Preserving Privacy in Semantic-Rich Trajectories of Human Mobility

Anna Monreale
KDD-Lab
University of Pisa, Italy
annam@di.unipi.it

Roberto Trasarti
KDD-Lab
ISTI CNR, Pisa, Italy
roberto.trasarti@isti.cnr.it

Chiara Renso
KDD-Lab
ISTI CNR, Pisa, Italy
chiara.renso@isti.cnr.it

Dino Pedreschi
KDD-Lab
University of Pisa, Italy
pedre@di.unipi.it

Vania Bogorny
UFSC
Florianopolis, SC,Brazil
vania@inf.ufsc.br

ABSTRACT

The increasing abundance of data about the trajectories of personal movement is opening up new opportunities for analyzing and mining human mobility, but new risks emerge since it opens new ways of intruding into personal privacy. Representing the personal movements as sequences of places visited by a person during her/his movements - semantic trajectory - poses even greater privacy threats w.r.t. raw geometric location data. In this paper we propose a privacy model defining the attack model of semantic trajectory linking, together with a privacy notion, called *c-safety*. This method provides an upper bound to the probability of inferring that a given person, observed in a sequence of non-sensitive places, has also stopped in any sensitive location. Coherently with the privacy model, we propose an algorithm for transforming any dataset of semantic trajectories into a *c-safe* one. We report a study on a real-life GPS trajectory dataset to show how our algorithm preserves interesting quality/utility measures of the original trajectories, such as sequential pattern mining results.

Categories and Subject Descriptors

H.2.8 [**Database Applications**]: Spatial databases; K.4.1 [**Public Policy Issues**]: Privacy

General Terms

Algorithms

Keywords

Anonymity, Privacy, Semantic Trajectory, Pattern Mining

1. INTRODUCTION

The increasing abundance of data about the trajectories of personal movement, obtained through mobile phones, navigation and GPS devices, and so on, is opening up new avenues for analyzing and mining human mobility, with applications in diverse scientific and social domains. Besides new opportunities, also new risks emerge, as knowing the whereabouts of people also opens new ways of intruding into their personal privacy; this observation is at the basis of some recent research works that addressed the problem of protecting privacy while disclosing trajectory data. However, the progress on mobile device technology, geographic information and mobility data analysis and mining are creating entirely new forms of trajectory data, with far richer semantic information attached to the traces of personal mobility: we are rapidly moving from raw trajectories, i.e., sequences of time-stamped generic points sampled during the movement of a sensed device, to what we call *semantic trajectories*, i.e., sequences of stops and moves of a person during her/his movements, where each location of stop can be attached to some semantics, or purpose - either by explicit sensing or by inference.

In this paper, we argue that the new form of data of semantic trajectories poses even greater privacy threats w.r.t. raw location data, and we propose a privacy model to face this challenging problem. The first problem introduced by this form of data is that, from the fact that a person has stopped in a certain sensitive location, e.g., an oncology clinic, an attacker can derive private personal information of the health of such person. So, in this context, a place is *sensitive* if it allows to infer personal sensitive information of an individual. Moreover, it is easy to show that guaranteeing the privacy in semantic trajectories is not trivial. In fact, just hiding a person's trajectory into a crowd, following the idea of k-anonymity, is not enough for a robust protection: when individuals in a crowd of people with similar trajectories stop in the same sensitive place, we can easly infer the individual sensitive information.

The problem resembles the discussion about k-anonymity and l-diversity in relational, tabular data. Here, we essentially devise a similar privacy model for semantic trajectories, with reference to a background knowledge defining which are the sensitive and non-sensitive locations corresponding to stops. We represent this background knowledge through a place taxonomy, describing sensitive and non-sensitive locations at different levels of abstraction (e.g., a

SPRINGL '10 November 2, 2010, San Jose,CA,USA

turistic landmark, a museum, the Louvre museum; a health-related service, a hospital, the Childrens' Hospital).

The main contribution of this paper is the definition of the attack model of semantic trajectory linking, which formalizes the mentioned privacy-violating inferences, together with a privacy notion, called *c-safety*, which provides an upper bound c to the probability of inferring that a given person, observed in a sequence of non-sensitive places, has also stopped in any sensitive location.

Coherently with the introduced privacy model, we propose an algorithm for transforming any dataset of semantic trajectories into a *c-safe* one, which can be safely published under the specified privacy safeguard. Our algorithm is based on the generalization of places driven by a place taxonomy, thus providing a way to preserve the semantics of the generalized trajectories. We conduct a study on a real-life GPS trajectory dataset, and show how our algorithm preserves interesting quality/utility measures of the original trajectories. In particular, we show that sequential pattern mining results are preserved.

The rest of the paper is organized as follows. Section 2 discusses the relevant related studies on privacy issues in movement data. In Section 3 some basic definitions and background information are given. Section 4 introduces the problem of anonymity in semantic trajectories. In Section 5 we describe the privacy model. Section 6 describes the semantic generalization approach for trajectories. In Section 7 we discuss the experimental results of the application of our method on the real-world moving object dataset. Finally, Section 8 concludes the paper.

