B \to K \\ t'_{10} & : B \to K \end{cases}$$

(a) Original (b) 3-anonymised

Fig. 1. Converting $\mathcal{D}_T$ to k-anonymised $\mathcal{D}'_T$ with $k = 3$

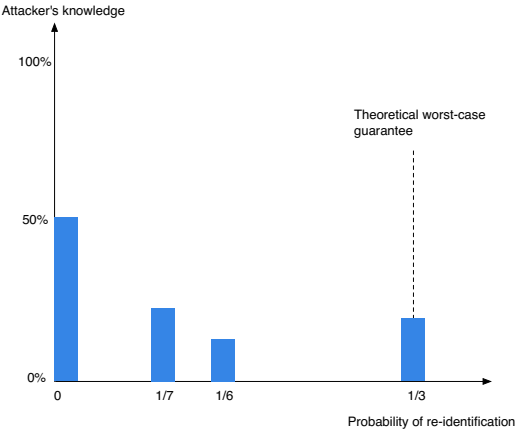
However, we observe from figure 2 that the actual probability of re-identification is often much lower than the theoretical worst-case scenario, but this fact is not demonstrated by the theoretical guarantee.

2.4 Empirical Risk Model for Anonymised Trajectory Data

In the last sub-section, we demonstrated that the theoretical worst-case guarantee does not demonstrate the distribution of attack probabilities. The worst-case

$$
\begin{aligned}
\Pr(re\_id|A) &= 1/7 \\
\Pr(re\_id|B) &= 1/6 \\
\Pr(re\_id|C) &= 0 \\
\Pr(re\_id|D) &= 1/7 \\
\Pr(re\_id|E) &= 0 \\
\Pr(re\_id|F) &= 0 \\
\Pr(re\_id|A \to B) &= 1/3 \\
\Pr(re\_id|A \to D) &= 1/6 \\
\Pr(re\_id|A \to K) &= 0 \\
&\dots \\
\Pr(re\_id|A \to D \to E) &= 0 \\
&\dots \\
\Pr(re\_id|A \to D \to E \to F) &= 0
\end{aligned}
$$

(a) Re-identification probabilities



(b) Probability density distribution

Fig. 2. Probability distribution of re-identification

scenario also does not illustrate the fact that a large majority of the attacks have far lower probabilities of success than the worst-case guarantee. Thus, we propose an empirical risk model for anonymised sequence data. If t' represents attacker's knowledge; $h = |t'|$ denotes the number of observations in the attacker's knowledge then the intent is to approximate a probability density and a cumulative distribution of $\Pr(re\_id|t')$ for each value of h. This can be achieved by iterating over every value of $h = 1, \dots, M$ where M is the length of the longest trajectory in $\mathcal{D}_T$. For each value of h, we consider all the sub-trajectories $t' \in \mathcal{D}_T$ of length h and compute the probability of re-identification $\Pr(re\_id|t')$ as described in Algorithm 1. In particular, for each value of h a further iteration can be run over each value of t' of length h, in which we compute $N_{\mathcal{D}'_T}(t')$, $N_{\mathcal{D}_T}(t')$ and the probability of re-identification by following the reasoning described in Section 2.2 for the computation of this probability. Algorithm 1 presents the pseudocode of the attack simulation.

The advantages of this approach is that this model supports arguments such as: (a) "98% of the attacks have at most 10^{-5} probability of success"; and (b) "only 0.001% of the attacks have a probability close to $\frac{1}{k}$". The disadvantages of this model are: (a) a separate distribution plot is necessary for each value of h; and (b) the probability of re-identification increases with the increase in h. The illustration in Figure 3 demonstrates the aforementioned advantages and disadvantages of the risk model.

For the simulation of the attack we need to select a set of trajectories BK_T from the original dataset of trajectories. The optimal solution would be to take the all possible sub-trajectories in the original dataset and compute the probability of re-identification. Since the set of attack trajectories can be quite large, in order to avoid a combinatorial explosion, two strategies can be adopted.

