TRANSACTIONS ON DATA PRIVACY 4 (2011) 73–101

C-safety: a framework for the anonymization of semantic trajectories

Anna Monreale*, **Roberto Trasarti****, **Dino Pedreschi***, **Chiara Renso****, **Vania Bogorny*****

*KDD-Lab, University of Pisa, Italy.

**KDD-Lab, ISTI CNR, Pisa, Italy.

***UFSC, Florianopolis, SC, Brazil.

E-mail: {annam,pedre}@di.unipi.it,{roberto.trasarti,chiara.renso}@isti.cnr.it, vania@inf.ufsc.br

Abstract. The increasing abundance of data about the trajectories of personal movement is opening new opportunities for analyzing and mining human mobility. However, 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 great privacy threats. In this paper we propose a privacy model defining the attack model of semantic trajectory linking and a privacy notion, called *c-safety* based on a generalization of visited places based on a taxonomy. 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 visited 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 two real-life GPS trajectory datasets to show how our algorithm preserves interesting quality/utility measures of the original trajectories, when mining semantic trajectories sequential pattern mining results. We also empirically measure how the probability that the attacker's inference succeeds is much lower than the theoretical upper bound established.

Keywords. Semantic trajectories, anonymization of trajectories datasets, taxonomy, sequential patterns

1 Introduction

The increasing abundance of data about the trajectories of personal movement, obtained through mobile phones and other location-aware devices that accompany our daily routines, is opening up new avenues for analyzing and mining human mobility, with applications in diverse scientific and social domains, ranging from urban planning to transportation management to epidemic prevention. 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 [25, 2, 28, 1]. This problem, however, becomes ever more challenging as the fast progress on mobile device technology, geographic information and mobility data analysis and mining creates

entirely new forms of trajectory data, with far richer semantic information attached to the traces of personal mobility. In fact, 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 is called in the literature *semantic trajectories*, i.e., sequences of stops and moves of a person during her/his movement [30]. In semantic trajectories each location of stop can be attached to some contextual information such as the visited place or the purpose - either by explicit sensing or by inference. An example of semantic trajectory is the sequence of places visited by a moving individual such as *Supermarket, Restaurant, Gym, Hospital, Museum*.

In this paper, we argue that the new form of data of semantic trajectories poses important privacy threats. In fact, by definition the semantic trajectories represent the most important places visited by a person and therefore we explicitly represent the person interests in the movement data itself. 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 (that person's health status, in the example). A place is considered *sensitive* if it allows to infer personal sensitive information of the tracked individual.

However, just hiding a person's trajectory into a crowd, following the idea of k-anonymity, is not enough for a robust protection. When all individuals in a crowd with indistinguishable trajectories visit the same sensitive place, an attacker just needs to know that the person belongs to that crowd to infer that he/she visited the sensitive place.

This problem reflects 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 nonsensitive locations corresponding to the places where the persons stop. We represent this background knowledge through a place taxonomy, describing sensitive and non-sensitive locations at different levels of abstraction (e.g., a touristic landmark, a museum, the Louvre museum, a health-related service, a hospital, the Children 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 the place taxonomy, thus providing a way to preserve the semantics of the generalized trajectories. We conduct a study on two real-life GPS trajectory datasets, 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. Also, we empirically measure how the probability that the attacker's inference succeeds is much lower than the theoretical upper bound established, thus bringing further evidence that our method achieves an interesting trade-off between privacy and utility.

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

works.

2 Background

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

2.1 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 [30] 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 [6]. 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 $t_0 < t_1 < t_2 < \ldots < t_N$.

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).

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 [5] 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) [27] 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

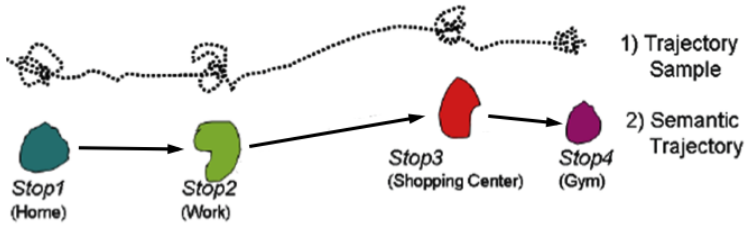


Figure 1: Example of Trajectory Sample and Semantic Trajectory

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 they are not associated to any interesting place. Another work has been proposed by Manso et al. [18] which developed an algorithm to compute stops based on the variation of the direction of the trajectory.

2.2 Domain Ontologies and Taxonomies

An ontology is defined in [11] 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". An ontology determines 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. The basic elements of an ontology 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 satisfy 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 semantic hierarchy of geographical places of interest. Here, the set of *stop places* obtained from the computation of semantic trajectories are the *leaves* of Tax. For example, we have that *Restaurant "Da Mario"* is a *kind of* Restaurant which is a *kind of* Entertainment. In general, each concept in the taxonomy describes the semantic categories of the geographical objects of interest for a given application domain. We will see in the remaining of the paper that the place taxonomy is essential during the generalization phase in the anonymization algorithm. Figure 2 depicts an example of the taxonomy of places of interest in a city used as example through the paper. Here, all the places represented by red nodes are sensitive locations, as discussed later.

It's important to notice that the taxonomy is global and the leaves represent the public points of interest defined by a domain expert and considered *relevant* w.r.t. a specific application. As a consequence, the private places such as the *homes of the users* are not represented. Moreover, we consider also the special case of points of interests which can be considered private for a user such as *his/her workplace*. We solve the problem by suppressing

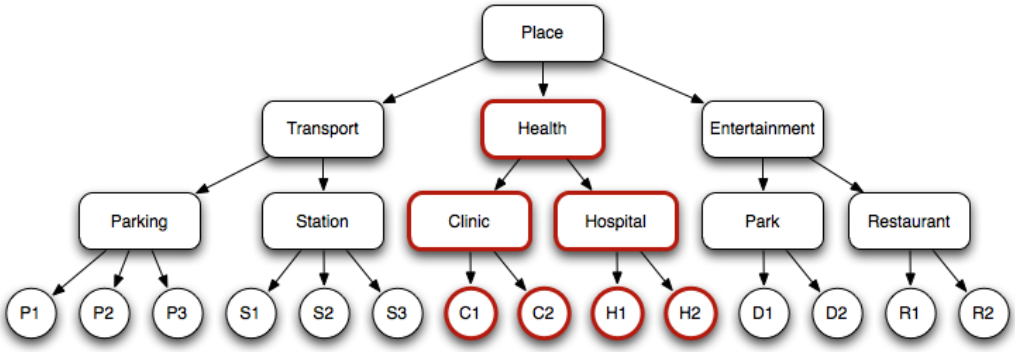


Figure 2: The Places taxonomy

these places in the user trajectories (see Section 5).

3 Privacy Model

In this paper we provide a framework that, given a dataset of semantic trajectories, generates an *anonymous semantic trajectory dataset*. This new dataset guarantees that 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 we propose is based on the generalization, driven by the place taxonomy, of the places visited by a user. The use of a domain taxonomy to generalize places allows to preserve some degree of 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 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 their intrinsic geographical nature, 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 places, with respect to some applications, can be sensitive because an attacker can derive personal sensitive information. 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. In the relational model there are three categories of attributes: quasi-identifiers, sensitive and non-confidential. But it is often found in the literature a simple and conservative approach which considers all the non-sensitive attributes as quasi-identifiers. In the present work we apply this simplification for the places. The place taxonomy is used to set this distinction: some concepts are tagged as "sensitive" and, as a consequence, all the remaining concepts are "quasi-identifiers". We assume the labeled 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 $\mathcal{ST}$ and the privacy places taxonomy $PTax$ describing the categories of the geographical objects of interest for an application domain, we define a framework to transform $\mathcal{ST}$ in its anonymous version $\mathcal{ST}^*$ 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* $S_Q = q_1, \ldots, q_h$, where $h > 0$ 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.

The semantic knowledge associated to the dataset allows the classification of the set of places in sensitive and quasi-identifier places: this clear distinction is not possible without additional semantic knowledge. Of course this new setting requires the definition of a suitable privacy model that takes into account this additional information.

The taxonomy can be used to transform a semantic trajectory into its generalized version. Intuitively, the stop places may be replaced by one of their ancestors in the taxonomy (e.g. "Restaurant da Mario" may be replaced by "Restaurant" or by "Entertainment"). Obviously, the generalized version looses the specific information of the original semantic trajectory, but tends to preserve some level of semantics (e.g. we know a person stopped at a restaurant but we don't know exactly which one). The formal definition of a generalized semantic trajectory is given below.

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.

The containment concept, defined below, expresses the basic relation to compare two semantic trajectories even if some of places are generalized.

Definition 5 (Contained)**.** 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 ($A \preceq_g T_g$) 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.

Definition 6 (Set Contained)**.** Let $T_g = g_1, g_2 \ldots, g_n$ be a generalized semantic trajectory and $A = \{p_1, \ldots, p_m\}$ a set of places with $m \leq n$. We say that all the places in *A are contained* in T_g ($A \sqsubseteq T_g$) if there exists a set B of m places appearing in T_g and $\forall g_i \in B \; \exists \, p_j \in A$ such that either $g_i = p_j$ or the relation $HC(p_j, g_i)$ holds.

