Basu *et al. Journal of Trust Management* (2015) 2:9
DOI 10.1186/s40493-015-0020-6

Journal of Trust Management
a SpringerOpen Journal

RESEARCH Open Access

A risk model for privacy in trajectory data

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

*Correspondence: basu@kddilabs.jp
†Equal Contributors
[1] Information Security Group, KDDI R&D Laboratories, 2-1-15 Ohara, Fujimino, Saitama 356-8502, Japan
Full list of author information is available at the end of the article

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 and then, we show how the empirical evaluation of the privacy risk has a different trend in synthetic data describing random movements.

Keywords: Privacy, Risk, Utility, Model, Anonymisation, Sequential data

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 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. In this paper, we focus mostly on a specific PPDP method – *k*-anonymity.

A well-known anonymisation model typically used for PPDP is the *k*-anonymity model [1, 2]. It states that in the *worst-case*, where the attacker has knowledge of the

Springer

© 2015 Basu et al. **Open Access** This article is distributed under the terms of the Creative Commons Attribution 4.0 International License (http://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made.

full set of quasi-identifiers chosen at the time of data release, the attacker will find either zero or at least k (and no less) users in a k-anonymised dataset with the same values of the quasi-identifier attributes. Thus, the re-identification probability for any single user, in the worst-case, is equal to $1/k$. In general, a quasi-identifier is a piece of information (e.g., age of a person), which by itself is not a unique identifier, but can be combined with other quasi-identifiers to identify a unique entity. 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. Our experiments highlight that the empirical evaluation of the protection guaranteed by the algorithm of anonymisation on real-world data is much higher than the theoretical protection. This happens because in real life the user movements are influenced by a lot of external constraints such as the existence of one or more streets, the direction of a specific street, the traffic intensity, and so on. As an example, if in a specific area we have only one street for going from the place A to the place B all people will cross the same street and will produce similar trajectory data. This helps the result of the empirical privacy protection evaluation. To prove this fact we also generate some synthetic movement data without using any of those constraints and as expected we found that in this kind of data the empirical privacy protection is lower than that one on real data and the data quality decreases. We also discuss how the empirical risk model can be adapted to semantic trajectories anonymized by considering the privacy model c-safety [3].

This paper is an extension of our earlier work presented at an international conference in 2014, the proceedings of which were published by Springer. In particular, this paper describes an extension of our risk model for k-anonymity applied to c-safety, which is a framework for anonymisation of semantic trajectories. We have also described the risk analysis on null models.

The rest of the paper is organised as follows. In Sections “From theoretical guarantees to an empirical risk model for k-anonymity”, “Data utility measures: coverage and precision” and “Privacy-by-design for data-driven services”, we propose our empirical risk model with a running example based on k-anonymous trajectory data the inadequacy of worst-case risk evaluation. We describe an extension of this risk model to c-safety in Section “An empirical risk model for c-safety”. We validate our empirical model by tests on real world trajectory data and synthetic data in Section “Experimental validation” followed by the state-of-the-art related to the information privacy and its measurements in Section “The state-of-the-art” before concluding the paper in Section “Conclusions”.

From theoretical guarantees to an empirical risk model for *k*-anonymity

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 $N_{\mathcal{D}_T}(t') = \big|\{t \in \mathcal{D}_T | t' \preceq t\}\big|$.

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 [4–6]. 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 [6], 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 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.

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 as shown in Fig. 1. 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 $\forall t'$, $\Pr(re\_id|t') \leq \frac{1}{3}$.

Figure 2 illustrates the probability that a given observed trajectory (i.e., attacker's knowledge) can be uniquely identified in the anonymised dataset, while Fig. 3

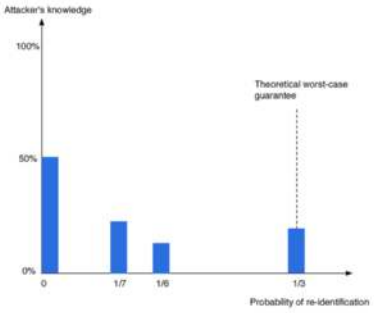
$$\mathcal{D}_T = \begin{cases} t_1 & : & A \to B \to C \to D \to E \to F \\ t_2 & : & A \to B \to C \to D \to E \to F \\ t_3 & : & A \to B \to C \to D \to E \to F \\ t_4 & : & A \to D \to E \to F \\ t_5 & : & A \to D \to E \to F \\ t_6 & : & A \to D \to E \\ t_7 & : & B \to K \to S \\ t_8 & : & B \to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

(a) Original (b) 3-anonymised

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

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

(a) Re-identification probabilities



(b) Probability density distribution

Fig. 2 Probability distribution of re-identification. The actual probabilities of re-identification and the corresponding probability distribution

shows the cumulative distributions of probabilities with h denoting the number of observations in the attacker's knowledge. We notice in Fig. 2 and Fig. 3 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.

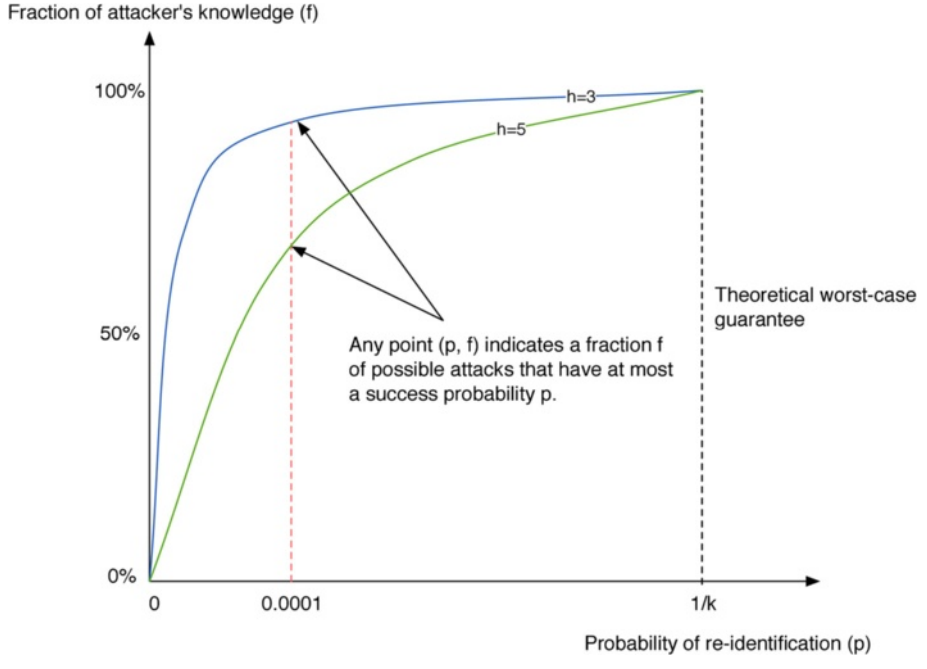


Fig. 3 Representative cumulative density distribution for attacks in the toy example. The cumulative distribution corresponding to the probability distribution in Fig. 2

Empirical risk model for anonymised trajectory data

In the last sub-section, we described that the theoretical worst-case guarantee does not demonstrate the distribution of attack probabilities. The worst-case 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 trajectory 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 $h = 1, \ldots, M$ 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 "The k-anonymity framework for trajectory data" for the computation of this probability. Algorithm 1 presents the pseudocode of the attack simulation.