Algorithm 1. Attack Simulation

Require: The k-anonymised dataset $\mathcal{D}'_T$, the original dataset $\mathcal{D}_T$, the set of trajectories for the attacks BK_T and anonymity threshold k.

for $h = 1, \ldots, M$ where M is the length of the longest trajectory in $\mathcal{D}_T$ **do**
  for t' of length h in BK_T **do**
    $N(t')_{\mathcal{D}_T} \leftarrow |\{t \in \mathcal{D}_T | t' \preceq t\}|$.
    $N(t')_{\mathcal{D}'_T} \leftarrow |\{t \in \mathcal{D}'_T | t' \preceq t\}|$.
    if $N(t')_{\mathcal{D}_T} \geq k$ *and* $N(t')_{\mathcal{D'}_T} \leq N(t')_{\mathcal{D}_T}$ **then**
      $\Pr(re\_id|t') \leftarrow 1/N(t')_{\mathcal{D}_T}$.
    else
      $\Pr(re\_id|t') \leftarrow 1/N(t')_{\mathcal{D'}_T}$.
    end if
  end for
end for
return Cumulative Distribution of $\Pr(re\_id|t')$ for all h.

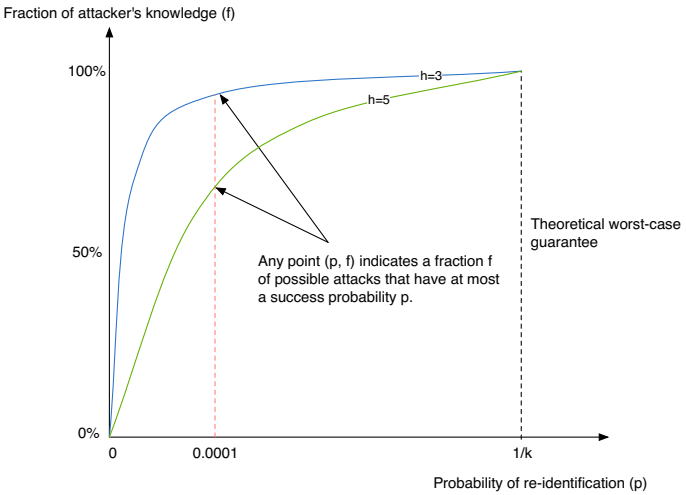


Fig. 3. Representative cumulative density distribution for attacks in the toy example

First, we can extract from the original dataset of trajectories a random subset of trajectories that we can use as background knowledge for the attacks to estimate the distributions. Secondly, we can use a prefix tree to represent in a compact way the original dataset and then, by incrementally visiting the tree we can enumerate all the distinct sequences for using them as an adversary's background knowledge.

Risk versus Cost. One of the most important open problems that makes the communication between the experts in law and in computer science hard is how to evaluate whether an individual is identifiable or not, i.e., the evaluation of privacy risks for an individual. Usually, the main legal references to this problem

suggests to measure the difficulty in re-identifying the data subject in terms of "time and manpower". This definition is surely suitable for traditional computer security problems. As an example, we can measure the difficulty to decrypt a message without the proper key in terms of how much time we need to try all possible keys i.e., the time and resources required by the so-called *brute force* attack. In the field of privacy the computer science literature shows that the key factor affecting the difficulty to re-identify an anonymous data is the *background knowledge* available to the adversary. Thus, we should consider the difficulty to acquire the knowledge that enables the attack to infer some sensitive information. If we are able to measure the *cost* of the acquisition of the background knowledge then we can provide a single risk indicator that takes into consideration both the probability of success of an attack and its cost. Combining the two factors and providing one single value could help the communication of a specific privacy risk in the legal language.

We propose three methods for measuring the cost of an attack and a way to combine it with the probability of re-identification. We also propose to normalise the probability of re-identification $\Pr(re\_id|t')$ with the cost of gaining the knowledge of t' by the attacker. The longer the t', the higher the cost to acquire such knowledge. Thus, $\Pr(t') = \Pr(re\_id|t')/C(t')$ where $C(t')$ is the cost function proportional to the length of t'. We can then estimate the distribution of $\Pr(t')$ over all t' to obtain a unique combined measurement of risk over all possible attacks.

The cost function $C(t')$ can be derived from various alternatives. (1) One option would be to use a sub-linear cost function akin to that incurred in machine-operated sensing. The initial costs of setting up the sensing equipment are high but subsequent observations are cheaper and cheaper. Thus, $C(t') = 1 + log(|t'|)$ is a good approximation. (2) Another option is a linear cost where a spying service is paid a fixed fee per observation, leading to $C(t') = \alpha|t'|$. (3) A third alternative is a super-linear cost where the attacker directly invests time and resources to sensing, thus making the cost function $C(t') = e^{-\beta|t'|}$.