Notice that the difference between Definition 5 and Definition 6 is that the second one does not take into account the order of the places of A in the generalized semantic trajectory T_g. We refer to the number of generalized semantic trajectories in $\mathcal{ST}^*$ containing a sequence of places A as *support of A* denoted by $supp_{\mathcal{ST}^*}(A)$. More formally, $supp_{\mathcal{ST}^*}(A) = |\{T_g \in \mathcal{ST}^* | A \preceq_g T_g\}|$.

3.1 Adversary Knowledge and Attack Model

An intruder who gains access to $\mathcal{ST}^*$ may possess some background knowledge allowing he/she to conduct attacks making 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 7 (Adversary Knowledge)**.** The attacker has access to the generalized dataset $\mathcal{ST}^*$ and knows: (a) the algorithm used to anonymize the data, (b) the privacy place taxonomy *PTax* , (c) that a given user is in the dataset and (d) a quasi-identifier place sequence S_Q visited by the given user.

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 visited sensitive places.

Definition 8 (Attack Model)**.** The attacker, given a published semantic trajectory dataset $\mathcal{ST}^*$ where each trajectory is uniquely associated to a de-identified respondent, tries to identify the semantic trajectory in $\mathcal{ST}^*$ associated to a given respondent U, based on the additional knowledge introduced in Definition 7. The attacker, given the quasi-identifier sequence S_Q constructs a set of candidate semantic trajectories in $\mathcal{ST}^*$ containing S_Q and tries to infer the sensitive leaf places related to U. We denote by $Prob(S_Q, S)$ the probability that, given a quasi-identifier place sequence S_Q related to a user U, the attacker infers his/her set of sensitive places S which are the leaves of the taxonomy $PTax$.

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 9 (C-Safety)**.** The dataset $\mathcal{ST}$ is defined *c-safe* with respect to the place set Q if for every quasi-identifier place sequence S_Q, we have that for each set of sensitive places S the $Prob(S_Q, S) \leq c$ with $c \in [0, 1]$.

It is crucial to understand that in this scenario the focus is on the *real place instances* (the leaves of the taxonomy) which represent the most sensitive information. Therefore, the inner nodes of the taxonomy are used to generalize the semantic trajectory in order to reduce the probability for the attacker to identify the real places.

4 Problem Statement

Given the definitions introduced in the previous section, we formulate the problem statement as follows.

Definition 10 (Problem Statement)**.** Given a dataset $\mathcal{ST}$ of semantic trajectories and a protection probability threshold that we want to guarantee $c \in [0, 1]$, we want to build a c-safe version $\mathcal{ST}^*$ of $\mathcal{ST}$.

In other words, we want to fix a threshold at the probability that an adversary who access $\mathcal{ST}^*$ can use a sequence of quasi-identifier places visited by a user to correctly infer any visited sensitive place. The problem that we want to address is very similar to l-diversity [17], but the particular nature of the data makes the problem different and this framework

cannot be directly applied to this case. First of all, the semantic trajectories do not have fixed length. In particular, given any two semantic trajectories T_i and T_j belonging to the same database, the number of quasi-identifiers stop places 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 a-priori for all trajectories. It it worth pointing out that a trajectory may be composed only by quasi-identifier places or only by sensitive places. However, these two cases do not introduce further privacy issues: 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.

It is interesting to notice that a place tagged in the privacy place taxonomy as sensitive by the domain expert could be for some user a quasi-identifier. For instance, a medical center is a sensitive place for patients, whereas is a quasi-identifier for doctors and nurses. This may lead to privacy flows, therefore we choose to explicitly suppress from the trajectories the places which are tagged in the taxonomy as sensitive, but are, in fact, quasi identifiers for the user.

There could be different ways to construct a *c-safe* dataset of semantic trajectories. We propose a method that is based on a generalization of places, driven by the privacy place taxonomy, which preserves some data utility. The main steps of the method can be summarized as follows: [a] suppressing from the dataset each sensitive place when, for that given user, that place is a quasi-identifier; [b] grouping semantic trajectories in groups of a predefined size, *m*; [c] building a generalized version of each semantic trajectory in the group generalizing the quasi-identifier places. In each group the quasi-identifiers of the generalized trajectories should be identical. Sensitive places are generalized when the quasi-identifiers generalization is not enough to get a *c-safe* dataset.

This method generates a *c-safe* version of a dataset of semantic trajectories keeping under control both the probability to infer sensitive places and the generalization level (thus the information loss) introduced in the data. In other words, the obtained dataset guarantees the *c-safety* and maintains the information useful for the data mining tasks, as much as possible. The taxonomy defined by the domain expert is crucial in this process. In fact, having more levels of abstraction allows the method in finding a better generalization in terms of information loss (see Section 5). Before going into details of the algorithm, we first discuss two crucial aspects: how to compute the probability to disclose sensitive information and how to measure the cost, in terms of information loss, of the anonymization.

4.1 Disclosure Probability

During the anonymization process, the algorithm checks if the probability to infer sensitive places is less than the fixed threshold c. In the following, we give the basic definitions to compute this probability. Given a set of sensitive places $S = \{s_1, \ldots, s_h\}$ ($S \subseteq SP$), where each s_i is a leaf node of the privacy place taxonomy, and a quasi-identifier sequence S_Q, the probability to infer S is the *conditional probability* $P(S_Q, S) = P(S|S_Q)$. This probability is computed as follows:

$$P(S_Q, S) = \frac{1}{supp_{\mathcal{ST}^*}(S_Q)} \times \sum_{\forall t \in G_{S_Q}} P_t(S)$$

where G_{S_Q} is the set of (generalized) semantic trajectories which support the quasi-identifier sequence S_Q and $P_t(S)$ is the probability to infer the set of places in S in the trajectory t. In

the following, given the set $S' = \{s_i \in S | \forall j \neq i . place_t(s_i) \neq place_t(s_j) \wedge s_j \in S\}$, we define:

$$P_t(S) = \begin{cases} 0 & \text{if } S \sqsubseteq t \\ \frac{v}{\prod_{\forall s_i \in S'} leaves(place_t(s_i))^{o(place_t(s_i))}} & \text{otherwise} \end{cases}$$

where:

- $place_t(s_i)$ is equal to s_i if the trajectory t contains the leaf place s_i otherwise it is equal to the generalized concept of the taxonomy (inner node) used to generalize s_i in the semantic trajectory t.
- $leaves(place_t(s_i))$ denotes the number of places *represented* in the taxonomy by the specific node returned by $place_t(s_i)$. More in detail, we define $leaves(place_t(s_i))$ equal to 1 when $place_t(s_i)$ is a leaf of the privacy place taxonomy, whereas when $place_t(s_i)$ is an internal node (a generalized concept) it is equal to the number of leaves of the sub-tree with root $place_t(s_i)$.
- $o(place_t(s_i))$ is the number of occurrences of $place_t(s_i)$ in the semantic trajectory t.
- v is the number of combinations of sensitive places that can be represented in the semantic trajectory t and which support the set S.

The proposed CAST algorithm, introduced in the following, guarantees that for each sensitive place $s_i \in S$ $P(s_i|S_Q) \leq c$. In Section 5.1 we will show that this also guarantees $P(S|S_Q) \leq c$.

4.2 Data Utility

The main aim of the privacy-preserving data publication techniques is to protect respondents' individual privacy. However, the recipients of the transformed dataset ST^* could be interested in applying different analytical tools to extract from data some useful knowledge. As an example, they could apply some data mining tools to extract common and frequent human behaviors. Therefore it is essential that anonymization techniques should be able to maintain a good quality of the data. For this reason, we introduce a method that combines anonymization process with data utility. Precisely, our interest is to preserve as much as possible the sequential pattern mining results. Therefore, when our method chooses the grouping of trajectories to be generalized it tries to minimize the cost of generalization, i.e., the cost to transform the original trajectories belonging to the same group into the generalized ones. The question is: "how can we express this cost?" We decided to measure this cost using two different distance functions that measure the cost to transform an original semantic trajectory into a generalized one, based on the taxonomy. These functions measure the distance in steps from two places in the taxonomy tree (Hops-based distance), and the information loss based on the number of leaves of a node (Leaves -based distance).

These measures are formally defined as follows.

Definition 11 (Hops-based Distance)**.** Let $T = p_1, p_2, \ldots p_h$, $T_g = g_1 \ldots g_h$ and *PTax* be a semantic trajectory in the original dataset $\mathcal{ST}$, a generalized semantic trajectory and a privacy place taxonomy, respectively. The *Hops-based distance* is defined as:

$$Distance_H(T, T_g, \ PTax\ ) = \frac{\sum_{1\ldots h}^{k} Hops(p_k, g_k, \ PTax\ )}{MaxDeep(\ PTax\ ) * h}$$

where $Hops(p_k, g_k,\ PTax\ )$ is the number of steps needed to transform the item p_k into g_k w.r.t. the taxonomy $PTax$; $MaxDeep(\ PTax\ )$ is the maximum depth of the taxonomy tree. The Hops-based distance is defined only between semantic trajectories of the same length. In the other cases the distance is 1.