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.

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 Fig. 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. 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. In particular, for each each trajectory length value h we extract a random subset of trajectories BK_T^h and then, the union of all BK_T^h represents the global background knowledge BK_T used in the attack simulation.

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 grow in some exponential fashion, such as $C(t') = e^{-\beta|t'|}$.

These cost models are not exhaustive. There can be other factors, beyond the scope of this paper, that can have perceptible effects on the costs of attacks.

An empirical risk model for *c*-safety

In this section we discuss how it is possible to adapt the empirical risk model, presented above, to *semantic trajectories* anonymized by considering the privacy model *c-safety* introduced in [3].

Semantic trajectories

In the previous sections we considered a trajectory as the spatio-temporal evolution of the position of a moving entity that is represented as a discrete sequence of points. An interpolation function between two consecutive points approximates the movements between two sample points. Recently, in [7] a new concept of trajectory has been introduced for reasoning on trajectories from a semantic point of view, called *semantic trajectory*, that based on the notion of stops and moves. Stops are the *important parts* of a trajectory where the moving object has stayed for a minimal amount of time. They correspond to the set of x, y, t points of a trajectory which are important from an application point of view. Stops correspond to places and can be different *types* of geographic locations as hotels, restaurants, museums, etc; or different *instances* of geographic places, like Ibis Hotel, Louvre Museum, and so on. Moves are the sub-trajectories describing the movements between two consecutive stops. Based on the concept of *stops* and *moves* the user can enrich trajectories with semantic information according to the application domain [8].

Semantic Trajectory Given a set of important places $\mathcal{I}$, a *semantic trajectory* $T = p_1, p_2, \ldots, p_n$ with $p_i \in \mathcal{I}$ is a temporally ordered sequence of important places, that the moving object has visited.

Figure 4 (2) illustrates the concept of semantic trajectory for the trajectory shown in Fig. 4 (1). In the semantic trajectory the moving object first was at home (stop 1), then he went to work (stop 2), later he went to a shopping center (stop 3), and finally the moving object went to the gym (stop 4).

The important parts of the trajectories (stops) are application dependent, and are not known a priori, therefore they have to be computed. Different methods have been proposed for computing important parts of trajectories [9–11].

The *C-safety* model for semantic trajectories

In [3] authors provide a framework that, given a dataset of semantic trajectories, generates an *anonymous semantic trajectory dataset*. This new dataset guarantees that

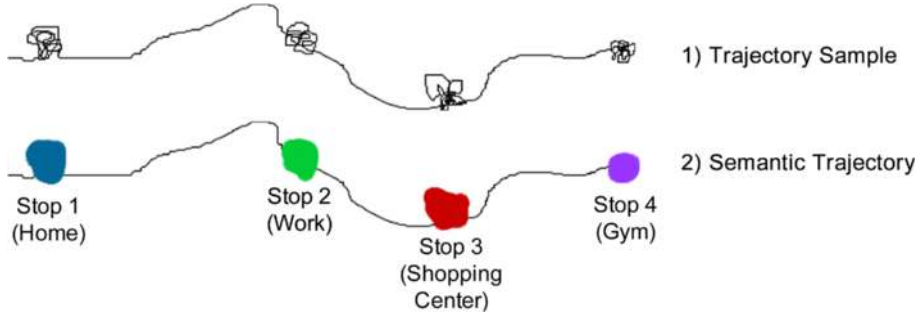


Fig. 4 An example trajectory and its semantic equivalence. Example of Semantic Trajectory

it is not possible to infer the identity of a user and the visited sensitive places with a probability greater than a fixed threshold, set by the data owner. The method is based on the generalization, driven by a place taxonomy, of the places visited by a user. A taxonomy of places of interest represents the semantic hierarchy of geographical places of interest. The set of *stop places* obtained from the computation of semantic trajectories are the *leaves* of taxonomy. In general, each concept in the taxonomy describes the semantic categories of the geographical objects of interest for a given application domain. For example, we have that *Restaurant "Da Mario"* is a *kind of* Restaurant which is a *kind of* Entertainment. Figure 5 depicts an example of the taxonomy of places of interest in a city. Here, all the places represented by red nodes are sensitive locations. Indeed, to avoid the identification of sensitive places visited by a user, the taxonomy specifies which places are *sensitive* and which are *non-sensitive*.

A place is considered sensitive when it allows to infer personal information about the person who has stopped there. For example, a stop at an oncology clinic may indicate that the user has some health problem. Other places (such as parks, restaurants, cinemas, etc) are considered as quasi-identifiers. Note that, any non-sensitive place is assumed to be a quasi-identifier. Given a dataset of semantic trajectories D_T and the privacy places taxonomy *Ptax* describing the categories of the geographical objects of interest for an application domain, the goal of the data release is to transform D_T in its anonymous version D'_T by using a method based on the generalization of places driven by the taxonomy.

An attacker may access the dataset D'_T and may know the privacy place taxonomy *Ptax*, the quasi-identifier place sequence S_Q visited by a specific person and could use this information to infer the sensitive places visited by a that person. Given the attacker's background knowledge of quasi-identifier places S_Q and $c-$*safe* version of the semantic trajectories has to guarantee that for each set of sensitive places S the $Pr(S|S_Q) \leq c$ with $c \in [0,1]$. Here, the parameter c is a given threshold that reflects the expected level of privacy.

To guarantee $c-$*safety* the approach proposed in [3] generates by generalization (driven by the taxonomy) groups of m trajectories having the same sequence of quasi-identifier places and guaranteeing that for each sensitive place $Pr(s_i|S_Q) \leq c$.

Empirical risk model for *c*-safe semantic trajectories

As in the case of k-anonymity also in this case actual probability of inferring an exact sensitive place is often much lower than the theoretical worst-case scenario. This

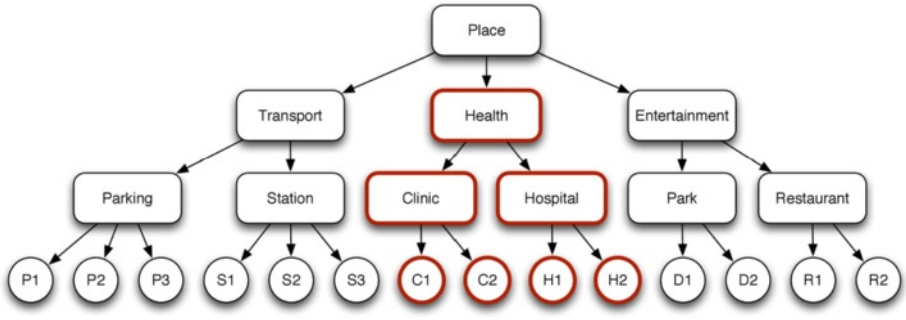


Fig. 5 An example taxonomy of places of interest. Example of Place Taxonomy

is due to the fact that the attacker could knows only a subset of the user quasi-identifier places and so, in the c-safe dataset he could find more than one group with m trajectories matching the known quasi-identifier places. Since, each group guarantees that for each sensitive place the probability of inference is at most c then could happen that the protection becomes higher. In the following example we show this point.