3 Data Utility Measures: Coverage and Precision

Alongside the risk versus cost estimations, it is also important to identify the usability of the anonymised data and show the relation between usability and privacy risk. In this context, we introduce two usability measures: *coverage* and *precision*. While a trajectory can consist of multiple hops, it can also be seen as a chain of smaller trajectories, each of which just contain the start point (the origin) and the end point (the destination). We call these smaller trajectories as $ODpairs$ (or, origin-destination pairs). Given a k-anonymisation function that maps $\mathcal{D}_T$ into $\mathcal{D}'_T$, we define *coverage*:

$$coverage = |ODpairs(\mathcal{D}_T) \cap ODpairs(\mathcal{D}'_T)|/|\mathcal{D}_T| \quad (1)$$

and *precision* as:

$$precision = |ODpairs(\mathcal{D}_T) \cap ODpairs(\mathcal{D}'_T)|/|ODpairs(\mathcal{D}'_T| \quad (2)$$

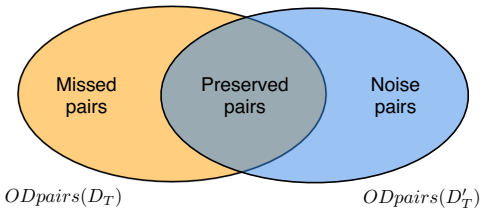


Fig. 4. Diagrammatic representation of coverage and precision

The coverage versus risk for a given risk threshold can be estimated as follows. Given an anonymised dataset $\mathcal{D}'_T$ and a specified probability threshold p where $0 \leq p \leq \frac{1}{k}$, all trips t containing attack based on t' with $\Pr(re\_id|t') > p$ can be retrieved as:

$$RiskyTrips(p) = \{t \in \mathcal{D}'_T | \exists\, t' : \Pr(re\_id|t') > p \text{ and } t' < t\} \tag{3}$$

Thus, the coverage of the dataset $\mathcal{D}_T$ with respect to the risk threshold p is defined as follows

$$coverage = |ODpairs(\mathcal{D}'_T) \setminus ODpairs(RiskyTrips(p))|/|\mathcal{D}'_T| \tag{4}$$

The characteristics of the mobility data that are preserved with high fidelity if we measure a high coverage rate are: (a) presence (of users in locations), (b) flows (i.e., the number of trips between any origin-destination pair), and (c) overall distance travelled in *all* trips.

The characteristics that are not necessarily preserved include the properties of sequences of individual trips, e.g., distribution of trip length and routine trips.

4 Privacy-by-Design for Data-Driven Services

The *privacy-by-design* model for privacy and data protection has been recognised in legislation in the last few years years. Privacy-by-design is an approach to protect privacy by inscribing it into the design specifications of information technologies, accountable business practices, and networked infrastructures, from the very start. It was developed by Ontario's Information and Privacy Commissioner, Dr. Ann Cavoukian, in the 1990s.

Privacy officials in Europe and the United States are embracing this paradigm as never before. In Europe, in the comprehensive reform of the data protection rules, proposed on January 25, 2012 by the EC, the new data protection legal framework introduces, with respect to the Directive 95/46/EC, the reference to data protection by design and by default (Article 23 of the Proposal for a Regulation and Article 19 of the Proposal for a Directive). These articles compel

the controller to "*implement appropriate technical and organizational measures and procedures in such a way that the processing will meet the requirements of this Regulation and ensure the protection of the rights of the data subject.*" and to "*implement mechanisms for ensuring that, by default, only those personal data are processed which are necessary for each specific purpose of the processing ...*".

In [6] Monreale at al. define a methodology for applying the *privacy-by-design* principle in the context of data analytics. This work states that one of the most important points to take into consideration for releasing technological frameworks that offer *by-design* the privacy protection is the trade-off between privacy guarantees and the data quality.

The model presented in above sections provides a methodology for the evaluation of this trade-off. Indeed, the availability of this model allows us to define a methodology of risk evaluation of datasets that have to be used for specific services; and this methodology allows us to establish a well-defined relation between the risks of re-identification of any individual represented in the data and the usability of the anonymous data for the specified services.