Intuitively, this definition expresses the distance between a given semantic trajectory and a generalized one in terms of number of steps in the privacy place taxonomy needed to transform the original version of the semantic trajectory into the generalized one.

Definition 12 (Leaves-based Distance)**.** Let $T = p_1, p_2, \dots p_h$, $T_g = g_1 \dots g_h$ and $PTax$ be a semantic trajectory in the original dataset $\mathcal{ST}$, a generalized semantic trajectory and a privacy place taxonomy, respectively. The *Leaves-based distance* is defined as:

$$Distance_L(T, T_g,\ PTax\ ) = \frac{\sum_{1\dots h}^{k} Loss(p_k, g_k,\ PTax\ )}{Leaves(\ PTax\ ) * h}$$

where $Loss(p_k, g_k,\ PTax\ )$ is the information loss between two items on the taxonomy $PTax$. This is quantified as the difference between the number of leaves of the subtree of g_k and the number of leaves of the subtree of p_k. $Leaves(\ PTax\ )$ is the total number of leaves in the taxonomy. The Leaves-based distance is defined only between semantic trajectories of the same length. In the other cases the distance is 1.

The distance defined above allows to measure the distance between an original semantic trajectory and a generalized one taking into account that some concepts in the taxonomy (internal nodes) generalize more than other having more leafs in its subtree. Considering this fact is important since generalizing to a parent having in its subtree few leaves implies less information loss than generalizing to a parent having in its subtree more leaves.

In the remaining of the paper we will use the generic term *distance* to represent one of the two distances defined above.

The term *generalization cost* indicates the distance between a set of semantic trajectories and its generalized version:

Definition 13 (Generalization Cost)**.** Let $\mathcal{T}$, $\mathcal{T}^*$ and $PTax$ be a set of semantic trajectories of a dataset $\mathcal{ST}$, the generalized version of $\mathcal{T}$ and a privacy place taxonomy, respectively.

The *generalization cost* is a value in the interval $[0, 1]$ defined as:

$$GenCost(\mathcal{T}, \mathcal{T}^*,\ PTax\ ) = \frac{\sum_{\forall T_i \in \mathcal{T} \wedge T_i^* \in \mathcal{T}^*} distance(T_i, T_i^*,\ PTax\ )}{|\mathcal{T}|}$$

where each sequence T_i^* is the generalized version of the semantic trajectory T_i.

5 CAST Algorithm

We now tackle the problem of constructing a *c-safe* version of dataset $\mathcal{ST}$ of semantic trajectories. The algorithm, called CAST (C-safe Anonymization of Semantic Trajectories), finds the best trajectory grouping in the dataset which guarantees the c-safety. However, since this problem is computationally hard, the implementation of the method considers an additional parameter m indicating the size of the trajectory groups in which the *c-safety* must be guaranteed. The pseudo-code of CAST is provided in Algorithm 1.

The first step of the algorithm calls the function $SuppressFalseSP(\mathcal{ST},\ PTax\ )$ which verifies automatically if a sensitive place of the privacy taxonomy $PTax$ is a quasi-identifier

Algorithm 1: CAST($\mathcal{ST}$, c, m, $P\text{-}Tax$)

Input: A semantic trajectory database $\mathcal{ST}$, a probability c, a group size m, a privacy place taxonomy *PTax*
Output: A c-safe semantic trajectory database $\mathcal{ST}^*$

```
ST' = SuppressFalseSP(ST, PTax );
Ordering(ST');
minLength = minLength(ST');
maxLength = maxlength(ST');
ST* = ∅;
for minLength ≤ i ≤ maxlength do
    CurSeq = ExtractSubSet(ST', i);
    ST' = ST' \ CurSeq;
    while CurSeq.size > m do
        G = FindBestGroup(CurSeq, m, c, PTax );
        ST* = ST* ∪ Generalize(G);
        CurSeq = CurSeq − G;
    ST' = ST' ∪ Cut(CurSeq);
return ST*
```

for some user. In this case, the place is suppressed from the sequence of places visited by that particular user. We recall that a sensitive place can be a quasi identifier when a place, tagged as sensitive in the taxonomy, is the work place of the user. To automatically detect these particular cases we use the system described in [24]. This system analyses the trajectories of an individual to identify the most interesting visited places, such as his/her place of work and inferring the trajectory behaviour. This process realizes the step [a] of the method described in Section 4. $Ordering(\mathcal{ST}')$ is the function which temporarily removes from the semantic trajectories the sensitive places. The aim is to obtain sequences of places containing only quasi-identifiers sorting the resulting semantic trajectories by length. $ExtractSubSet(\mathcal{ST}', i)$ function extracts the semantic trajectories obtained at the previous step having length *i*. The function $FindBestGroup(CurSeq, m, c,$ *PTax* $)$ finds the groups of semantic trajectories of the same length which minimize the *generalization cost* (Definition 13). These three methods implement the step [b] described above. The method $Generalize(G)$ generalizes the quasi-identifiers of the trajectories within a group to obtain the identical sequences of places (and generalizes the sensitive places when needed) to guarantee the *c-safety*, thus realizing the step [c].

The following example shows the generalization step using the *Hops-based Distance* introduced in the Definition 11. Consider the taxonomy presented in Figure 2 and a group G composed of the following three semantic trajectories (therefore the group size is set to $m = 3$):

$$T_1 = S_1, R_2, H_1, R_1, C_2, S_2$$
$$T_2 = S_3, D_1, R_1, C_2, S_2$$
$$T_3 = S_1, P_3, C_1, D_2, S_2$$

Let us assume that we want to guarantee 0.45-safety (*c=0.45*). First of all, the algorithm generalizes the quasi-identifiers of the semantic trajectories in order to obtain trajectories with identical sequences. To do this, the algorithm removes temporarily the sensitive places

and finds the minimal ancestor in the taxonomy of each place of the semantic trajectories, thus obtaining:

$T_1 = Station, Place, Entertainment, S_2\ (H_1, C_2)$
$T_2 = Station, Place, Entertainment, S_2\ (C_2)$
$T_3 = Station, Place, Entertainment, S_2\ (C_1)$

At this point, the algorithm computes the disclosure probability (defined in Section 4.1) for each sensitive place: $P(S_Q, H_1) = 1/3$, $P(S_Q, C_2) = 2/3$ and $P(S_Q, C_1) = 1/3$. Note that in this case S_Q is the sequence of places $\langle Station, Place, Entertainment, S_2\rangle$. In this example, only the probability of place C_2 is higher than the *c-safety* threshold, therefore we need to generalize it to a higher representation level in the taxonomy which is $Clinic$. Considering that we have only two clinics leaves in the taxonomy, after the generalization the probability of C_2 becomes $2/5$ which is below the threshold, and therefore 0.45-safe. After that, the algorithm rebuilds the semantic trajectories restoring the sensitive places in their original positions:

$T_1 = Station, Place, H_1, Entertainment, Clinic, S_2$
$T_2 = Station, Place, Entertainment, Clinic, S_2$
$T_3 = Station, Place, Clinic, Entertainment, S_2$

As a general rule, we can notice that when a place is generalized, all the other places having the same parent in the taxonomy are generalized too. For this reason, C_1 is replaced with $Clinic$: not generalizing C_1 may lead the attacker to infer that $Clinic$ stands for C_2.

5.1 Privacy Analysis

We now discuss the privacy guarantees obtained by applying the proposed anonymization approach. We will formally show that the CAST Algorithm guarantees that the disclosed dataset $\mathcal{ST}^*$ is a $c\text{-}safe$ version of $\mathcal{ST}$.

Lemma 14. *Let $S = \{s_1, s_2, \ldots, s_h\}$ and S_Q be a set of sensitive places and a quasi-identifier place sequence, respectively. If $\forall s_i \in S\ P(S_Q, s_i) \leq c$ then $P(S_Q, S) \leq c$.*

Proof. We assume that for any $s_i \in S$ we have $P(S_Q, s_i) \leq c$. We recall that

$$P(S_Q, s_i) = \frac{1}{supp_{\mathcal{ST}^*}(S_Q)} \times \sum_{\forall t \in G_{S_Q}} P_t(s_i)$$

where

$$P_t(s_i) = \begin{cases} 0 & \text{if } \{s_i\} \sqsubseteq t \\ \frac{v'}{leaves(place_t(s_i))^{o(place_t(s_i))}} & \text{otherwise} \end{cases}$$

Now we have to show that

$$P(S_Q, S) = \frac{1}{supp_{\mathcal{ST}^*}(S_Q)} \times \sum_{\forall t \in G_{S_Q}} P_t(S) \leq c$$

where

$$P_t(S) = \begin{cases} 0 & \text{if } S \sqsubseteq t \\ \frac{v}{\prod_{\forall s_i \in S'} leaves(place_t(s_i))^{o(place_t(s_i))}} & \text{otherwise} \end{cases}$$

First of all we show that

$$P(S_Q, S) \leq P(S_Q, s_i) \tag{1}$$

To this end we show that $\forall t \in G_{S_Q}$, i.e., for each trajectory selected by the quasi-identifier sequence S_Q, $P_t(S) \leq P_t(s_i)$.