2. RELATED WORK

Many research studies have focused on the design of techniques for privacy-preserving data mining [2] and for privacy-preserving data publishing. The basic operation for data publishing is to replace personal identifiers with pseudonyms. However, in [19] authors showed that this simple operation is insufficient to protect privacy. They proposed k-anonymity to make each record indistinguishable with at least $k-1$ other records thus protecting data against the *linking attack*. The k-anonymity framework is the most popular method for the anonymization of spatio-temporal data and, for relational datasets, is based on the attribute distinction among *quasi identifiers* (attributes that could be used for linking with external information) and *sensitive* attributes (information to be protected) [21]. Although it has been shown that finding an optimal k-anonymization is NP-hard [12] and that k-anonymity has some limitations [11, 10], this framework is still very relevant and it is often used in the studies on privacy issues in transactional databases [9] and in location-based services (LBSs) [8, 13, 14], as well as on the anonymity of trajectories [1, 16, 24, 15].

In [9] authors present a k-anonymization approach for transactional database. This method is based on a top-down generalization and presents all the limitation deriving from the k-anonymity that we solve with our approach.

In [1], the authors propose the notion of (k, δ)-anonymity for moving object databases, where δ represents the possible location imprecision. The authors also proposed an approach, called *Never Walk Alone* based on trajectory clustering and spatial translation. In [16] Nergiz et al. addressed privacy issues regarding the identification of individuals in static trajectory datasets. They provide privacy protection by first enforcing k-anonymity and then randomly reconstructing a representation of the original dataset from the anonymization. In [24] different objects may have different quasi-identifiers and thus, anonymization groups associated with different objects may not be disjoint. Therefore, an innovative notion of k-anonymity based on spatial generalization is provided in order to generate anonymity groups that satisfy the novel notion of k-anonymity: *Extreme Union* and *Symmetric Anonymization*. In [15] authors present a method for the anonymization of movement data combining the notions of spatial generalization and k-anonymity and show how the results of clustering analysis are faithfully preserved.

Another approach based on the concept of k-anonymity is proposed in [18], where a framework for the k-anonymization of sequences of regions/locations is presented. The authors also propose an approach that is an instance of their framework, which enables protected datasets to be published while preserving the data utility for sequential pattern mining tasks.

In [4, 22], suppression-based approaches for trajectory data are suggested. In the first one, the objective is to sanitize the input database in such a way that a set of sensitive patterns is hidden. In the second one, given the head of the trajectories, reduces the probability of disclosing the tail of the them. It is based on the assumption that different attackers know different and disjoint portions of the trajectories and the data publisher knows the attacker's knowledge.

However, all these approaches deal with the anonymization of trajectories from the geometric point of view. So far, to the best of our knowledge, no approaches face the problem of anonymizing semantic trajectories. In the context of LBS the work [6] proposes a solution to protect personal location information when the adversary is aware of the *semantic locations*. The main difference between this work and ours is that we anonymize a dataset of semantic trajectories for a safe publication while [6] anonymizes a user's location during the communication with a LBS provider upon a service request.

Lastly, in [23] Valls et al. consider the problem of privacy preserving for sequence of events, which has similar characteristic to semantic trajectories data. Their approach finds clusters of records and then for each cluster constructs a prototype used to substitute the original values in the masked version of the data of the cluster.

3. BACKGROUND

In this section we briefly recall some basic concepts which are useful to understand the proposed anonymization framework, namely the notion of semantic trajectory as sequence of stops and moves and an introduction to ontologies and taxonomies.

3.1 Semantic trajectories

A trajectory has been defined as the spatio-temporal evolution of the position of a moving entity. A trajectory is typically represented as a discrete sequence of points. An interpolation function between two consecutive points approximates the movements between two sample points. Recently a new trajectory concept has been introduced in [20] for reasoning over trajectories from a semantic point of view, the *semantic trajectory*, 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. 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 [5]. To illustrate these concepts let us consider some basic definitions.

Definition 1 (Trajectory Sample). *A Trajectory Sample is a list of space-time points* $\langle x_0, y_0, t_0\rangle, \ldots, \langle x_N, y_N, t_N\rangle$*, where* $x_i, y_i \in R$*,* $t_i \in R^+$ *for* $i = 0, 1, \ldots, N$*, and* $t_0 < t_1 < t_2 < \cdots < t_N$*.*

Important parts of a trajectory, i.e., stops, correspond to the set of x, y, t points of a trajectory sample that are important from an application point of view. The important parts correspond to places that 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. For every type of important place a minimal amount of time is defined, such that a sub-trajectory should continuously intersect this place for it to be considered a stop. A set of important places characterizes a semantic trajectory.

Definition 2 (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 1 (2) illustrates the concept of semantic trajectory for the trajectory sample shown in Figure 1(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).

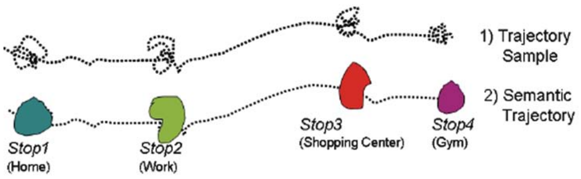

Figure 1: Example of Trajectory Sample and Semantic Trajectory

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. For instance, the method SMoT [3] verifies the intersection of the trajectory with a set of user defined geographic places, for a minimal amount of time. The method CB-SMoT (Clustering-based stops and moves of trajectories) [17] is a more sophisticated method that computes important places based on the variation of the speed of the trajectory. The important places are those in which the speed is lower than the average speed of the trajectory. After the low speed clusters have been computed, the method verifies for each cluster if it intersects the user defined geographic places, i.e., the possible places that were visited by the user. In positive case, this place is added to the sub-trajectory that intersects this place, building a semantic trajectory. Low speed clusters which do not intersect any geographic place are labeled as *unknown stops*. For the purpose of this paper, the unknown stops are simply omitted since not associated to any interesting place.