Suppose that the privacy transformation create the following two groups of semantic trajectories generalized by the taxonomy in Fig. 5. The transformed dataset is 0.66-safe, that means that for each sensitive place the disclosure probability is at most equal to 0.66. Computing the disclosure probability for different size and composition of quasi-identifiers S_Q we have that only the sensitive place C_1 we have the worst-case value 0.66, in fact assuming that the attacker knows $S_Q = \textit{Station, Restaurant, Park}$ $Pr(C_1|S_Q) = \frac{2}{3} = 0.66$. In the other cases the guarantees is greater that 0.66. In particular, we can note that when we consider the background knowledge $S_Q = \textit{Station, Restaurant}$ the guarantee becomes higher. In fact, we decrease also $Pr(C_1|S_Q)$ that becomes 0.5.

We propose an empirical risk model also for c-safe datasets of semantic trajectories. In this case, the intent of the attacker is to approximate the cumulative distribution of the disclosure probability of sensitive places in a datasets $Pr(s|S_Q)$, knowing a sequence of quasi-identifiers S_Q with length h. This can be achieved by iterating over every value of $h = 1, \ldots, M$ where M is the length of the longest quasi-identifier sequence in the background knowledge BK_T. The idea is that for each h value, we select the group of semantic trajectories in D'_T containing the sequence of quasi-identifier places known by the attacker, called G. Then, for each sensitive place s in the semantic trajectories in G we compute the disclosure probability $Pr(s|S'_Q)$ and compute the cumulative distribution (Algorithm 2).

Algorithm 2 Attack simulation for c-safe semantic trajectories

Require: c-safe dataset D'_T, the set of semantic trajectories for the attacks BK_T (only quasi-identifier places), the privacy threshold c.

1: **for** $h = 1, \ldots, M$ where M is the length of the longest sequence of quasi-identifiers in BK_T. **do**
2: **for** $S'_Q \leq S_Q$ s.t. $S_Q \in BK_T$ and S'_Q has length h **do**
3: $G \leftarrow \{t \in D'_T | S'_Q \preceq t\}$.
4: **for** sensitive place s in the group of trajectories G **do**
5: $P \leftarrow Pr(s|S'_Q)$.
6: **end for**
7: **end for**
8: **end for**
9: **return** Cumulative Distribution of P for all h.

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*. This is

visually illustrated in Fig. 6. While a trajectory can consist of multiple hops, it can also be seen as a chain of smaller trajectories, each of which just contains 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*:

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

and *precision* as:

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

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:

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

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

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

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.

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.

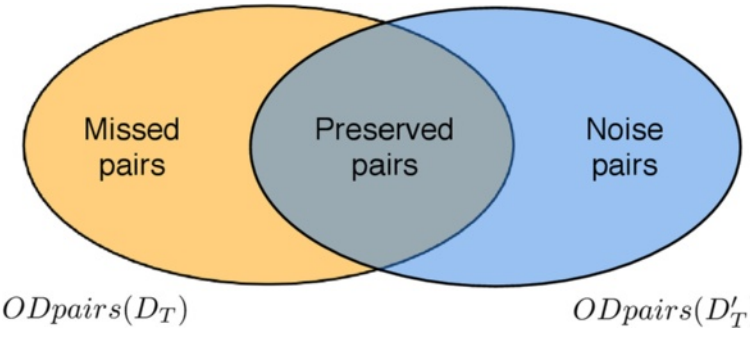


Fig. 6 Diagrammatic representation of coverage and precision. Coverage and precision illustrated

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 [12] Monreale et 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 consider in technological frameworks that offer the *by-design* 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.

In Fig. 7 we depict this methodology that is composed of three phases: (a) data preparation, (b) data anonymisation, and (c) risk evaluation.

The cycle, illustrated in Fig. 7 needs to be repeated with respect to the different dimensions (e.g., spatial and temporal granularity, refresh window) obtaining a collection of anonymised datasets $\mathcal{D}'^{i}_{T}$ with associated risks R^i. Given a class of services that are to be facilitated by the published data, the anonymised dataset $\mathcal{D}'^{i}_{T}$ will be chosen for which the associated risk R^i is *minimal* with acceptable utility of the published data.

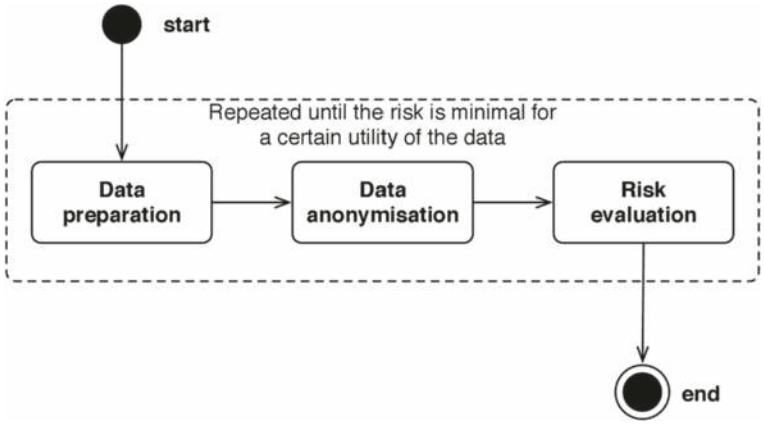


Fig. 7 Refining privacy and risk until the risk is minimal for a certain utility of the data. The privacy by design paradigm

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. 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 shown in Fig. 8a.

To begin with, the privacy-sensitive locations captured through GPS readings were obfuscated using Voronoi tessellation [13]. Each trajectory is then translated into a sequence of locations represented by the centers of the Voronoi tessellation. In Fig. 8b those locations are shown and a visual metaphor of the existing *ODpairs* in the data is depicted representing the existing connection between the locations (almost a complete graph). Moreover, the data was further subjected to k-anonymisation for $k = 3$, $k = 5$, and $k = 10$ by using the method proposed in [6]. Before applying this anonymisation, we subjected the trajectory 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 9 illustrates an example of a user trajectory observed at an hour-level and at the 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 Section “Data utility measures: coverage and precision”.