To do this, we analyze the three possible cases:

(a) "each sensitive place $s_i \in S = \{s_1, \ldots, s_h\}$ is generalized in the semantic trajectory t with the same ancestor": this implies that in t there are $m \geq h$ occurrences of the same generalized place and so $S' = \{s_i \in S | \forall j \neq i . place_t(s_i) \neq place_t(s_j) \wedge s_j \in S\}$ contains one of the s_i places. In this case, we have

$$P_t(S) = \begin{cases} 0 & \text{if } \{s_i\} \sqsubseteq t \\ \frac{v}{leaves(place_t(s_i))^m} & \text{otherwise} \end{cases}$$

$$P_t(s_i) = \begin{cases} 0 & \text{if } \{s_i\} \sqsubseteq t \\ \frac{v'}{leaves(place_t(s_i))^m} & \text{otherwise} \end{cases}$$

It is easy to derive that $P_t(S) \leq P_t(s_i)$ because the denominators of the two formulas are equal and $v' \geq v$.

(b) "for each sensitive place $s_i \in S$ the value of $place_t(s_i)$ is different and in the semantic trajectory t there is only one occurrence for each value": this implies that $S = S' = \{s_i \in S | \forall j \neq i . place_t(s_i) \neq place_t(s_j)\}$. In this case we have:

$$P_t(S) = \begin{cases} 0 & \text{if } S \sqsubseteq t \\ \frac{1}{\prod_{\forall s_i \in S'} leaves(place_t(s_i))} & \text{otherwise} \end{cases}$$

$$P_t(s_i) = \begin{cases} 0 & \text{if } \{s_i\} \sqsubseteq t \\ \frac{1}{leaves(place_t(s_i))} & \text{otherwise} \end{cases}$$

We can observe that $P_t(S) \leq P_t(s_i)$ because the numerators of the two formulas are equal and the denominator of $P_t(S)$ is greater than the one of $P_t(s_i)$.

(c) "for each sensitive place $s_i \in S$ in the semantic trajectory t there can be more than one occurrence of $place_t(s_i)$": this implies that $S' = \{s_i \in S | \forall j \neq i . place_t(s_i) \neq place_t(s_j)\} \subset S$. We can rewrite the set S in the following way: $S = \{S_1, \ldots, S_m\}$ with $m \leq h$ where each $S_i = \{s_j \in S | place_t(s_i) = place_t(s_j) \; and \; s_i \in S\}$. Moreover, it easy to see that $P_t(S) = P_t(S_1) \times P_t(S_2) \times P_t(S_m)$. We also know, as shown in the point (a), that $P_t(S_i) \leq P_t(s_{i_j})$ for each $s_{i_j} \in S_i$ therefore we can conclude that $P_t(S) \leq P_t(s_{1_j}) \times P_t(s_{2_j}) \times P_t(s_{m_j})$ and therefore $P_t(S) \leq P_t(s_{i_j})$.

Since, we shown $P(S_Q, S) \leq P(S_Q, s_i)$ and we know $P(S_Q, s_i) \leq c$ the lemma follows. □

Theorem 15. *Given a semantic trajectory dataset $\mathcal{ST}$ and a protection probability threshold to be guaranteed $c \in [0, 1]$, the CAST algorithm defined in 1 produces a dataset $\mathcal{ST}^*$ that is a c-safe version of $\mathcal{ST}$.*

Proof. The correctness of the theorem is implied by the Lemma 14. Indeed, CAST algorithm generates groups of m semantic trajectories which are identical with respect to the quasi-identifier sequence S_Q. For each group it guarantees, by construction, that for each sensitive place s_i belonging to a sensitive place sequence in the group, we have $P(S_Q, s_i) \leq c$; by Lemma 14 we can also derive that $P(S_Q, S) \leq c$. As a consequence the theorem follows. □

5.2 Complexity Analysis

In this section we study the complexity of the CAST algorithm. In the following analysis we use the same notation of the Algorithm 1 and introduce some new measures: n is the size of the the semantic trajectory database $\mathcal{ST}$, l is the maximum length of the semantic trajectories in terms of the number of places and k is the maximum number of trajectories with the same size. Therefore, considering the different parts of the Algorithm 1, we have:

SuppressFalseSP: this function computes the support of each place visited by the user applying a temporal constraint to detect workplaces, this can be done in linear time w.r.t. the number of users' trajectories, but since it is done for all the users this covers the whole dataset and then the complexity is $O(n)$.

Ordering by length: a merge sort is used, therefore the complexity is $O(n \log n)$;

Extraction of trajectory subsets : considering that the database is sorted by length, this operation is very simple and can be done in a constant time. We represent it with $O(c)$.

FindBestGroup: given a subset of trajectories K the algorithm analyzes the m-permutations of K. For the complexity analysis we consider the worst case where every set K contains the maximum number of trajectories k, therefore the number of m-permutations is $d = k(k-1)(k-2)\dots(k-m+1)$ and then the complexity of this step becomes $O(d)$.

Generalization: the selected semantic trajectories are generalized to guarantee the c-$safety$ analyzing each place; in this case considering the worst case where the length of the trajectories is l the complexity becomes $O(l \times m)$.

To compute the overall complexity it is important to take into consideration that the last two operations are repeated $\lfloor \frac{n}{m} \rfloor$ times obtaining:

$$O(n) + O(n \log n) + O(\frac{n}{m}(d + lm)) = O(n + n \log n) + O(\frac{n}{m}d + nl) = O(n \log n + \frac{n}{m}d + nl)$$

It is important to notice that in real cases the average length of semantic trajectories is small due the fact that they represent only the stops occurring in places described by the taxonomy $PTax$. Moreover, the m value is usually small due the fact that a high value leads to a stronger generalization which decreases the resulting dataset usefulness. These aspects will be highlighted in Section 6.

6 Experimental Evaluation

This section summarizes the results of running CAST to anonymize two different trajectory datasets, both representing GPS traces of private cars. The first dataset has been collected in

the center of Milano, a city in the north of Italy, whereas the second one has been collected in a wider area which includes Pisa, a small city in the center of Italy[1]. Details on the two dataset are depicted in Table 1. These two datasets, despite being the same kind of data, present different characteristics. In fact the Milano dataset covers a smaller area compared to the Pisa dataset, therefore the latter one results more spatially sparse. In addition, the taxonomies used for the two experiments are different. The taxonomy used in the Pisa case study is larger in number of concepts, levels and leaves (i.e. places instances). We will see later, when discussing the plots, that these characteristics affect the quality of the results. In the experiments we have analyzed the following aspects: (a) the data quality after the data transformation by measuring the quality of the sequential patterns extracted from the anonymized data; (b) the disclosure probability on the data; and (c) the runtime execution of the proposed algorithm. Before going into the details of the two case studies, we give some preliminary definitions of the measures we used to evaluate the data quality.

Dataset	*N. trajectories*	*Avg length of traj (Km)*	*Covered area* (Km^2)	*Avg Stops per Traj*	*Avg. Sensitive Stops per Traj*
Milano	6225	8.977	386.021202	5.2	1.28
Pisa	7231	16.502	4675.936938	7.1	1.63

Table 1: Statistics about the experimental datasets

6.1 Measuring the quality of the sequential pattern results

To determine how sequential pattern results of the original dataset are preserved after the anonymization phase, we designed *ad hoc* quality measures: *coverage coefficient* and *distance coefficient*.

Definition 16 (Covered Predicate)**.** Given two sequential patterns $P = p_1 \dots p_h$ and $T = t_1 \dots t_m$ and a privacy place taxonomy $PTax := \langle C, HC, \lambda \rangle$, the *covered predicate* is defined as:

$$covered(P, T, PTax) = \begin{cases} true & \text{if } (h = m) \wedge \forall_{k=1\dots h}(p_k = t_k \vee HC(p_k, t_k)) \\ false & \text{otherwise} \end{cases}$$

Definition 17 (Coverage Coefficient)**.** Given two sets of sequential patterns $\mathcal{P}$, $\mathcal{P}^*$ and a privacy place taxonomy $PTax$, the *coverage coefficient* is a value in $[0, 1]$ defined as:

$$coverage(\mathcal{P}, \mathcal{P}^*, PTax) = \frac{|CovSet(\dots)|}{|\mathcal{P}|}$$

where

$$CovSet(\mathcal{P}, \mathcal{P}^*, PTax) = \{P \in \mathcal{P} | \exists\, P^* \in \mathcal{P}^*.\, covered(P, P^*, PTax) = true\}$$

Intuitively, the coverage coefficient measures how many patterns extracted from the non-anonymized dataset are covered at least by the patterns extracted from the anonymized dataset with a certain level of generalization. It's important to notice that the coverage does not measure *how much* the patterns are generalized, but only if they are covered by a

[1]The two datasets have been collected by OctoTelematics Spa and they are not public available

pattern obtained from the anonymized dataset or not. This means that a pattern composed by place generalized to the root of the taxonomy (i.e. $\langle Place\ Place\ Place \rangle$) will cover all the patterns with the same length. We have defined three levels of coverage:

- *Coverage upper bound* considers the whole set of patterns $\mathcal{P}^*_{ub}$ extracted from the anonymized dataset.
- *Coverage* considers a subset of the patterns $\mathcal{P}^*$, extracted from the anonymized dataset, which does not include the patterns composed only of root items.
- *Coverage lower bound* considers only the subset of patterns $\mathcal{P}^*_{lb}$, extracted from the anonymized dataset, which does not contain any root item.