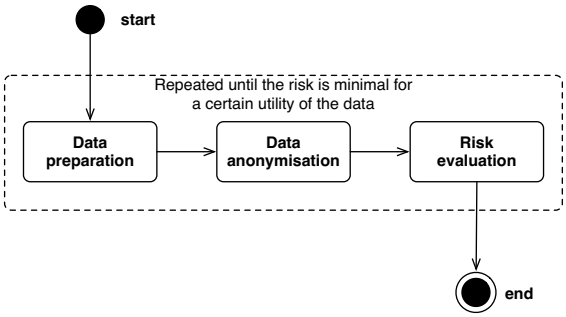


Fig. 5. Refining privacy and risk until the risk is minimal for a certain utility of the data

In Figure 5 we depict this methodology that is composed of three phases: (a) data preparation, (b) data anonymisation, and (c) risk evaluation.

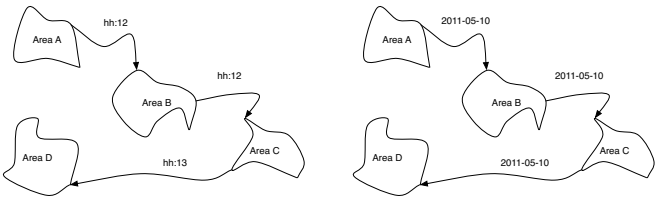
The cycle, illustrated in figure 5 needs to be repeated with respect to the different dimensions (e.g., spatial and temporal granularity, refresh window) obtaining a collection of anonymised datasets $\mathcal{D}_T'^i$ with associated risks R^i. Given a class of services that are to be facilitated by the published data, the anonymised dataset $\mathcal{D}_T'^i$ will be chosen for which the associated risk R^i is *minimal* with acceptable utility of the published data.

5 Experimental Validation

In this section we present a detailed evaluation of the proposed risk model by using real-world mobility data. We used a large dataset of real GPS traces from

vehicles, collected during the period between May 1 and May 31, 2011, donated by an Italian company called *OctoTelematics*. The dataset contains the GPS traces collected in the geographical areas around Pisa and Florence, in central Italy, for around 18,800 vehicles making up around 46,000 trips. For our simulations, we extracted from the whole dataset the data on May 10, 2011 that contained 8,330 participating users and 15,345 trajectories.

To begin with, the privacy-sensitive locations captured through GPS readings were obfuscated using Voronoi tessellation [7]. The trajectory data containing tessellated locations (signifying vertices in a trajectory graph) was further subjected to k-anonymisation for $k = 3$, $k = 5$, and $k = 10$ by using the method proposed in [5]. Before applying this anonymisation, we subjected the sequence data to two further steps: generalisation of temporal information and transformation of trajectories. The first step – generalisation of the temporal information associated with each location visited by the user – consisted of two levels of generalisations: one that contains sequences of Voronoi areas where the time associated with each location is generalized at an hour-level (*hour-level data*) and another one where the time is at a day-level (*day-level data*). Figure 6 illustrates an example of a user trajectory observed at an hour-level and at the day-level.