Risk analysis

In order to evaluate the privacy risks on the two anonymised trajectory datasets we applied the methodology described in Section “Empirical risk model for anonymised trajectory data”. Therefore, we estimated the cumulative distribution of

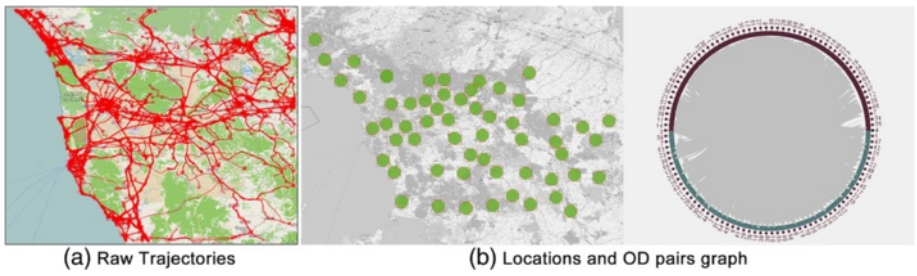


Fig. 8 User trajectory data in the different tessellated areas. An example of user trajectory through the different tessellated areas observed at an hour-level and at a day-level

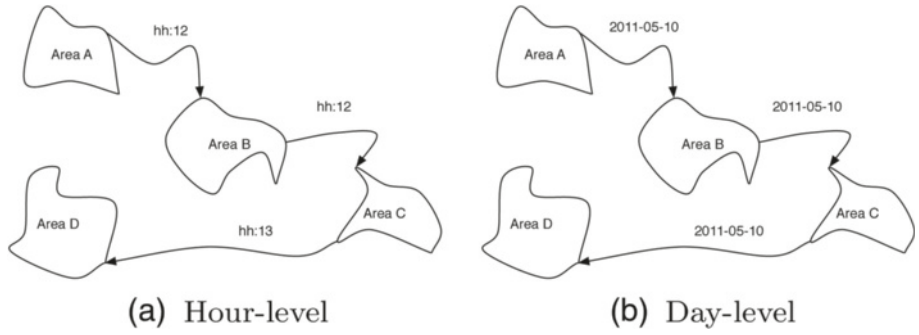


Fig. 9 User trajectory illustrated. An illustration of user trajectory through the different tessellated areas observed at an hour-level and at a day-level

the probability of re-identification for each value of $h = |t'|$, which 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 10 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 "Empirical risk model for anonymised trajectory data"). Figure 11 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 Fig. 11b 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 $\left(\frac{1}{20}\right)$ for 90 % of the attacks.

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

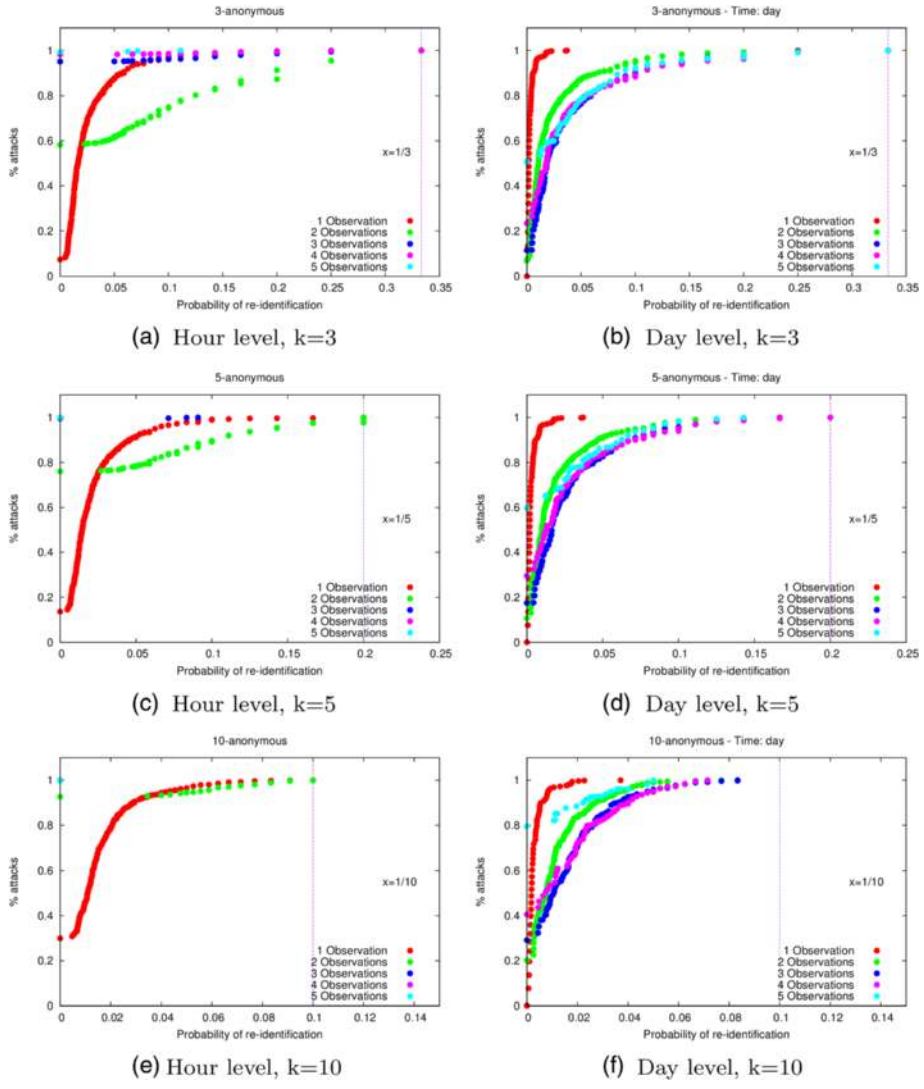


Fig. 10 Cumulative distribution of the re-identification probability

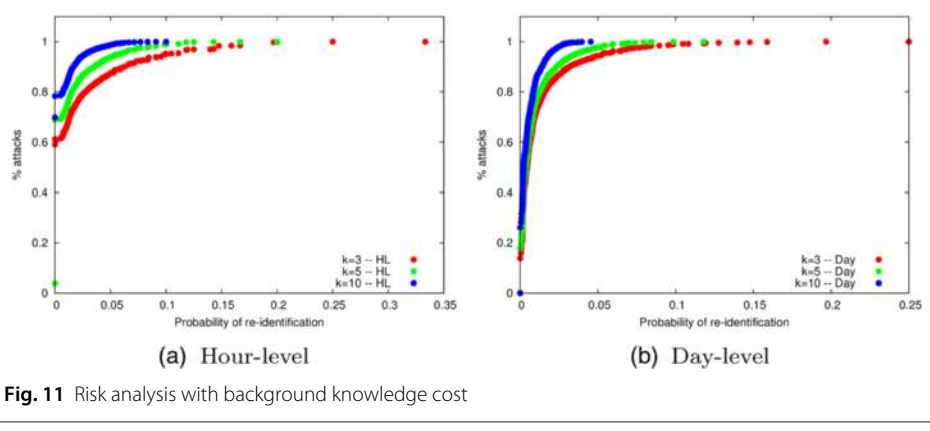


Fig. 11 Risk analysis with background knowledge cost

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

k	Precision	Coverage
	(a) Time: hour-level	
3	1.00	0.27
5	1.00	0.15
10	1.00	0.04
	(b) Time: day-level	
3	0.98	0.87
5	0.97	0.83
10	0.96	0.72

k-anonymous versions of the *hour-level dataset* while Table 1(b) shows the same information for the *day-level dataset*.