The aim of these three levels is to better describe which kind of generalization is performed and the consequences on the patterns, including or not the room items. We recall that the presence of the root item indicates that we loose the knowledge about the stop place. The *distance coefficient* is a measure that represents the distance between the set of patterns extracted from the original dataset and the set of patterns extracted from the anonymized dataset. The definition of this coefficient is very similar to the *generalization cost*. The difference consists in the fact that when we generalize a dataset of trajectories we keep a correspondence between the original trajectory and the anonymized one. Instead, in the case of patterns extracted from the two versions of the dataset (original and anonymized), we do not know a priori which pattern is the generalized version of a specific pattern since we loose the one-to-one correspondence. In other words, given a pattern extracted from the original dataset, we can have a set of patterns extracted from the anonymous dataset that can be its generalized version. Therefore, the *distance coefficient* definition has to take into consideration this fact. In particular, for each pattern extracted from the original dataset, this measure computes the distance from the pattern extracted from the anonymous dataset and that minimizes the *distance* value (Definition 11 or Definition 12).

Definition 18 (Distance coefficient)**.** Let $\mathcal{P}$, $\mathcal{P}^*$ and *PTax* be the set of sequential patterns extracted from the original dataset, the set of sequential patterns extracted from an anonymized dataset and a privacy place taxonomy, respectively. The *distance coefficient* is a value in $[0, 1]$ defined as:

$$Dis(\mathcal{P}, \mathcal{P}^*, PTax) = \frac{\sum_{P_i \in \mathcal{P}} Min_{P^*_j \in \mathcal{P}^*}(distance(P_i, P^*_j, PTax))}{|\mathcal{P}|}$$

This measure shows a different aspect of the effect that the data transformation has on the patterns. Indeed, it expresses the cost of transforming a pattern from the non-anonymized dataset to a pattern extracted from the anonymized one. In other words, it measures how much information we loose, from the semantical point of view, when we apply sequential pattern to the anonymized dataset. In the following we use the terms $Distance_H$ *coefficient* and $Distance_L$ *coefficient* to denote the *Distance coefficient* that uses the Hops-based distance and Leaves-based distance, respectively.

In the following, we study how these two measures may vary depending on different settings of the problem in two different experiments.

6.2 Case study on Milan urban area

In this section we illustrate the effects of the anonymization process on the sequential patterns extracted from the datasets using Coverage and Distance Coefficients. We have applied the *CAST* algorithm on a dataset containing the movements of 17000 moving cars in

Milan urban area, collected through GPS devices in one week (Figure 3). This is a sample of data donated by a private company [26] devoted to collect them as a service for insurance companies.

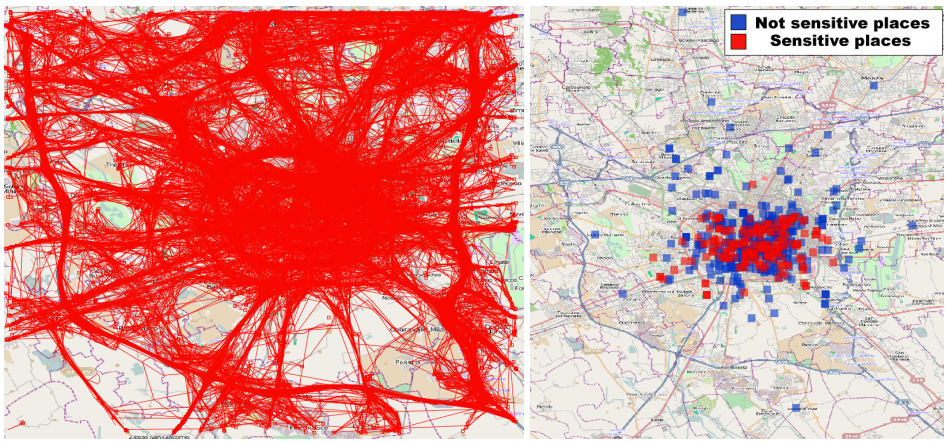
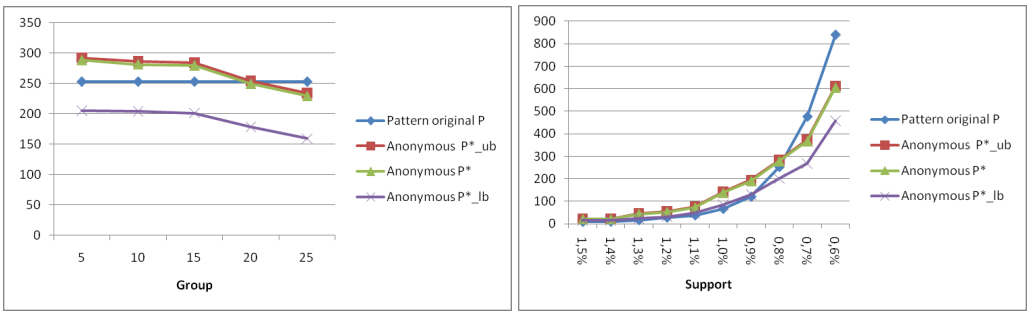


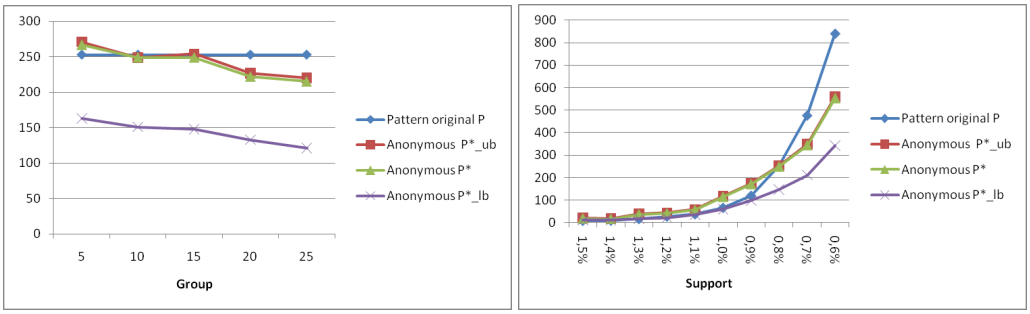
Figure 3: Milan urban area trajectories dataset (left) and the Points Of Interest, blue indicates the non-sensitive and red the sensitive ones (right)

Starting from the original dataset containing GPS traces, we have computed the semantic trajectories detecting the stops as described in Section 2. The set of places has been downloaded manually selecting the Points of Interest from *Google Earth* [10]. Figure 3 shows a plot of sensitive (red squares) and non-sensitive places (blue squares). The taxonomy used in the experiment is represented in Figure 2 with 530 instances of places (leaves) instead of the 14 leaves depicted in that figure. In the taxonomy the red nodes represent the places considered sensitive. We have obtained a dataset of 6225 semantic trajectories with an average length equal to 5.2 stops. We have run the sequential pattern mining algorithm on both the original semantic trajectories dataset and the anonymized versions varying the minimum support threshold value (parameter of the mining algorithm).