3.2 Domain Ontologies

The definition given by [7] is used to define ontology as "a technical term denoting an artifact that is *designed* for a purpose, which is to enable the modeling of knowledge about *some* domain, real or imagined". Such ontologies determine what can be represented and what can be inferred about a given domain, using a specific formalism of concepts. Usually, the term *domain ontology* is used to refer to ontologies describing the main concepts and relations of a given domain, i.e. urban or medical. Ontology basic elements are: *concepts* (or *classes*), which describe the common properties of a collection of individuals; *properties* are binary relations between concepts; *instances* represent the actual individuals of the domain. We say that a given individual is an *instance of* a concept when the individual properties satisfies the concept definition. A special property called is_a represents the $kind\_of$, or specialization, relationship between concepts. An ontology having only is_a relationships is called *taxonomy*. Formally, a taxonomy is a 2-tuple $Tax := \{C, HC\}$, where C is a set of concepts, HC is a taxonomy or concept-hierarchy, which defines the is_a relations among concepts ($HC(c1, c2)$ means that c_1 is a sub-concept of c_2). A taxonomy of places of interest represents the geographical places of interest in a given domain and is used during the generalization phase in the anonymization algorithm presented later in the paper. Here, the set of stop places obtained from the computation of semantic trajectories are the *leaves* of Tax, namely the stops places are semantically organized in a hierarchy. For example, we have that *R1* is a kind of Restaurant which is a kind of Enternainment, etc. Each concept in the taxonomy describes the categories of the geographical object of interest for the application domain. Figure 2 depicts an example of the taxonomy of places of interest in the urban used as example thought the paper. Here, the dotted square identifies the sensitive places, as discussed later.

4. ANONYMITY IN SEMANTIC TRAJECTORIES

Our main idea is to provide a framework to generate an *anonymous semantic trajectory dataset* which guarantees that it will not be possible to infer the identity of a user and the sensitive places visited by him/her with a probability greater than a given threshold set by the data owner. To this aim we propose a method based on the generalization of the places visited by a user driven by the place taxonomy, allowing to preserve semantic information in the anonymized dataset.

To avoid the identification of sensitive places visited by a user first of all we need to specify which places are *sensitive* and which one are *non-sensitive*. In relational data the sensitivity notion is defined on the table attributes where it is possible to specify the quasi-identifier (public information that can be used to discover private information) and sensitive attributes (information to be protected). In semantic trajectories, due to the particular format of the data and the intrinsic geographical nature of data, this distinction is among the *stop places*. A place is considered sensitive when it allows to infer personal information about the person who has stopped there. This means that some

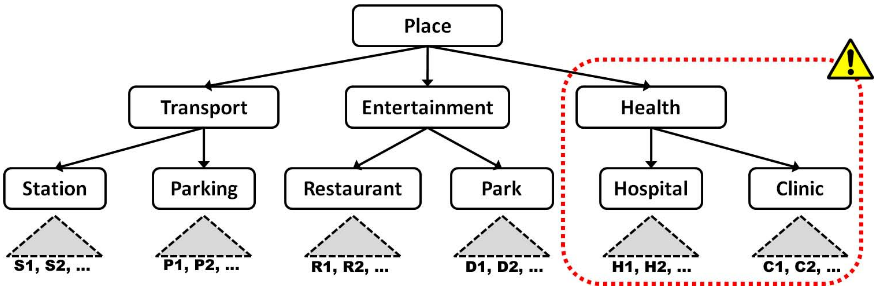

Figure 2: The Places taxonomy

places, such as an oncology clinic, in some applications can be sensitive because an attacker can derive that a person has stopped there probably has the cancer; whereas others places (such as parks) can be considered as quasi-identifiers. We propose here to exploit the taxonomy to establish this distinction, therefore some taxonomy concept can be tagged as "sensitive" and, as a consequence, all the remaining concepts as quasi-identifiers. We assume the labelled taxonomy is given by the domain expert who tags each concept with the "sensitivity" label. To formally introduce the sensitivity label of a concept we define a *privacy place taxonomy* as an extension of the places taxonomy Tax with a function λ which assigns a concept of C with a label belonging to the set $L = \{s, q\}$, where s means "sensitive" and q means "quasi-identifier". Hence, the taxonomy becomes a triple $PTax := \langle C, HC, \lambda \rangle$.

Given a dataset of semantic trajectories ST and the privacy places taxonomy $PTax$, we define a framework to anonymize ST by using a method based on the generalization of places driven by the taxonomy.

Now, we introduce the notion of *quasi-identifier place sequence* in the context of semantic trajectories. We use Q and SP to denote the set of quasi-identifiers places and the set of sensitive places defined in the taxonomy $PTax$, respectively.