(a) Hour-level (b) Day-level

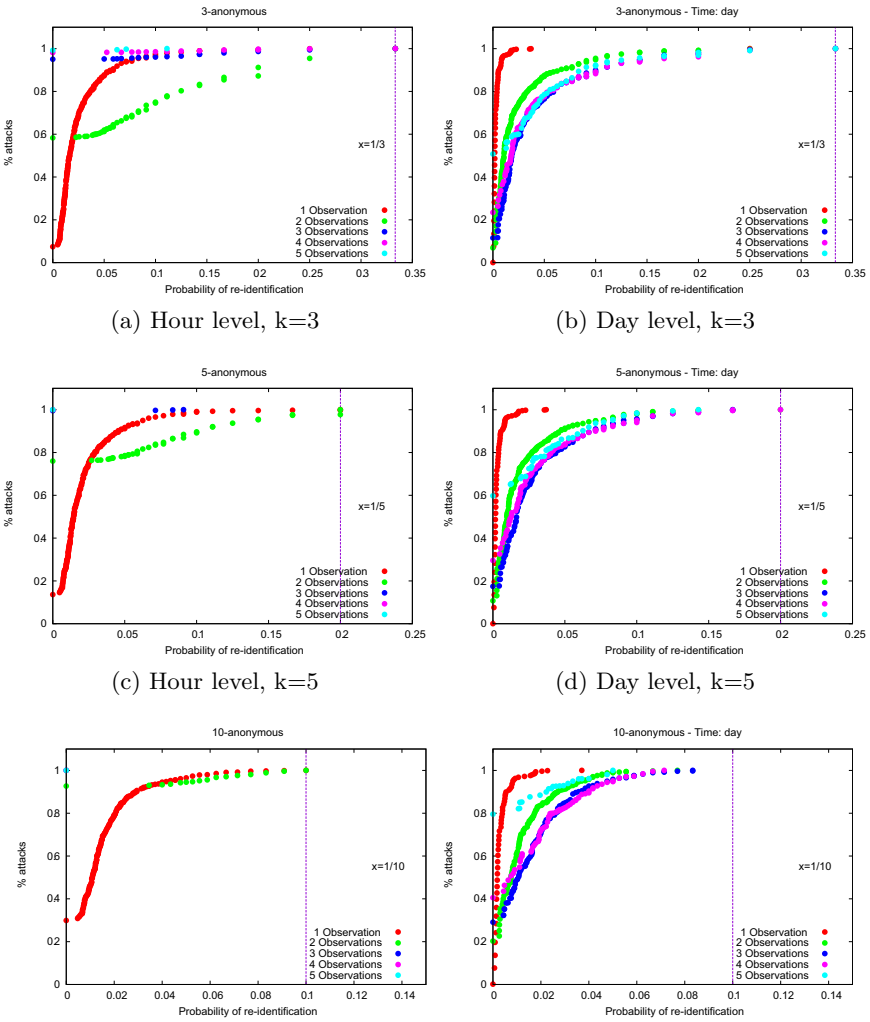
Fig. 6. An example of user trajectory through the different tessellated areas observed at an hour-level and at a day-level

The second step consisted of the transformation of the generalised trajectories into sequences of *ODpairs*; in particular, we divided the whole user sequence into smaller sequences and for each small sequence we extracted its origin and its destination.

In our evaluation we performed two different analyses. First, we applied our risk model showing the evaluation of the privacy risks obtained from the two anonymised datasets described above, and then, we measured the data utility in terms of *precision* and *coverage* described in §3.

5.1 Risk Analysis

In order to evaluate the privacy risks on the two anonymised sequence datasets we applied the methodology described in §2.4. Therefore, we estimated the cumulative distribution of the probability of re-identification for each value of $h = |t'|$, which



(a) Hour level, k=3 (b) Day level, k=3

(c) Hour level, k=5 (d) Day level, k=5

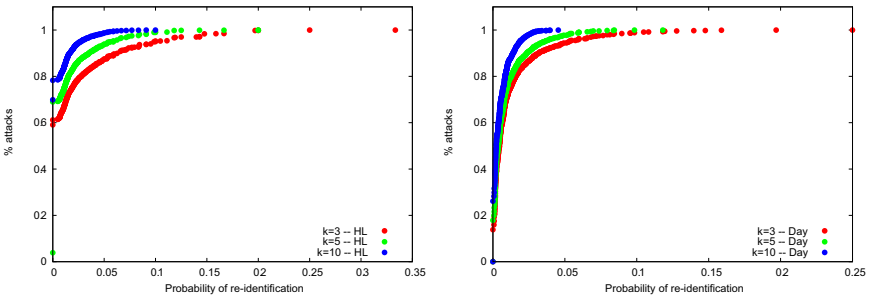
(e) Hour level, k=10 (f) Day level, k=10

Fig. 7. Cumulative distribution of the re-identification probability

denotes the number of observations in the attacker's knowledge. We simulated a set of attacks by randomly selecting from the original database a subset of trajectories and using them as background knowledge. In particular, in our experiment for each h, we drew from the original database, $10,000$ sub-sequences with length h. We considered $h = 1, \ldots, 5$ because the longest sequence in the original data has length 5. Figure 7 shows the results obtained with this attack simulation. The first column of images contains the plots related to the cumulative distributions related to the *hour-level dataset* while the second column contains the results obtained from the *day-level dataset*.