A complete view of how the results change varying the parameters of CAST algorithm is presented in Figure 4, Figure 5 and Figure 6. Here, we have compared the results using the two distance functions, *Hops-based Distance* ($Distance_H$) and *Leaves-based Distance* ($Distance_L$). Figure 4(b) and Figure 4(d) highlight the fact that the generalization have a double effect on the patterns: (i) the frequency of generalized places increases, (ii) the frequency of leaf places of the taxonomy decreases. Therefore, with a high support threshold the difference between the patterns created and removed by the generalization phase is positive and this increases the size of the resulting patterns set. However, after a certain threshold, due to the smaller number of generalized places, the decrease of patterns becomes predominant. We notice a reduction of patterns with lower support and this is accentuated by the cut of semantic trajectories during the anonymization process. These effects are highlighted also in Figure 5(c): we notice that, decreasing the support, the coverage coefficient tends first to increase and then to decrease. From the comparison of the two distances it is clear that the *Hops-based distance* produces more patterns than the *Leaves-based Distance*. This is due to the fact that using the *Leaves-based Distance* the algorithm tends to *not generalizing* the subtrees with high number of leaves or rather *generalize* faster the subtrees with a low number of leaves. This behavior tends to produce a greater number of distinct items and therefore a low number of patterns. For the same reason, the lower



(a) $c = 0.3, distance = Distance_H, support = 0.8\%$ (b) $c = 0.3, m = 15, distance = Distance_H$



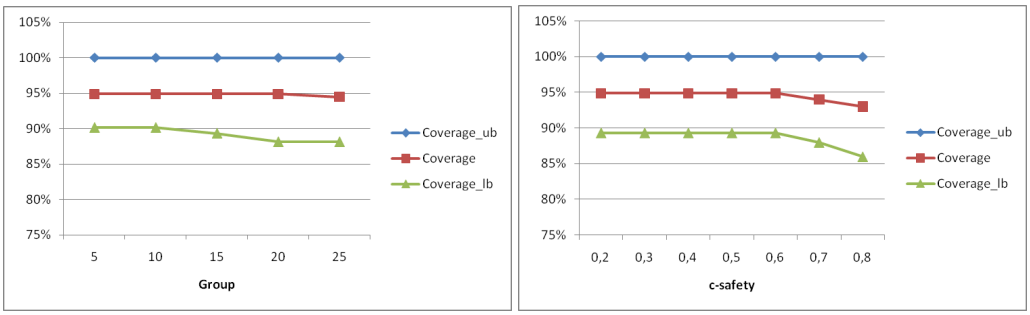
(c) $c = 0.3, distance = Distance_L, support = 0.8\%$ (d) $c = 0.3, m = 15, distance = Distance_L$

Figure 4: Number of patterns extracted from Milan Data varying the group size and the support threshold.

bound of the number of patterns decreases since the subtrees with a low number of leaves reaches faster the root item. Figures 4(a) & 4(c) demonstrate this effect.

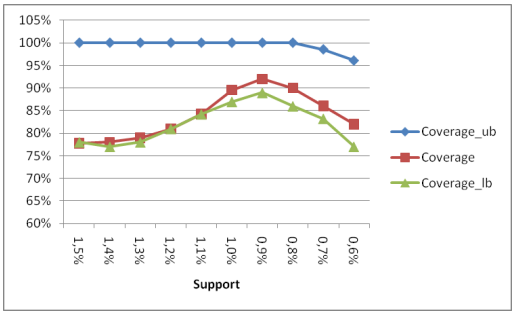
In Figure 5 we illustrate how the coverage coefficient changes varying the group size, the c-safety and the support threshold. In Figures 5(a) & 5(d) the *coverage lower bound* gives us another hint of how the patterns are transformed after the anonymization using the two different distance functions. We can notice that, using the *Hops-based distance* and when the group size is between 5 and 20, they are generalized without root items (the presence of root items means we completely lost the information on the stop place). When the group size exceeds 20, the patterns contain at least one root item, therefore the lower bound coverage value decreases fast. In the other case, using the *Leaves-based Distance*, it is likely that the place is generalized to the root when its subtree has a low number of leaves. As a consequence, the lower bound decreases when the group size has value 5.

In the other hand all the cases shown in Figure 5 the coverage value of Leaves-based Distance is greater than the Hops-based Distance even if, as shown before, the number of patterns extracted is lower. In other words, from these results emerges that the Leaves-based Distance leads to a more *intelligent* anonymization which maintains the patterns. Furthermore the analysis of the average distance in Figure 6 highlights that the patterns are not only better covered using the Leaves-based Distance but also the distance is lower.

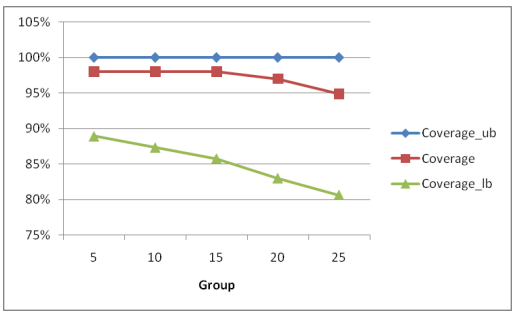


(a) $c = 0.3, distance = Distance_H, support = 0.8\%$

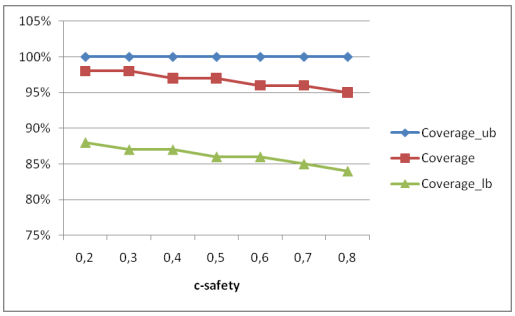
(b) $m = 15, distance = Distance_H, support = 0.8\%$



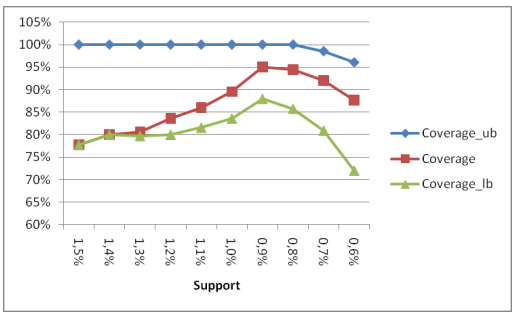
(c) $c = 0.3, m = 15, distance = Distance_H$



(d) $c = 0.3, distance = Distance_H, support = 0.8\%$



(e) $m = 15, distance = Distance_L, support = 0.8\%$

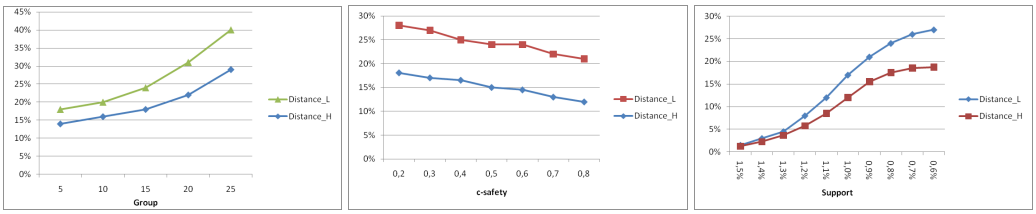


(f) $c = 0.3, m = 15, distance = Distance_L$

Figure 5: Study of the coverage on Milan Data varying the group size, the c-safety threshold and the support threshold.

6.3 Case study on Pisa

In this section we illustrate an experiment similar to the previous one on a dataset with different characteristics. The Pisa dataset contains the movements of 25078 moving cars in Pisa and the surrounding area which covers almost a quarter of Tuscany. Data is collected from the same provider of the Milano dataset. The dataset and the points of interest used in the taxonomy are plotted in a map in Figure 7. The main differences between Pisa and Milano experiments are three: (i) the size of the dataset, since the Pisa dataset is smaller than the Milano one; (ii) since the spatial area where the movements have been collected is wider than Milano, Pisa dataset and corresponding places are more sparse, and (iii) the taxonomy used in the Pisa experiment is deeper and larger than the previous one, contain-



(a) $c = 0.3$, $support = 0.8\%$ (b) $m = 15$, $support = 0.8\%$ (c) $m = 15$, $c = 0.3$

Figure 6: Study of the coefficient distance value varying the group size and the c-safety threshold and the support threshold.

ing 851 stop places (Figure 8). Similarly to the previous case, red subtrees represent the sensitive places. We have computed the semantic trajectories from the Pisa dataset with the new taxonomy obtaining 7231 trajectories with an average length of 7.1 stops.

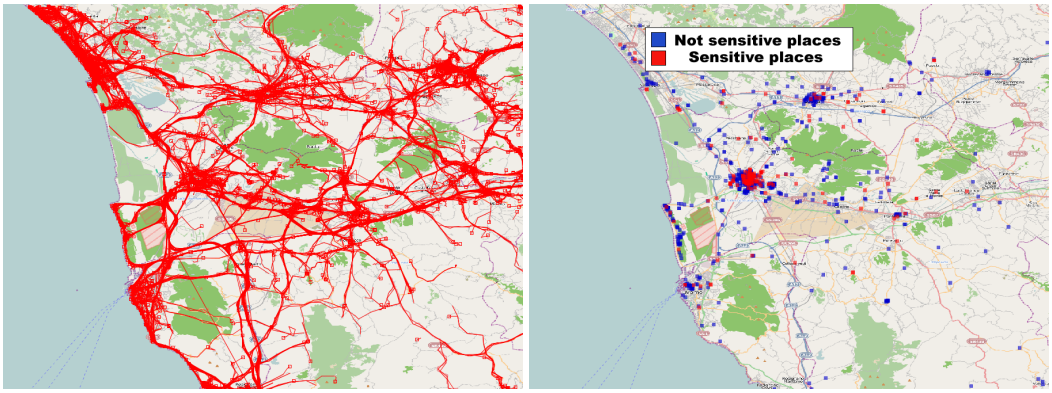


Figure 7: Pisa trajectories dataset (left) and the Points Of Interest, blue indicates the non-sensitive ones and red the sensitive ones (right)

Figure 9 shows that the number of patterns is quite lower even if the dataset is bigger than the previous one. Figure 9(d) highlights that using the Leaves-based Distance the number of patterns is always below the original ones. Moreover, the Leaves-based Distance maintains a very high coverage w.r.t. the Hops-based Distance as shown in Figures 10(c) & 10(f). This experiment mainly confirmed the results obtained for the Milan urban area. In some cases some behaviors appear accentuated, such as the case illustrated in Figure 10 where we can get evidence of the *lower bound effects* of the Leaves-based Distance.