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 12 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.

Another way to evaluate the real effect of the data quality is a visual representation of the *ODpairs* graph. Figure 13 shows how the graphs degrade increasing the *k* parameter considering the hour-level and day-level aggregations.

Here, green locations represent vertices of the graph having both in- and out-going edges, on the contrary blue and orange locations represent respectively vertices with only out-going and in-going edges. In the original data all the locations are green and

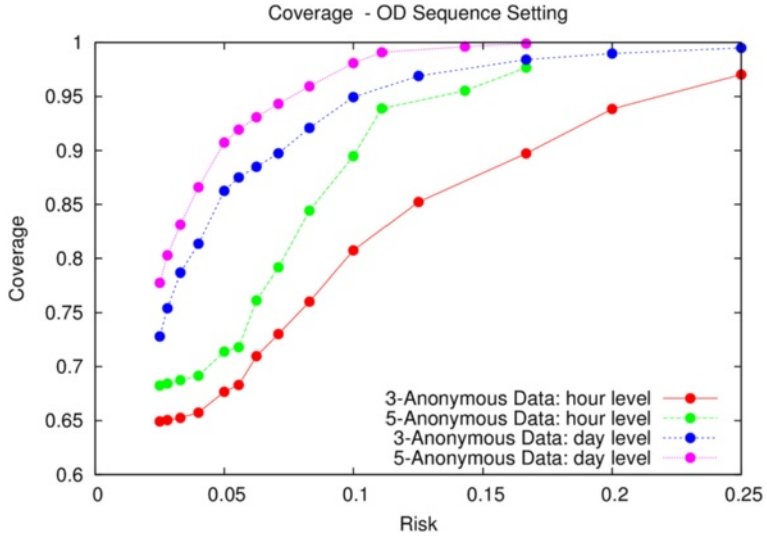


Fig. 12 Coverage with respect to privacy risk

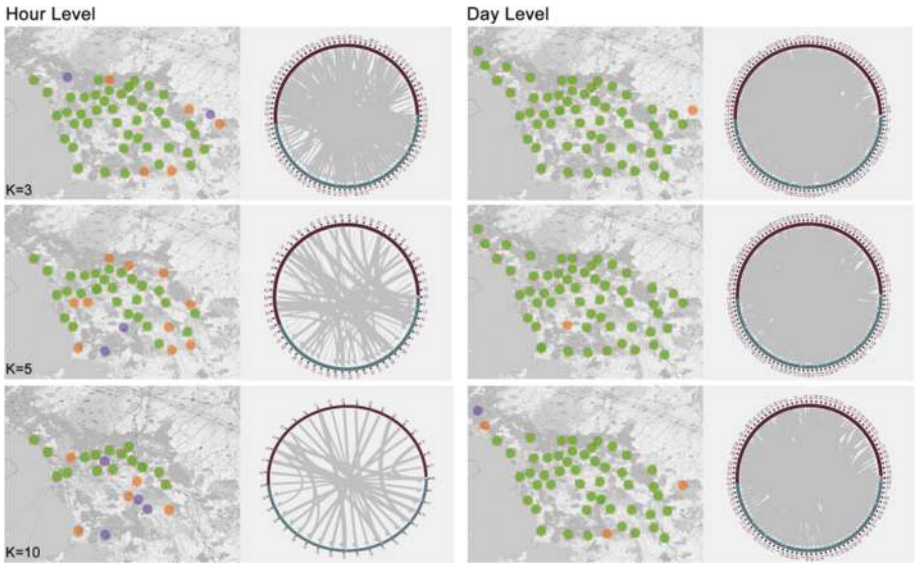


Fig. 13 *ODpairs* graph

the *ODpairs* graph is almost complete (see Fig. 8b) but in the case of hour-level even with $k = 3$ some of the nodes lose their connections becoming blue and orange and the relative graph becomes less dense. Using $k = 10$ the data is completely destroyed. At day-level the situation changes completely and with all the values of k the data remains similar to the original one: the locations do not disappear and the graph remain almost complete.

Risk analysis on null model

The risk analysis results shown above highlight that the empirical protection guaranteed by the algorithm of anonymisation is much higher than the theoretical protection. Our claim is that this happens because real human data such as movement data, describe the behavior of users who during the daily activity have to respect specific constraints that depend on different factors and this can bring to generate data describing similar behavior. Example of constraints are streets network topology or traffic during rush hours. All these factors constraint people movements generating trajectories which are similar and then with an high frequency. To prove this we generated a set of random null models starting from hypothetical locations and users moving without any constraint. Applying the empirical risk model to these null models we should have two main effects: first, the empirical protection guaranteed by the algorithm of anonymisation should not be so high like in the real-data and, second, the data quality after the anonymisation should decay because the algorithm hide more information to guaranteed the same level of privacy. The null models used consist in a set of randomly generated trajectories over a set of pseudo locations in a way that the number of users, the number of locations and the distribution of trajectories length (in terms of number of locations traversed) are equal to the original data. After the application of the anonymisation process over this synthetic data we measured the *coverage* and *precision* obtaining the values shown in Table 2. As predicted the hour-level values are significantly lower (by order of magnitudes) than the

Table 2 Precision versus coverage of the *k*-anonymised synthetic data

k	Precision	Coverage
	(a) Time: hour-level	
3	1.00	0.14
5	1.00	0.058
10	1.00	0.005
	(b) Time: day-level	
3	1	1
5	1	1
10	1	1

ones computed using the original data. The same does not happen for the day-level due to the fact that (i) at this level the number of possible *ODpairs* is lower because by generalised the time the data becomes denser, and (ii) the synthetic trajectories are generated with an uniform distribution over the locations, e.g. as said before they do not have any constraint to be followed. Therefore, the generated dataset is very dense with an uniform frequency distribution. In the case of *ODpairs* this frequency values are greater than 10, therefore all with the attack simulation we find everything safe.