Our analyses highlight that the empirical protection guaranteed by the algorithm of anonymisation is much higher than the theoretical protection. Only few attacks have a protection very close to $\frac{1}{k}$. We observe as an example that when the *day-level dataset* is anonymised with $k = 5$ our empirical risk analysis shows that 90% of the attacks have at most a risk of re-identification of $\frac{1}{10}$. The findings are similar in the other anonymised datasets. Moreover, we note that when the number of observations increases too much the probability of re-identification becomes very low and often zero because these sequences are infrequent in the original database. These long sequences no longer exist in the published database since the process of anonymisation tends to eliminate the outliers (i.e., sequences with a very low frequency). This effect is more evident in the case of the *hour-level data*.

We also estimated the cumulative distribution of the re-identification probability normalised with the cost of obtaining the background knowledge (see Section 2.4). Figure 8 depicts the cumulative distribution of our single risk indicator obtained considering a sub-linear cost for the acquisition of the attacker's knowledge. We observe that if we assign a cost to the attack then the protection guaranteed is higher; thus allowing us to express in a very simple way the risk to the individuals if the whole dataset is published. Indeed, as an example figure 8(b) shows that when the *day-level dataset* is anonymised with $k = 5$ the probability of re-identification considering also the attack cost is at most about 0.025 ($\frac{1}{20}$) for 90% of the attacks.



(a) Hour-level (b) Day-level

Fig. 8. Risk analysis with Background Knowledge Cost

5.2 Data Quality Evaluation

In our experiment we also evaluated the data quality by measuring the *precision* and the *coverage* defined above. Table 1(a) shows these two measures for the k-anonymous versions of the *hour-level dataset* while table 1(b) shows the same information for the *day-level dataset*.

Table 1. Precision versus coverage of the k-anonymised experimental data

(a) Time: hour-level

k	Precision	Coverage
3	1.00	0.27
5	1.00	0.15
10	1.00	0.04

(b) Time: day-level

k	Precision	Coverage
3	0.98	0.87
5	0.97	0.83
10	0.96	0.72

As expected the anonymisation preserves very well the precision of the *ODpairs*; this means that the data transformation does not introduce noise, while it tends to suppress some *ODpairs* and this affects the data coverage. This behaviour is more evident in the *hour-level dataset*. Lastly, we also analysed how the *coverage* changes by varying the risk in the dataset. Figure 9 outlines the results. In line with our expectations, the *coverage* increases with the privacy risk. However, we observe that with a risk of re-identification of 0.1 we can have a *coverage* of about 90% in the *hour-level dataset* anonymized with $k = 5$. The situation improves a lot in the *day-level dataset*. Thus, this is a good tool for managing the trade-off between privacy and data utility.

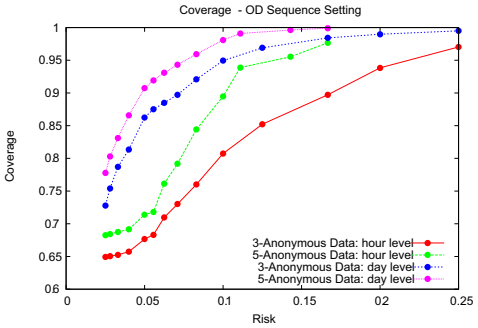


Fig. 9. Coverage with respect to privacy risk

6 The State-of-the-Art

Research in information privacy consists of a vast corpus of multi-disciplinary work combining results from the fields of psychology, law, computer science amongst others. Privacy in information systems has been often governed by a set of fair practices that help organisations manage users' information in responsible manners [8]. There often exists a disconnection between the interpretation of privacy needs from the perspective of the user and the prescribed privacy preserving mechanisms offered by devices and systems. Hong et al. [9] presented privacy risk models for ubiquitious systems in order to convert privacy from an abstract concept into specific issues relating to concrete applications. Kosa et

al. [10], in an attempt to represent and measure privacy, presented an interesting finite state machine based representation of at most nine privacy states for any individual in a computer system. A recent work by Kiyomoto et al. [11] proposes a privacy policy management mechanism whereby a match is made between user's personal privacy requirements and organisational privacy policies. PrivAware [12] was presented as a tool to detect and report unintended loss of privacy in a social network. Krishnamurthy et al. [13] measured the loss of privacy and the impact of privacy protection in web browsing both at a browser level as well as a HTTP proxy level. Tao et al. [14] put forward a model for quality of service (QoS) for web services that quantified users' privacy risks in order to make the service selection process manageable. Banescu et al. [15] came up with a privacy compliance technique for detecting and measuring the severity of privacy infringements.