6.4 Measuring the Disclosure Probability

In this section we want to study empirically the disclosure probability of a sensitive place by an hypothetical attacker which knows some of the quasi-identifiers of an specific user. This probability is the actual *c-safety* which is guaranteed in the anonymized dataset. We generated 10,000 instances of attacks and we computed the average of those probabilities. In the experiments we varied three parameters of the CAST algorithm, i.e., the c-safety threshold, group size and the distance function. We replicated these experiments for the two datasets. The overall result is that, generally, the empiric disclosure probability is

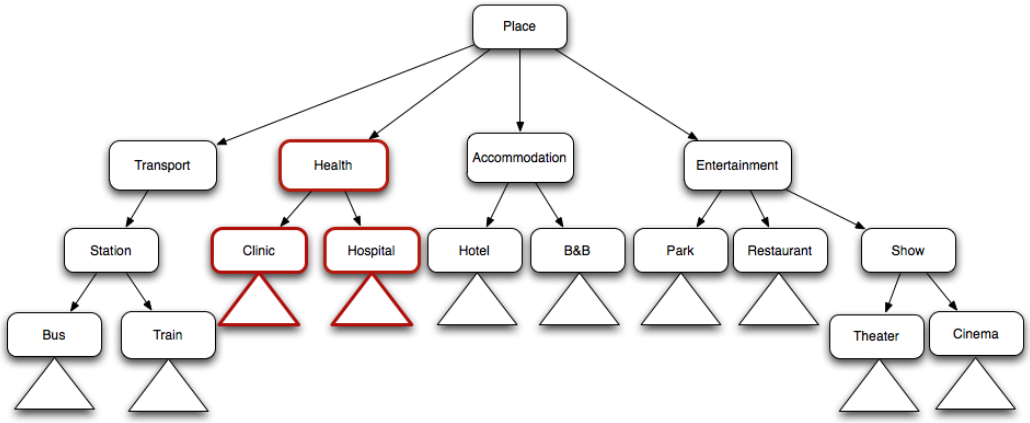
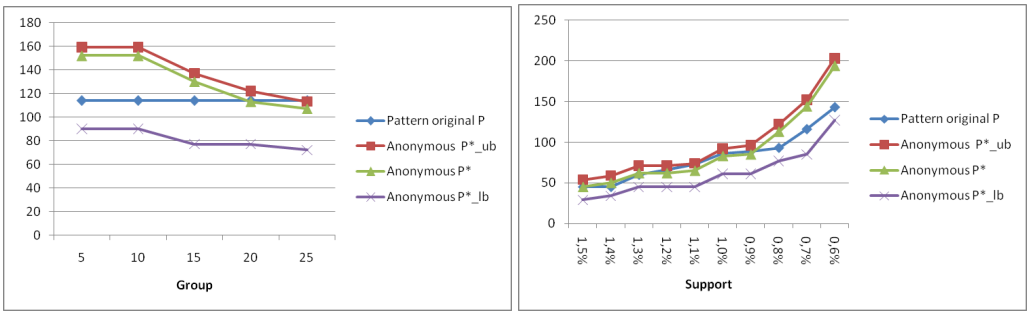
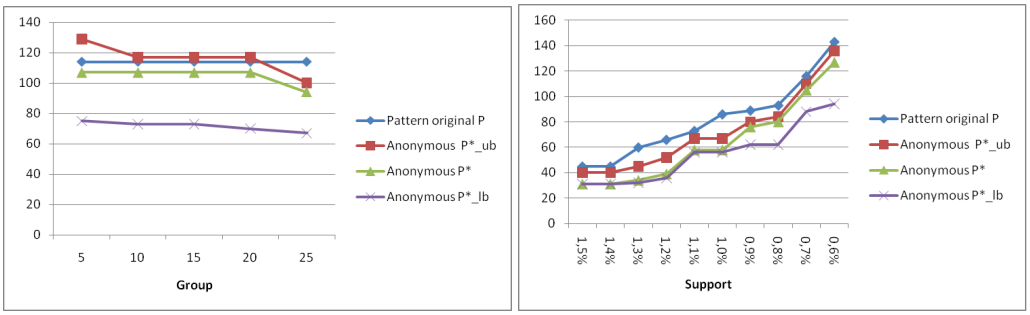


Figure 8: The taxonomy used in the experiments on Pisa dataset.



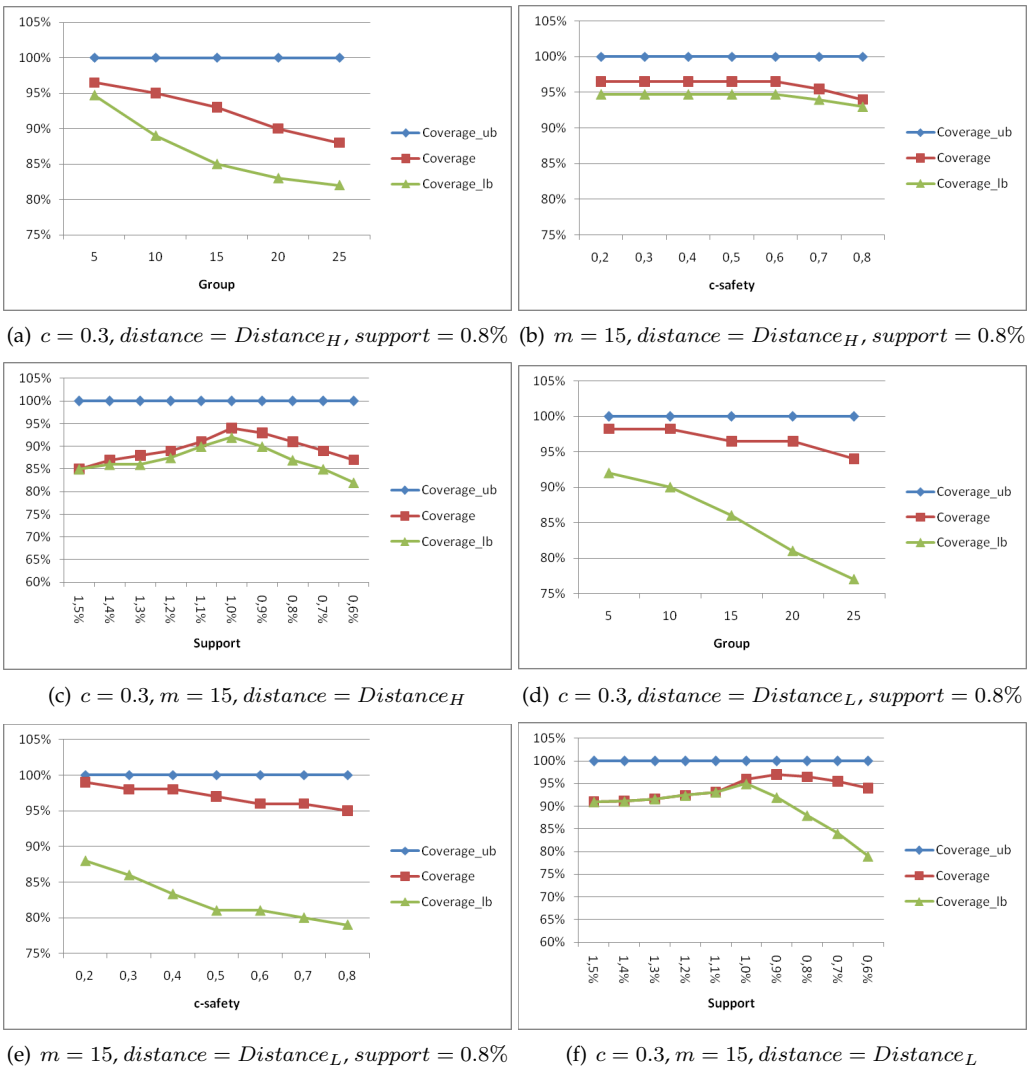
(a) $c = 0.3$, $distance = Distance_H$, $support = 0.8\%$ (b) $c = 0.3$, $m = 15$, $distance = Distance_H$



(c) $c = 0.3$, $distance = Distance_L$, $support = 0.8\%$ (d) $c = 0.3$, $m = 15$ $distance = Distance_L$

Figure 9: Number of patterns extracted from Pisa Data varying the group size and the support threshold.

lower than the given c-safety threshold. In particular, Figure 12 highlights that the probability to infer the sensitive place given one quasi-identifier place is below 0.03 and then the probability increases when the number of known places augments; this is due to the fact that the set of semantic trajectories containing the exact sequence of known quasi-identifiers becomes smaller. Another effect to be noticed is that when the *group* parameter