Definition 3 (Quasi-identifier place sequence). *A quasi-identifier sequence $S_Q = q_1, \ldots, q_n$, where $n > 0$ and $q_i \in Q$ is a temporally ordered sequence of stop places of the privacy place taxonomy $PTax$ labelled as quasi-identifier and that can be joined with external information to re-identify individual trajectories with sufficient probability.*

We assume that quasi-identifier places are known based on specific knowledge of the domain.

5. PRIVACY MODEL

Let ST denote the original dataset of semantic trajectories. The dataset owner applies a transformation on ST to obtain ST^*. Our privacy scheme is based on generalization of all the semantic trajectories driven by a privacy place taxonomy $PTax$ that describes the categories of the geographical object of interest for the application domain.

Definition 4 (Generalized Semantic Trajectory). *Let $T = \{p_1, p_2, \ldots, p_n\}$ be a semantic trajectory. A generalized version of T, obtained by the place taxonomy $PTax$, is a sequence of places $T_g = g_1, g_2 \ldots, g_n$ where $\forall i = 1, \ldots, n$ we have that $HC(p_i, g_i)$ holds or $g_i = p_i$.*

In other words, a place g_i of a generalized semantic trajectory can be either an ancestor of the original place p_i or p_i itself.

Definition 5. *Let $T_g = g_1, g_2 \ldots, g_n$ be a generalized semantic trajectory and $A = p_1, \ldots, p_m$ a sequence of places. We say that A* is contained *in T_g ($A \preceq T_g$) if there exist integers $1 \leq i_1 < \ldots < i_m \leq n$ such that $\forall 1 \leq j \leq m$ we have $g_{i_j} = p_j$ or the relation $HC(p_j, g_{i_j})$ holds.*

We refer to the number of generalized semantic trajectories in ST^* containing a sequence of places A as *support of A* and denote it by $supp_{ST^*}(A)$. More formally, $supp_{ST^*}(A) = |\{T_g \in ST^* | A \preceq T_g\}|$ and $supp_{ST^*}(A, B) = |\{T_g \in ST^* | A \preceq T_g \wedge B \preceq T_g\}|$.

5.1 Adversary Knowledge

An intruder who gains access to ST^* may possess some background knowledge allowing he/she to conduct attacks that may allows her/him to make inferences on the dataset. We generically refer to any of these agents as an attacker. We adopt a conservative model and in particular we assume the following adversary knowledge.

Definition 6 (Adversary Knowledge). *The attacker has access to the generalized dataset ST^* and knows: (a) the schema used to anonymize the data, (b) the privacy place taxonomy $PTax$, that describes the levels of abstraction, (c) that a given user is in the dataset and (d) a quasi-identifier place sequence S_Q visited by this user.*

5.2 Attack Model

What is the information that has to remain private? In our model, we keep private all the sensitive places visited by a given user. Therefore, the attack model considers the ability to link the released data to other external information enabling to infer sensitive places visited by a given user.

Definition 7. *The attacker, given a published semantic trajectory dataset ST^* where each trajectory is uniquely associated to a de-identified respondent, tries to identify the semantic trajectory in ST^* associated to a given respondent U, based on the additional knowledge introduced in Definition 6. The attacker, given the quasi-identifier sequence S_Q constructs a set of candidate semantic trajectories in ST^* containing S_Q and tries to infer the sensitive places related to U. We denote by $Prob(S_Q, S)$ the probability that, with a quasi-identifier place sequence S_Q related to a user U, the attacker infers the sequence of sensitive places S visited by the user.*

From a data protection perspective, we aim at controlling the probability $Prob(S_Q, S)$. To prevent the attack defined above we propose to release a *c-safe* dataset.

Definition 8 (c-Safe Dataset). *The dataset ST is said c-safe with respect to the place set Q if for every quasi-itentifier place sequence S_Q, we have that for each set of sensitive place S $Prob(S_Q, S) \leq c$ with $c \in [0, 1]$.*

Given these definitions, we formulate the problem statement as follows:

Problem Statement. Given a dataset ST of semantic trajectories and a protection probability threshold that we want to guarantee $c \in [0, 1]$, find a c-safe version ST^* of ST.

In other words, we want to avoid that an adversary can use a sequence of quasi-identifier places visited by a user to correctly infer any sensitive places after accessing ST^*. The problem that we want to address is very similar to l-diversity [11], but the particular nature of the data makes the problem different and such framework cannot be directly applied to this case. First of all, the semantic trajectories does not have fixed length. In particular, given any two semantic trajectories S_i and S_j belonging to the same database, the number of quasi-identifiers place stops contained in T_i is in general different from T_j. Moreover, in a semantic trajectory we can have more than one sensitive place stop, and their number is not fixed for all trajectories. Note also that it is possible that a trajectory is composed only by quasi-identifier places or only sensitive places. In the first case, we don't need to protect any sensitive location visited by the user, whereas in the second case an attacker cannot use any quasi-identifier place sequence to discover the sensitive places visited by the user.

There could be different ways to construct a *c-safe* dataset of semantic trajectories. In this paper, we propose a method based on generalization of stops place driven by a privacy places taxonomy. The main steps of the method are: (a) generate groups of semantic trajectories; (b) generalize the quasi-identifiers places within each group, and the sensitive places when the generalization quasi-identifiers place is not enough to get a *c-safe* dataset.