In Fig. 14 show the empirical evaluation of the privacy guarantees over the null models. The first column of images contains the plots with the cumulative distributions of the re-identification probability related to the *hour-level dataset* while, the second column contains the results obtained from the *day-level dataset*. We observe that in the *hour-level dataset* (left images), when the number of attacker's observations is greater than 2 almost 100 % of attacks has a probability of success equal to 0 %, especially for $k = 5$ and $k = 10$. This means that the anonymisation algorithm tends to suppress a lot of information to provide good privacy levels. The curve representing a number of observations equal to 2 (green curve) describes a lower guarantee with respect to the same case in the real-world data (Fig. 10). On the other side, at day-level we can notice how the uniform distribution of the locations frequencies is evident: considering an attacker knowledge equal to 1 observation the dataset is practically safe. Moving towards a richer knowledge of the attacker, i.e. from 2 to 5 observations, the results change dramatically following the same trends seen for the hour-level. This happens because the frequency of a sequences is clearly lower than a single location, therefore when the frequencies become less than k the data will be destroyed by the algorithm.

To better understand this effect of and to prove that the results shown before are due to the density of the generated data, we modified the null model definition varying the number of locations used. The objective is to reduce the density of the *ODpairs* by means of having more combinations and maintaining the same number of users. The results shown in Fig. 15 confirm our hypothesis, in fact at both levels and for each k when the density reaches a critical value the coverage of the anonymized data decreases drastically destroying completely the data. Here the number of locations are: 64, 144, 256, 400, 576, 784 and 1024.

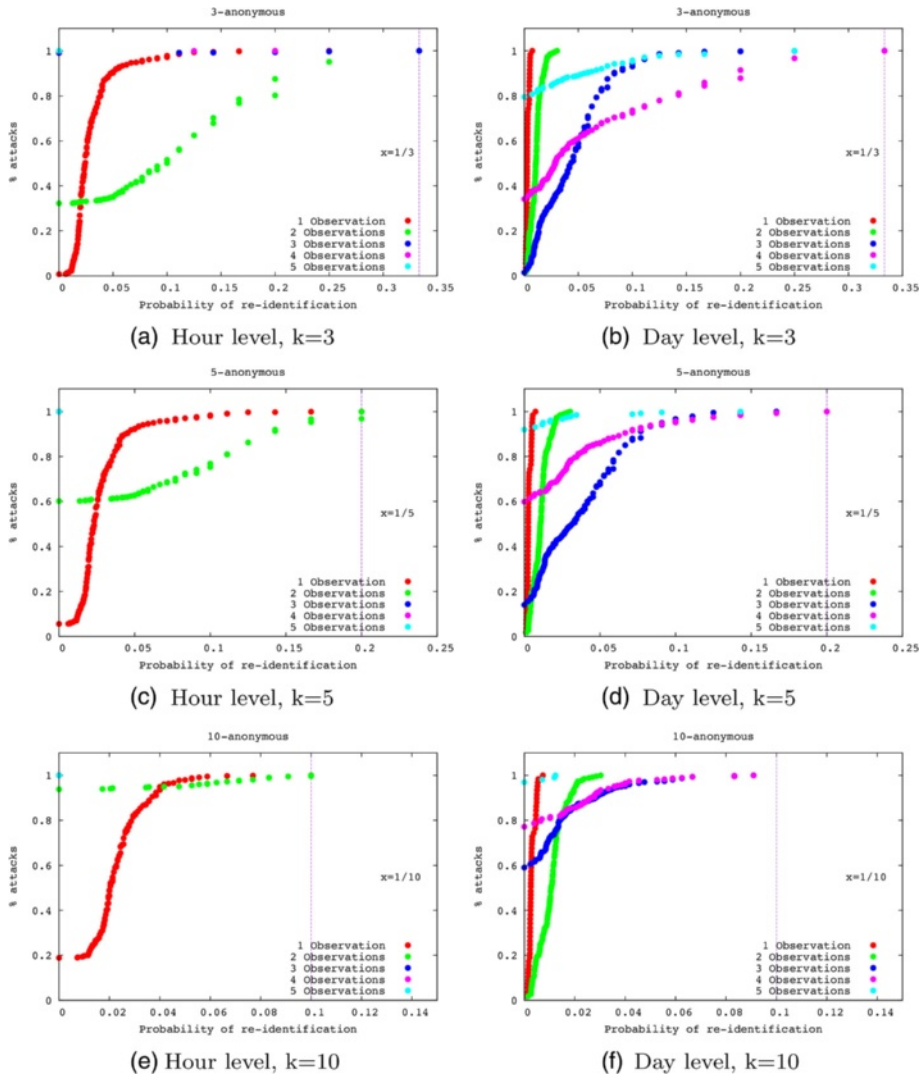


Fig. 14 Comparison with the null model. Cumulative distribution of the re-identification probability on data generated by the null model

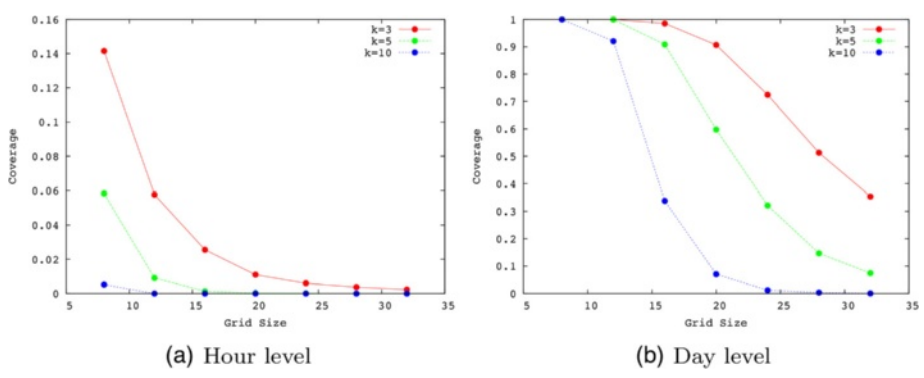


Fig. 15 Coverage by varying the grid size

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 [14]. 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. [15] presented privacy risk models for ubiquitous systems in order to convert privacy from an abstract concept into specific issues relating to concrete applications. Kosa et al. [16], 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. [17] proposes a privacy policy management mechanism whereby a match is made between user's personal privacy requirements and organisational privacy policies. PrivAware [18] was presented as a tool to detect and report unintended loss of privacy in a social network. Krishnamurthy et al. [19] 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. Yu et al. [20] 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. [21] 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 [22–25]. Some notable data anonymisation work include k-anonymity [2], l-diversity [26], t-closeness [27], p-sensitive k-anonymity [28], (α, k)-anonymity [29] and ϵ-differential privacy [30]. The k-anonymity model has been also studied and adapted in the context of movements data in different works: [4] exploits the inherent uncertainty of the moving object's whereabouts; [5] proposes a technique based on *suppression* of the dangerous observations from each trajectory; and [6] proposes a data-driven spatial generalization approach to achieve k-anonymity. A critique by Domingo-Ferrer and Torra [31] 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 [32, 33]. Sramka et al. [34] 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.

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. We also evaluate the model on synthetic data we have used as a null model to prove that the empirical evaluation of privacy protection is much better in real-world data because these data describe the behavior of users that during their activity must respect specific (external) constraints that influence the generated data.

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.

Competing interests
The authors declare that they have no competing interests.