With richer user data available for data mining, work in privacy preserving data mining and privacy preserving data publishing have gained momentum in the recent years. Techniques such as adding random noise and perturbing outputs while preserving certain statistical aggregates are often used [16–19]. Some notable work data anonymisation work include k-anonymity [2], l-diversity [20], t-closeness [21], p-sensitive k-anonymity [22], (α, k)-anonymity [23] and ϵ-differential privacy [24]. The k-anonymity model has been also studied and adapted in the context of movements data in different works: [3] exploits the inherent uncertainty of the moving object's whereabouts; [4] proposes a technique based on *suppression* of the dangerous observations from each trajectory; and [5] proposes a data-driven spatial generalization approach to achieve k-anonymity. A critique by Domingo-Ferrer and Torra [25] analyses the drawbacks of some of those anonymisation methods. The trade-off between the privacy guarantees of anonymisation models and the data mining utility have been considered by authors in [26, 27]. Sramka et al. [28] compared data utility versus privacy based on two well known privacy models – k-anonymity and ϵ-differential privacy.

Our proposed empirical risk model draws inspirations from the existing research in the privacy preserving data publishing domain. We envision that our model provides a clear understanding of privacy (or the lack of it) in released but anonymised data with relation to risk, privacy, cost of attacks and data utility.

7 Conclusions

In this paper we have proposed an empirical risk model that provides a complete and realistic view on the privacy risks, which can be derived from the release of trajectory data. Our model is able to empirically evaluate the real risks of re-identification taking into account also the cost of any attack on privacy as well as the relation between the risk and the utility of the data. With legislature becoming increasingly detailed about data protection, it is essential to be able to communicate well how privacy, risk and cost of attacks are associated when applying mathematical models for privacy preserving data release. We have presented promising evaluations of our model for the well-known k-anonymisation

applied to real trajectory data from the Italian cities of Pisa and Florence. In the future, we plan to evaluate our model with different types of real data of sequential nature. Furthermore, we intend to investigate risk models suitable for other types of data.

References

1. Samarati, P.: Protecting respondents identities in microdata release. IEEE Transactions on Knowledge and Data Engineering 13(6), 1010–1027 (2001)
2. Sweeney, L.: k-anonymity: A model for protecting privacy. International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems 10(05), 557–570 (2002)
3. Abul, O., Bonchi, F., Nanni, M.: Never walk alone: Uncertainty for anonymity in moving objects databases. In: The 24th IEEE International Conference on Data Engineering (ICDE), pp. 376–385 (2008)
4. Terrovitis, M., Mamoulis, N.: Privacy preservation in the publication of trajectories. In: MDM, pp. 65–72 (2008)
5. Monreale, A., Andrienko, G.L., Andrienko, N.V., Giannotti, F., Pedreschi, D., Rinzivillo, S., Wrobel, S.: Movement data anonymity through generalization. TDP 3(2), 91–121 (2010)
6. Monreale, A., Pedreschi, D., Pensa, R., Pinelli, F.: Anonymity preserving sequential pattern mining. Artificial Intelligence and Law (to appear, 2014)
7. Voronoï, G.: Nouvelles applications des paramètres continus à la théorie des formes quadratiques. deuxième mémoire. recherches sur les parallélloèdres primitifs. Journal für die Reine und Angewandte Mathematik 134, 198–287 (1908)
8. Westin, A.F.: Privacy and freedom. Washington and Lee Law Review 25(1), 166 (1968)
9. Hong, J.I., Ng, J.D., Lederer, S., Landay, J.A.: Privacy risk models for designing privacy-sensitive ubiquitous computing systems. In: The 5th Conference on Designing Interactive Systems: Processes, Practices, Methods, and Techniques, pp. 91–100. ACM (2004)
10. Kosa, T.A., EI-Khatib, K., Marsh, S.: Measuring privacy. Journal of Internet Services and Information Security (JISIS) 1(4), 60–73 (2011)
11. Kiyomoto, S., Nakamura, T., Takasaki, H., Watanabe, R., Miyake, Y.: PPM: Privacy policy manager for personalized services. In: Cuzzocrea, A., Kittl, C., Simos, D.E., Weippl, E., Xu, L. (eds.) CD-ARES Workshops 2013. LNCS, vol. 8128, pp. 377–392. Springer, Heidelberg (2013)
12. Becker, J.L., Chen, H.: Measuring privacy risk in online social networks (2009)
13. Krishnamurthy, B., Malandrino, D., Wills, C.E.: Measuring privacy loss and the impact of privacy protection in web browsing. In: The 3rd Symposium on Usable Privacy and Security, pp. 52–63. ACM (2007)
14. Yu, T., Zhang, Y., Lin, K.-J.: Modeling and measuring privacy risks in qos web services. In: The 3rd IEEE Conference on E-Commerce Technology and the 8th IEEE International Conference on and Enterprise Computing, E-Commerce, and E-Services, p. 4. IEEE (2006)
15. Banescu, S., Petković, M., Zannone, N.: Measuring privacy compliance using fitness metrics. In: Barros, A., Gal, A., Kindler, E. (eds.) BPM 2012. LNCS, vol. 7481, pp. 114–119. Springer, Heidelberg (2012)
16. Agrawal, R., Srikant, R.: Privacy-preserving data mining. ACM SIGMOD Record 29(2), 439–450 (2000)