The algorithm, during the anonymization process, checks if the probability to infer sensitive places is less than c, therefore we define how to compute this probability. Given a sequence of sensitive places $S = s_1, \ldots, s_h$ (where each s_i is either a leaf or an internal node of the privacy place taxonomy) and a quasi-identifier sequence S_Q the probability to infer S is the conditional probability, so $P(S_Q, S) = P(S|S_Q)$ that in our case is computed as follows:

$$\frac{supp_{ST^*}(S_Q, S)}{supp_{ST^*}(S_Q) + supp_{ST^*}(S_Q, S) \times (\prod_{\forall s_i \in S} places(s_i) - 1)}$$

where $places(s_i)$ denotes the number of places represented by s_i: this number is equal to 1 when s_i is a leaf of the privacy place taxonomy; when s_i is an internal node (a generalized concept) $places(s_i)$ is equal to the number of leaves of the sub-tree with root s_i.

Our algorithm, described in the next section, guarantees that for each sensitive place $s_i \in S$ $P(s_i|S_Q) \leq c$. It is straightforward to derive that guaranteeing that $\forall s_i \in S$ $P(s_i|S_Q) \leq c$ then we also guarantee that $P(S|S_Q) \leq c$.

6. CAST ALGORITHM

We now tackle the problem of constructing a *c-safe* version of dataset ST of semantic trajectories. The algorithm called CAST (C-safe Anonymization of Semantic Trajectories) should find the best grouping in the dataset which guarantees the c-safety, but this problem is computationally hard. For this reason the first implementation of the method consider an additional parameter m which assert the size of the groups to be found in which the *c-safety* must be guarantee. The pseudo-code of the algorithm follows:

```
INPUT <Dataset D, Probability c,
       Grouping m, Taxonomy T>
BEGIN
St:= BuildSemanticTrajectories(D,T);
Ordering(St);
minLength = St.minLength;
maxLength = St.maxlength;
R = 0;
For (i:=maxLength; i<=minLength; i--)
  CurSt:= ExtractSubset(St, i);
  St:= St-CurSt;
  While(CurSt.size>m)
    G:=FindBestGroup(St,m,c,T);
    R:= R + Generalize(G);
    CurSt:= CurSt-G;
  St:= St+Cut(CurSt);
END;
OUTPUT <Result R>
```

$Ordering(St)$ is the function which removes from the semantic trajectories the sensitive stops and orders the obtained semantic trajectories by length. $ExtractSubset(St, i)$ is the function that extracts the semantic trajectories obtained at the previous step having length i. The function $FindBestGroup(St, m, c, T)$ finds the groups of semantic trajectories of the same length which minimize the distance in the taxonomy between them (this distance measure will be detailed later in section 7). These three methods implement the step (a) described in previous section. The $Generalize(G)$ method generalizes the quasi-identifiers of the semantic trajectories to obtain the identical sequences (and generalizes the sensible places when needed) to guarantee the *c-safety*, thus realizing the step (b). The following example shows the generalization step. Consider the taxonomy presented in Fig.2 and a group G formed by three semantic trajectories:

$$S_1, R_2, H_1, R_1, C_4, S_2$$
$$S_3, D_1, R_1, C_4, S_2$$
$$S_1, P_3, C_3, D_2, S_2$$

Let us assume *c=0.45* thus we want to guarantee 0.45-safety. First of all, the algorithm generalize the quasi-identifiers of the semantic trajectories in order to obtain all identical trajectories. To do this, the algorithm removes temporarily the sensitive places and finds the minimal ancestor in the taxonomy of each item of the semantic trajectories in the corresponding position, thus obtaining:

$$Station, Place, Entertainment, S_2\ (H_1, C_4)$$
$$Station, Place, Entertainment, S_2\ (C_4)$$
$$Station, Place, Entertainment, S_2\ (C_3)$$

Therefore, the algorithm computes the probability of crack defined, in section 5, for each sensible places: $P(S_Q, H_1) =$

1/3, $P(S_Q, C_4) = 2/3$ and $P(S_Q, C_3) = 1/3$ where S_Q is $\langle Station, Place, Entertainment, S_2\rangle$. In this example, only the probability of item C_4 is higher then the *c-safety* threshold, therefore we need to generalize to the higher representation level in the taxonomy: *Clinic*. Considering having only two clinics leaves in the taxonomy, after the generalization the probability of C_4 become $2/5$ which is below our threshold and so it is safe. Then the algorithm rebuilds the semantic trajectories restoring the sensitive places in their original positions:

$$\begin{gathered} Station, Place, H_1, Entertainment, Clinic, S_2 \\ Station, Place, Entertainment, Clinic, S_2 \\ Station, Place, Clinic, Entertainment, S_2 \end{gathered}$$

As a general rule, when an item is generalized, all the other items having the same parent in the taxonomy are generalized too. For this reason, C_3 is replaced with *Clinic*: having both items can give to the attacker the opportunity to infer that *Clinic* stand for C_4.