Authors' contributions
AB, AN and RT conceptualised the paper along with DP, FG and SK. AB wrote most parts of the paper while AN and RT contributed the experimental results and some parts of the paper. JCC, SK, YM and TY contributed to proof-reading, discussions and criticisms about ideas and overall structure of the paper. All authors read and approved the final manuscript.

Authors' information
Anirban Basu is a Senior Researcher at KDDI R&D Laboratories in Japan. Prior to that, he worked as a Post-doctoral Researcher at Tokai University. He is also a Visiting Research Fellow at the University of Sussex. He holds a Ph.D. in Computer Science (2010) and a Bachelor of Engineering (Hons.) (2004) in Computer Systems Engineering from the University of Sussex. His research interests are in computational trust, privacy and security and peer-to-peer networks. He is particularly active within the IFIPTM computational trust management community.

Anna Monreale is an assistant professor at the Computer Science Department of the University of Pisa and a member of the Knowledge Discovery and Data Mining Laboratory (KDD-Lab), a joint research group with the Information Science and Technology Institute of the National Research Council in Pisa. Her research interests include big data analytics, social networks and the privacy issues raising in mining these kinds of social and human sensitive data. She earned her Ph.D. in computer science from the University of Pisa in 2011.

Roberto Trasarti graduated in Computer Science in 2006, at the University of Pisa. He discussed his thesis on ConQueSt: a Constraint-based Query System aimed at supporting frequent patterns discovery. He started the Ph.D. in Computer Science at the School for Graduate Studies "Galileo Galilei", (University of Pisa). In June 2010 he received his Ph.D. presenting the thesis entitled "Mastering the Spatio- Temporal Knowledge Discovery Process". Currently he is a researcher at ISTI-CNR as member of Knowledge Discovery and Delivery Laboratory. His interests regard data mining, mobility data analysis, artificial intelligence and automatic Reasoning.

Juan C Corena is a Software Engineer at Google Inc. Prior to that, he worked as a researcher at KDDI R&D Laboratories in Japan. He holds a Ph.D. in engineering from Keio University (2014). His research interests include: network security, applied cryptography and error-correcting codes.

Fosca Giannotti is a senior researcher at the Information Science and Technology Institute of the National Research Council at Pisa, Italy, where she leads the Knowledge Discovery and Data Mining Laboratory – KDD LAB – a joint research initiative with the University of Pisa, founded in 1995, one of the earliest European research groups specifically targeted at data mining and knowledge discovery. Her current research interests include data mining query languages, knowledge discovery support environment, web-mining, spatio-temporal reasoning, spatio-temporal data mining, and privacy preserving data mining.

Dino Pedreschi is a full professor of Computer Science at the University of Pisa. He has been a visiting scholar at the University of Texas at Austin (1989/90), at CWI Amsterdam (1993) and at UCLA (1995). His current research interests are in data mining and logic in databases, and particularly in data analysis, in spatio-temporal data mining, and in privacy-preserving data mining. He is a member of the program committee of the main international conferences on data mining and knowledge discovery and an associate editor of the journal Knowledge and Information Systems. He has been granted a Google Research Award (2009) for his research on privacy-preserving data mining and anonymity-preserving data publishing.

Shinsaku Kiyomoto received his B.E. in engineering sciences and his M.E. in Materials Science from Tsukuba University, Japan, in 1998 and 2000, respectively. He joined KDD (now KDDI) and has been engaged in research on stream ciphers, cryptographic protocols, and mobile security. He is currently a senior researcher at the Information Security Laboratory of KDDI R&D Laboratories Inc. He was a visiting researcher of the Information Security Group, Royal Holloway University of London from 2008 to 2009. He received his doctorate in engineering from Kyushu University in 2006. He received the IEICE Young Engineer Award in 2004. He is a member of JPS.

Yutaka Miyake received the B.E. and M.E. degrees of Electrical Engineering from Keio University, Japan, in 1988 and 1990, respectively. He joined KDD (now KDDI) in 1990, and has been engaged in the research on high-speed communication protocol and secure communication system. He received the Dr. degree in engineering from the University of Electro-Communications, Japan, in 2009. He is currently a senior manager of Information Security Laboratory in KDDI R&D Laboratories Inc. He received IPSJ Convention Award in 1995 and the Meritorious Award on Radio of ARIB in 2003.

Tadashi Yanagihara received the B.S. and M.S. degree in Computer Science from Keio University in 2002 and 2004, respectively. He was an associate research engineer at KDDI R&D Laboratories Inc. from 2005 to 2010, where he worked on recommendation systems and text mining, and a researcher at Toyota InfoTechnology Center Inc. from 2011 to 2015, where he worked on geo-mining and data mining based on large-scale automobile sensor data. He is currently an assistant manager at KDDI Corporation where is work is on R&D Strategies and Technology Development for connected car and artificial intelligence. He is a member of the IEICE, IPSJ, JSAI and ACM.

Acknowledgments
The authors thank the anonymous reviewers for their valuable time and feedback. The contributions of Juan Camilo Corena to this paper represent his work during his time at KDDI R&D; and are not related in any way to his work with his current employer - Google Inc.

Author details
[1]Information Security Group, KDDI R&D Laboratories, 2-1-15 Ohara, Fujimino, Saitama 356-8502, Japan. [2]KDD Lab, University of Pisa, Pisa, Italy. [3]KDD Lab, ISTI CNR, Pisa, Italy. [4]Toyota ITC, Fujimino, Tokyo, Japan.

Received: 26 December 2014 Accepted: 22 October 2015
Published online: 13 November 2015