17. Dinur, I., Nissim, K.: Revealing information while preserving privacy. In: The 22nd ACM SIGMOD-SIGACT-SIGART Symposium on Principles of Database Systems, pp. 202–210 (2003)
18. Evfimievski, A., Gehrke, J., Srikant, R.: Limiting privacy breaches in privacy preserving data mining. In: The 22nd ACM SIGMOD-SIGACT-SIGART Symposium on Principles of Database Systems, pp. 211–222 (2003)
19. Blum, A., Dwork, C., McSherry, F., Nissim, K.: Practical privacy: the sulq framework. In: The 24th ACM SIGMOD-SIGACT-SIGART Symposium on Principles of Database Systems, pp. 128–138 (2005)
20. Machanavajjhala, A., Kifer, D., Gehrke, J., Venkitasubramaniam, M.: l-diversity: Privacy beyond k-anonymity. ACM Transactions on Knowledge Discovery from Data (TKDD) 1(1), 3 (2007)
21. Li, N., Li, T., Venkatasubramanian, S.: t-closeness: Privacy beyond k-anonymity and l-diversity. In: The 23rd IEEE International Conference on Data Engineering (ICDE), pp. 106–115 (2007)
22. Truta, T.M., Vinay, B.: Privacy protection: p-sensitive k-anonymity property. In: The 22nd International Conference on Data Engineering Workshops, p. 94. IEEE (2006)
23. Wong, R.C.-W., Li, J., Fu, A.W.-C., Wang, K. (α, k)-anonymity: an enhanced k-anonymity model for privacy preserving data publishing. In: The 12th ACM SIGKDD international Conference on Knowledge Discovery and Data Mining, pp. 754–759 (2006)
24. Dwork, C.: Differential privacy. In: Bugliesi, M., Preneel, B., Sassone, V., Wegener, I. (eds.) ICALP 2006. LNCS, vol. 4052, pp. 1–12. Springer, Heidelberg (2006)
25. Domingo-Ferrer, J., Torra, V.: A critique of k-anonymity and some of its enhancements. In: The 3rd International Conference on Availability, Reliability and Security (ARES), pp. 990–993. IEEE (2008)
26. Rastogi, V., Suciu, D., Hong, S.: The boundary between privacy and utility in data publishing. In: The 33rd International Conference on Very Large Databases, pp. 531–542. VLDB Endowment (2007)
27. Brickell, J., Shmatikov, V.: The cost of privacy: destruction of data-mining utility in anonymized data publishing. In: The 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 70–78 (2008)
28. Sramka, M., Safavi-Naini, R., Denzinger, J., Askari, M.: A practice-oriented framework for measuring privacy and utility in data sanitization systems. In: EDBT/ICDT Workshops, p. 27. ACM (2010)