7. EXPERIMENTS

7.1 Datasets

We run the *CAST* algorithm on a trajectories dataset as illustrated in this section. The dataset adopted for that contains trajectories of 17000 moving cars in Milan, in one week, collected through GPS devices. This is a sample of data donated by a private company devoted to collect them as a service for insurance companies and other clients (Fig.3).

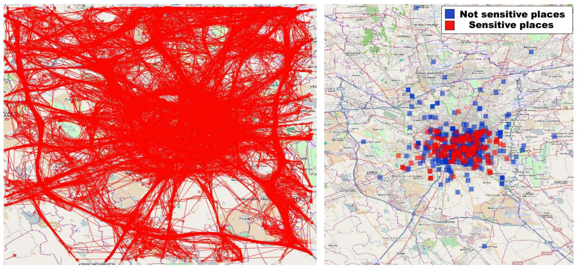

Figure 3: Trajectories dataset and the places used to build the semantic trajectories divided into not-sensitive and sensible

From this initial dataset we have computed for each trajectory the sequence of stops building the semantic trajectories as described in Section 3. The set of places has been downloaded from *Google Earth* and are shown in Fig.3 (bottom). Most of the places are in the city center of Milan and the places considered **sensitive** are represented by the red squares, the others in blue squares. The taxonomy used in the experiment is in represented in Fig.2, where the red dotted box represent the subtree considered sensitive. After the building step of the algorithm we obtain a dataset of 6225 semantic trajectories with an average length equal to 5.2 stops.

7.2 Measuring the quality of the sequential pattern results

Once obtained the *c-safe* semantic trajectory dataset, we check the data usefulness after the application of privacy-preserving framework. In this section we introduce two measures in order to verify the preserved sequential patterns among the obtained anonymous dataset. These measures are the *coverage coefficient* and the *distance coefficient*.

Definition 9 (Covered predicate). *Given two sequences $P^i = p^i_1 \dots p^i_n$ and $P^j = p^j_1 \dots p^j_m$ and a taxonomy $PTax := \langle C, HC, \lambda\rangle$, the predicate is defined as:*

$$\begin{aligned} &covered(P^i, P^j, PTax) = \\ &(|P^i| = |P^j|) \wedge \forall_{k=1\dots n}(p^i_k = p^j_k \vee HC(p^i_k, p^j_k)) \end{aligned}$$

Definition 10 (Coverage coefficient). *The coverage coefficient is a $[0, 1]$ value defined on two sets of sequences:*

$$coverage(PS_{orig}, PS_{anon}, PTax) = \frac{|CovSet(...)|}{|PS_{orig}|}$$

where

$$\begin{aligned} &CovSet(PS_{orig}, PS_{anon}, PTax) = \\ &\{\{P^i\} | P^i \in PS_{orig} \wedge \exists P^j \in PS_{anon}, covered(P^i, P^j, PTax)\} \end{aligned}$$

Intuitively, the coverage coefficient measures how many patterns extracted from the original dataset are covered at least by the patterns extracted in the anonymized dataset with a certain level of generalization. It's important to notice that the coverage does not measures how much the patterns are generalized, but only if they are covered by a pattern obtained from the anonymized dataset. This means that a pattern composed by item generalized to the root of the taxonomy (i.e. $\langle Place\ Place\ Place\rangle$) will cover all the pattern with the same length. To face this problem, we have defined three levels of coverage:

- *Coverage upper bound*: Considers all the patterns discovered PS^{ub}_{anon} including all the patterns extracted from the anonymized dataset.
- *Coverage*: Considers a subset of the extracted patterns PS_{anon} which not consider the patterns composed only by root item (*Place*).
- *Coverage lower bound*: Considers only patterns PS^{lb}_{anon} which does not contain any root items.

The aim of these three level is to better describe which kind of generalization is performed and the consequences on the patterns.

The other measure we will use is the *distance coefficient* which represent the distance in terms of steps in the taxonomy to transform the patterns from the set extracted on the original dataset and the one from the anonymized dataset. The coefficient is normalized on the maximum possible distance of the two sets.

Definition 11 (Sequence distance). *Given two sequences $P^i = p^i_1 \dots p^i_n$ and $P^j = p^j_1 \dots p^j_m$ and a taxonomy $PTax := \langle C, HC, \lambda\rangle$, the sequence distance is:*

$$SeqDis = \frac{Hops(P^i, P^j, PTax)}{MaxDeep(PTax) * 2 * Length(P^i)}$$

where $Hops(...)$ is the number of steps needed to transform the pattern P^i in the pattern P^j on the taxonomy $PTax$;

$MaxDeep(PTax)$ is the maximum depth of the taxonomy tree and $Length(P^i)$ is the number of items of the pattern. The Sequence distance is defined only between patterns of the same length, in the other cases the distance is 1.

Definition 12 (Distance coefficient). *The distance coefficient is a $[0, 1]$ value defined on two sets of sequences:*

$$Dis(PS_{orig}, PS_{anon}, PTax) = \frac{\sum_{P^i \in PS_{orig}} ArgMin_{P^j \in PS_{anon}}(SeqDis(P^i, P^j, PTax))}{|PS_{orig}|}$$

where PS_{orig} and PS_{anon} are the sets of patterns extracted from the original dataset and the anonymized dataset.

This measure shows the other aspect of the transformation applied on the patterns which is not highlighted by the coverage coefficient. In the next section we study how these two measures varies with different setting of the problem.