References

1. Samarati P (2001) Protecting respondents identities in microdata release. IEEE Trans Knowl Data Eng 13(6):1010–1027
2. Sweeney L (2002) k-anonymity: A model for protecting privacy. Int J Uncertain Fuzziness and Knowl-Based Syst 10(05):557–570
3. Monreale A, Trasarti R, Pedreschi D, Renso C, Bogorny V (2011) C-safety: a framework for the anonymization of semantic trajectories. Trans Data Priv 4(2):73–101
4. Abul O, Bonchi F, Nanni M (2008) Never walk alone: Uncertainty for anonymity in moving objects databases. In: The 24th IEEE International Conference on Data Engineering (ICDE). IEEE Computer Society, Cancun, Mexico. pp 376–385
5. Terrovitis M, Mamoulis N (2008) Privacy preservation in the publication of trajectories. In: MDM. pp 65–72
6. Monreale A, Andrienko GL, Andrienko NV, Giannotti F, Pedreschi D, Rinzivillo S, Wrobel S (2010) Movement data anonymity through generalization. TDP 3(2):91–121
7. Spaccapietra S, Parent C, Damiani ML, de Macêdo JAF, Porto F, Vangenot C (2008) A conceptual view on trajectories. Data Knowl Eng 65(1):126–146
8. Bogorny V, Wachowicz M (2009) A Framework for Context-Aware Trajectory. In: Cao L, Yu P, Zhang C, Zhang H (eds). Data Mining for Business Applications. Springer US, USA. pp 225–239. doi:10.1007/978-0-387-79420-4_16
9. Alvares LO, Bogorny V, Kuijpers B, de Macêdo JAF, Moelans B, Vaisman AA (2007) A model for enriching trajectories with semantic geographical information. In: 15th ACM International Symposium on Geographic Information Systems, ACM-GIS 2007, Seattle, Washington, USA, Proceedings. p 22
10. Palma AT, Bogorny V, Kuijpers B, Alvares LO (2008) A clustering-based approach for discovering interesting places in trajectories. In: Proceedings of the 2008 ACM Symposium on Applied Computing (SAC), Fortaleza, Ceara, Brazil. pp 863–868
11. Rocha JAMR, Times VC, Oliveira G, Alvares LO, Bogorny V (2010) Db-smot: A direction-based spatio-temporal clustering method. In: 5th IEEE International Conference on Intelligent Systems, IS 2010. University of Westminster, London, UK. pp 114–119
12. Monreale A, Pedreschi D, Pensa RG, Pinelli F (2014) Anonymity preserving sequential pattern mining. Artif Int Law. Springer 22(2):141–173
13. Voronoï G (1908) 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
14. Westin AF (1968) Privacy and freedom. Washington and Lee Law Review 25(1):166
15. Hong JI, Ng JD, Lederer S, Landay JA (2004) Privacy risk models for designing privacy-sensitive ubiquitous computing systems. In: Proceedings of the 5th Conference on Designing Interactive Systems: Processes, Practices, Methods, and Techniques, DIS '04. ACM, New York, NY, USA. pp 91–100. doi:10.1145/1013115.1013129
16. Kosa TA, El-Khatib K, Marsh S (2011) Measuring privacy. J Internet Serv Inf Secur (JISIS) 1(4):60–73
17. Kiyomoto S, Nakamura T, Takasaki H, Watanabe R, Miyake Y (2013) PPM: Privacy Policy Manager for Personalized Services. In: Cuzzocrea A, Kittl C, Simos D, Weippl E, Xu L (eds). Security Engineering and Intelligence Informatics, Lecture Notes in Computer Science. Springer Berlin Heidelberg, Germany Vol. 8128. pp 377–392. doi:10.1007/978-3-642-40588-4_26
18. Becker J, Chen H (2009) Measuring privacy risk in online social networks. In: Web 2.0 Security and Privacy (W2SP), Oakland, CA, USA
19. Krishnamurthy B, Malandrino D, Wills CE (2007) Measuring privacy loss and the impact of privacy protection in web browsing. In: Proceedings of the 3rd Symposium on Usable Privacy and Security, SOUPS '07. ACM, New York, NY, USA. pp 52–63. doi:10.1145/1280680.1280688
20. Yu T, Zhang Y, Lin K-J (2006) 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. IEEE, San Francisco, USA. pp 4–4

21. Banescu S, Petković M, Zannone N (2012) Measuring privacy compliance using fitness metrics. In: Barros A, Gal A, Kindler E (eds). Business Process Management, Lecture Notes in Computer Science. Springer Berlin Heidelberg, Germany Vol. 7481. pp 114–119. doi:10.1007/978-3-642-32885-5_8
22. Agrawal R, Srikant R (2000) Privacy-preserving data mining. ACM SIGMOD Record 29(2):439–450
23. Dinur I, Nissim K (2003) Revealing information while preserving privacy. In: Proceedings of the 22nd ACM SIGMOD-SIGACT-SIGART Symposium on Principles of Database Systems, PODS '03. ACM, New York, NY, USA. pp 202–210. doi:10.1145/773153.773173
24. Evfimievski A, Gehrke J, Srikant R (2003) Limiting privacy breaches in privacy preserving data mining. In: Proceedings of the 22nd ACM SIGMOD-SIGACT-SIGART Symposium on Principles of Database Systems, PODS '03. ACM, New York, NY, USA. pp 211–222. doi:10.1145/773153.773174
25. Blum A, Dwork C, McSherry F, Nissim K (2005) Practical Privacy: The SuLQ Framework. In: Proceedings of the 24th ACM SIGMOD-SIGACT-SIGART Symposium on Principles of Database Systems, PODS '05. ACM, New York, NY, USA. pp 128–138. doi:10.1145/1065167.1065184
26. Machanavajjhala A, Kifer D, Gehrke J, Venkitasubramaniam M (2007) l-diversity: Privacy beyond k-anonymity. ACM Trans Knowl Discov from Data (TKDD) 1(1):3
27. Li N, Li T, Venkatasubramanian S (2007) t-closeness: Privacy beyond k-anonymity and l-diversity. In: The 23rd IEEE International Conference on Data Engineering (ICDE). IEEE Computer Society, Istanbul, Turkey. pp 106–115
28. Truta TM, Vinay B (2006) Privacy protection: p-sensitive k-anonymity property. In: The 22nd International Conference on Data Engineering Workshops. IEEE, Atlanta, GA, USA. pp 94–94
29. Wong RC-W, Li J, Fu AW-C, Wang K (2006) (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
30. Dwork C (2006) Differential privacy. In: Bugliesi M, Preneel B, Sassone V, Wegener I (eds). Automata, Languages and Programming, Lecture Notes in Computer Science. Springer Berlin Heidelberg, Germany Vol. 4052. pp 1–12. doi:10.1007/11787006_1
31. Domingo-Ferrer J, Torra V (2008) A critique of k-anonymity and some of its enhancements. In: The 3rd International Conference on Availability, Reliability and Security (ARES). IEEE, Barcelona, Spain. pp 990–993
32. Rastogi V, Suciu D, Hong S (2007) The boundary between privacy and utility in data publishing. In: Proceedings of the 33rd International Conference on Very Large Data Bases, VLDB '07. VLDB Endowment, Vienna, Austria. pp 531–542. http://dl.acm.org/citation.cfm?id=1325851.1325913
33. Brickell J, Shmatikov V (2008) The Cost of Privacy: Destruction of Data-mining Utility in Anonymized Data Publishing. In: Proceedings of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, KDD '08. ACM, New York, NY, USA. pp 70–78. doi:10.1145/1401890.1401904
34. Sramka M, Safavi-Naini R, Denzinger J, Askari M (2010) A practice-oriented framework for measuring privacy and utility in data sanitization systems. In: Proceedings of the 2010 EDBT/ICDT Workshops, EDBT '10. ACM, New York, NY, USA. pp 27:1–27:10. doi:10.1145/1754239.1754270

Submit your manuscript to a SpringerOpen journal and benefit from:

- ► Convenient online submission
- ► Rigorous peer review
- ► Immediate publication on acceptance
- ► Open access: articles freely available online
- ► High visibility within the field
- ► Retaining the copyright to your article

Submit your next manuscript at ► springeropen.com