7.3 Experimental results

The experiments are performed on Milan dataset using an Intel Core 2 Duo T6400 at 2.00 Ghz. In this section we want to analyze the effects of the anonymization on the sequential patterns extracted on the datasets using the measures described above. Fig.4 shows how the coverage coefficient varies changing the support threshold used in the pattern mining algorithm. In the top of the figure we depict the number of pattern extracted both from the original dataset PS_{orig} and the anonymized one PS^{ub}_{anon}, PS_{anon}, PS^{lb}_{anon}.

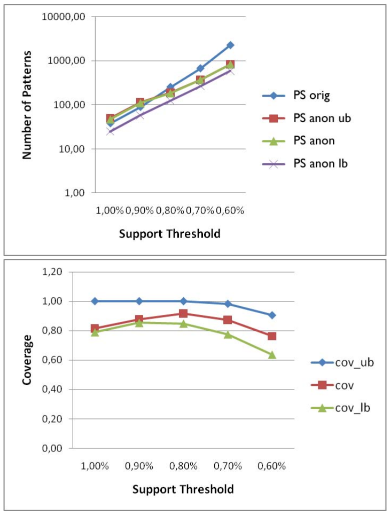

Figure 4: Number of patterns and the coverage measurement changing the support thresholds (m=10, l=.3)

We can notice that the generalization have a double effect on the patterns: (i) increases the frequency of generalized items, (ii) decreases the frequency of leaf items of the taxonomy. Therefore with an high support threshold the difference between the patterns created and the pattern removed by the generalization is positive (increasing the resulting size of the patterns set) but after a certain threshold, due to the smaller number of generalized items, the decreasing of patterns becomes predominant. The decreasing of patterns with lower support is accentuated by the cutting of semantic trajectories during the process. All this effects are highlighted by the coverage coefficients which show effectively the consequences of these behaviors.

In Fig.5 we study the coverage coefficient varying the group size and the c-safety value. In this case we can see that the variations are not so evident, since only the effect of the reduction of semantic trajectories change the coefficient. In other words we can say that the level of coverage is almost the same not considering the level of generalization. In the other hand, the Cov^{lb} give us another hint of how the patterns become: at the beginning, with a group size between 5 and 20, they are generalized but they don't contain root items. When the group size exceeds this limit the patterns contain at least a root item therefore the lower bound of the coverage fall.

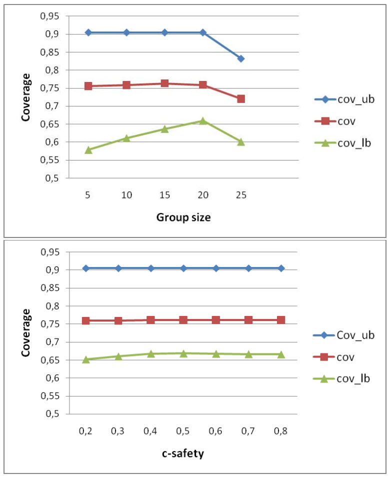

Figure 5: Coverage measurement changing the group sizes and c-safety (support threshold = 30)

As the coverage coefficient we can study the distance coefficient varying the other parameters. In Fig.6 we can see that the distance coefficient between the PS_{orig} and PS_{anon} increase with smaller support.

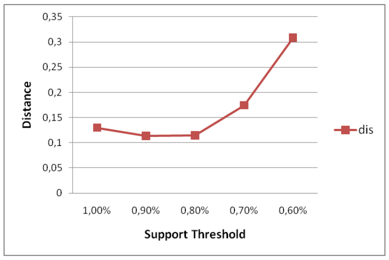

Figure 6: The distance coefficient changing the support thresholds (m=10, l.3)

Furthermore in Fig.7, when the group size or c-safety value increase the distance coefficient decrease. This is a clear effect of the generalization of the anonymized patterns which become more distant from the original set of pattern.

8. CONCLUSION AND FUTURE WORKS

In this paper we have investigated the problem of publishing semantic trajectories datasets while preserving the privacy of users. Here, the focus is on the semantics of the places visited distinguishing between sensitive places and

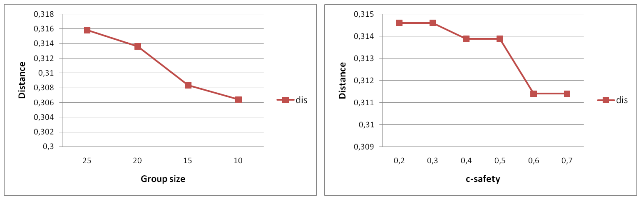

Figure 7: The distance coefficient changing the group sizes and c-safety (support threshold = 30)

quasi-identifier places. The introduced method exploits a places taxonomy to generalize the visited places to obtain a *c-safe* dataset. The use of a taxonomy encoding domain knowledge about the places tends to perform a generalization that preserves semantics of the trajectories. Through a set of experiments on a real-life spatio-temporal dataset, we have shown that our method, while guaranteeing a good protection, it also preserves the quality of the sequential pattern analysis. We are currently performing new statistical analysis on the dataset, in order to understand how the data properties are preserved after the anonymization. Further research includes the experimentation of new pattern mining methods on the anonymized trajectories. We will also investigate improved approaches to generate a *c-safe* version of a dataset of semantic trajectories, such as an algorithm that does not consider only groups of a fixed size. Another future research direction goes towards the exploitation of c-safe semantic trajectories dataset for semantic tagging of trajectories. How does the anonymization step affect the overall results of a trajectory semantic tagging inference? We believe that since the taxonomy tends to preserve semantics, the current approach should preserve some degree of semantics in the trajectory understanding and behavior classification.

9. ACKNOWLEDGMENTS

This work has been partially supported by MODAP FET CA (http://www.modap.org). Chiara Renso acknowledges support from CNR Short Term Mobility program and Dino Pedreschi acknowledges support by Google, under the Google Research Award program.

10. REFERENCES

[1] O. Abul, F. Bonchi, and M. Nanni. Never walk alone: Uncertainty for anonymity in moving objects databases. In *Int. Conf. on Data Engineering*, 2008.

[2] R. Agrawal and R. Srikant. Privacy-preserving data mining. In *SIGMOD*, pages 439–450. ACM, 2000.

[3] L. O. Alvares , V. Bogorny, B. Kuijpers, J. A. F. de Macedo , B. Moelans, and A. Vaisman. A model for enriching trajectories with semantic geographical information. In ACM-GIS, 2007.

[4] O. Abul, F. Bonchi, and F. Giannotti. Hiding Sequential and Spatio-temporal Patterns. *The TKDE Journal*, 2008.

[5] V. Bogorny and M. Wachowicz. A Framework for Context-Aware Trajectory Data Mining. Data Mining for Business Applications, Springer, 2008.

[6] M. L. Damiani, E. Bertino, C. Silvestri. The PROBE Framework for the Personalized Cloaking of Private Locations. In *TDP*, 3:2 (2010) 91 - 121.

[7] Gruber. T.R. (2008) Ontology. Entry in the Encyclopedia of Database Systems, Ling Liu and M. Tamer zsu (Eds.), Springer-Verlag.

[8] M. Gruteser and D. Grunwald. A methodological assessment of location privacy risks in wireless hotspot networks. In *First Int. Conf. on Security in Pervasive Computing*, 2003.

[9] Y. He and J. F. Naughton. Anonymization of Set-Valued Data via Top-Down, Local Generalization. In *PVLDB*, 2009.

[10] N. Li, T. Li, and S. Venkatasubramanian. t-closeness: Privacy beyond k-anonymity and l-diversity. In *Int. Conf. on Data Engineering*. IEEE, 2007.

[11] A. Machanavajjhala, J. Gehrke, D. Kifer, and M. Venkitasubramaniam. l-diversity: Privacy beyond k-anonymity. In *Int. Conference on Data Engineering*. IEEE, 2006.

[12] A. Meyerson and R. Williams. On the complexity of optimal k-anonymity. In *PODS '04*. ACM, 2004.

[13] M. F. Mokbel, C. Chow, and W. G. Aref. The new casper: Query processing for location services without compromising privacy. In *VLDB*, 2006.

[14] M. F. Mokbel, C. Chow, and W. G. Aref. The new casper: A privacy-aware location-based database server. In *Int. Conference on Data Engineering*, IEEE 2007.

[15] A. Monreale, G.Andrienko, N.Andrienko, F.Giannotti, D.Pedreschi, S. Rinzivillo, S. Wrobel. Movement Data Anonymity through Generalization. *Transactions on Data Privacy* 3:2 (2010) pp. 91 - 121,

[16] M. E. Nergiz, M. Atzori, and Y. Saygin. Perturbation-driven anonymization of trajectories. Technical Report 2007-TR-017, ISTI-CNR, Pisa, 2007.

[17] A.T. Palma, V. Bogorny, B. Kuijpers, and L. O. Alvares. A clustering-based approach for discovering interesting places in trajectories. In ACM-SAC, 2008.

[18] R. G. Pensa, A. Monreale, F. Pinelli, and D. Pedreschi. Pattern-preserving k-anonymization of sequences and its application to mobility data mining. In *Int. Workshop on Privacy in Location-Based Applications - PiLBA '08*, 2008.

[19] P. Samarati and L. Sweeney. Protecting privacy when disclosing information: k-anonymity and its enforcement through generalization and suppression. Technical report, SRI International, 1998.

[20] S. Spaccapietra, C. Parent M.L. Damiani, J. Macedo, F. Porto, C. Vangenot. A conceptual view on trajectories. DKE Journal 65(1): 126-146 (2008).

[21] L. Sweeney. Uniqueness of Simple Demographics in the U.S. Population, LIDAPWP4. Carnegie Mellon University, Laboratory for International Data Privacy, 2000. The Identifiability of Data.

[22] M. Terrovitis and N. Mamoulis. Privacy preservation in the publication of trajectories. In *Int. Conf. On Mobile Data Management*, 2008.

[23] A. Valls, C. Gómez-Alonso and V. Torra Generation of Prototypes for Masking Sequences of Events. In *Int. Conf. on Availability, Reliability and Security*, 2009.

[24] R. Yarovoy, F. Bonchi, L. V. S. Lakshmanan, and W. H. Wang. Anonymizing moving objects: how to hide a mob in a crowd? In *EDBT*, 2009.