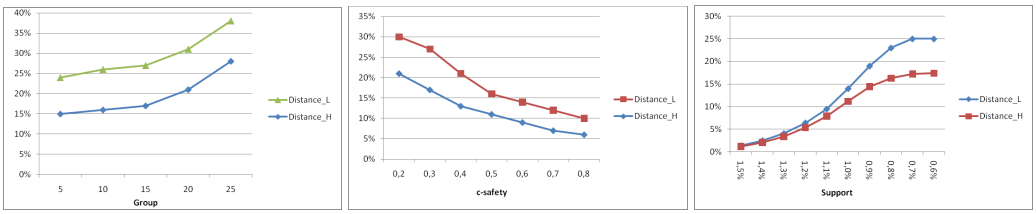
(a) $c = 0.3, distance = Distance_H, support = 0.8\%$ (b) $m = 15, distance = Distance_H, support = 0.8\%$

(c) $c = 0.3, m = 15, distance = Distance_H$ (d) $c = 0.3, distance = Distance_L, support = 0.8\%$

(e) $m = 15, distance = Distance_L, support = 0.8\%$ (f) $c = 0.3, m = 15, distance = Distance_L$

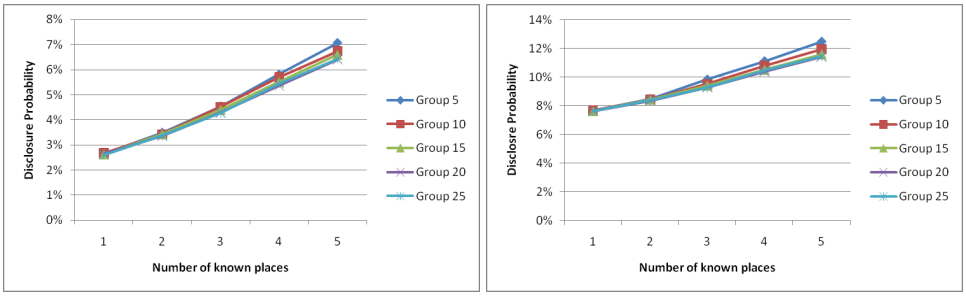
Figure 10: Study of the coverage on Pisa dataset varying the group size, the c-safety threshold and the support threshold.

becomes higher the probability decreases. Indeed, in this case, the semantic trajectories becomes more generalized by the algorithm, thus increasing the possibility of having similar anonymized trajectories. This effect becomes more evident in the Pisa dataset (Figure 13) where we have a larger taxonomy in terms of number of leaves and abstraction levels. We also highlight the fact that the results using the the two distance functions, Hops-based and Leaves-based, is similar. The only difference results from the experiments on Pisa case study, where the anonymized dataset using the *leaves-based* distance tends to generalize a smaller set of places (Figure 15).



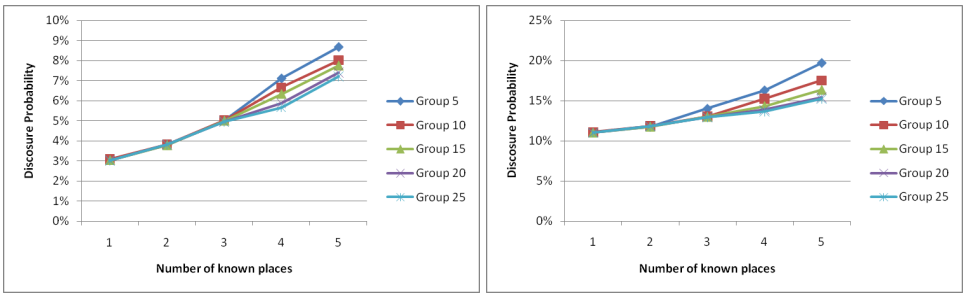
(a) $c = .3$, $support = 0.8\%$ (b) $m = 15$, $support = 0.8\%$ (c) $c = 0.3$, $m = 15$

Figure 11: Study of the distance coefficient varying the group size and the c-safety threshold and the support threshold.



(a) Milano Dataset, Probability threshold is c=0.3 (b) Milano Dataset, Probability threshold is c=0.6

Figure 12: Results of the experiments that measure the real disclosure probability on Milano dataset using Hops-based distance

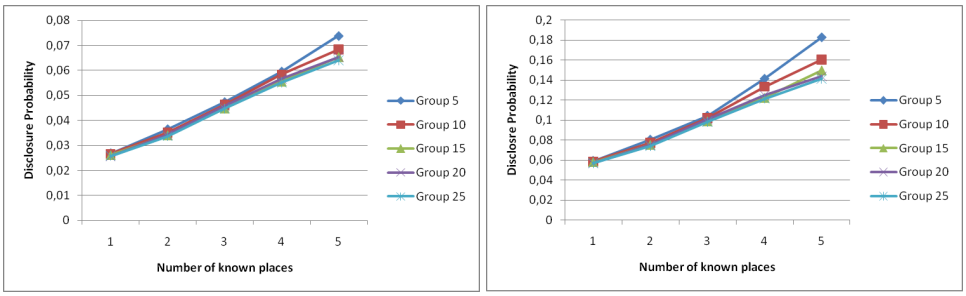


(a) Pisa Dataset - Probability threshold is c=.3 (b) Pisa Dataset - Probability threshold is c=.6

Figure 13: Results of the experiments that measure the real disclosure probability on Pisa dataset using the Hop-based distance

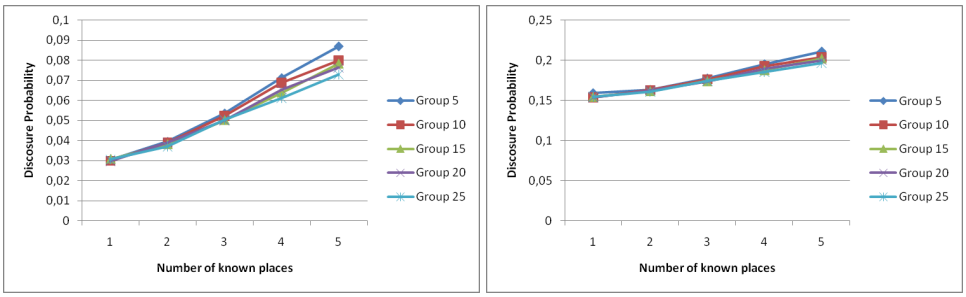
6.5 Runtime Analysis

The CAST algorithm has been implemented using Java 6.0. All experiments has been performed on an Intel Core 2 Duo T6400 with a 2.00GHz CPU over a Windows platform. We will now assess the total time needed to generate a *c-safe* database of semantic trajectories. Figure 16 reports the timings for anonymizing Milan dataset by using CAST and the Hops-based distance. We show the runtime behavior of CAST algorithm for different values of the size of the generalized groups and for various c-safety values. As expected, Figure



(a) Milano Dataset - Probability threshold is c=.3 (b) Milano Dataset - Probability threshold is c=.6, Leaves-based distance labelfig:KnownProb2F

Figure 14: Results of the experiments that measure the real disclosure probability on Milano dataset suing Leaves-based distance the



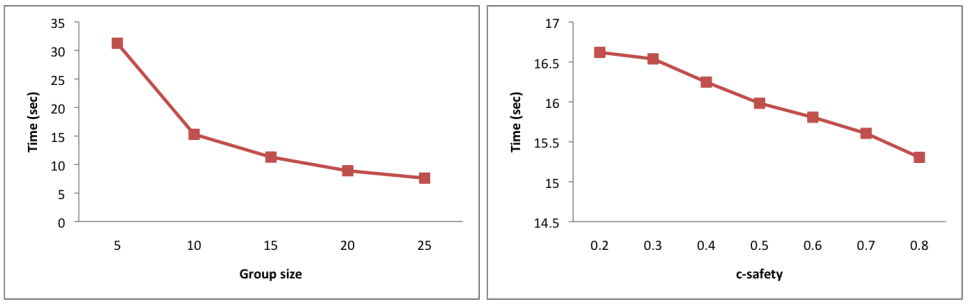
(a) Pisa Dataset - Probability threshold is c=.3 (b) Milano Dataset - Probability threshold is c=.6

Figure 15: Results of the experiments that measure the real disclosure probability on Pisa dataset using the Leaves-based distance

16(a) shows that when the size of the groups increases, the algorithm requires less execution time. This is due to the fact that bigger groups correspond in less groups to be handled. Analogously, Figure 16(b) shows that the execution time decreases when the c-safety value is higher. This behavior can be explained with the fact that a higher protection (that is, lower values of c) requires more generalization and so more time. We only show the execution time on Milan data with the use of Hops-based distance since the results obtained by using the Leaves-based distance and those obtained on Pisa dataset are very similar.

7 Related work

Many research studies have focused on the design of techniques for privacy-preserving data mining and for privacy-preserving data publishing [4, 3, 23]. The basic operation for data publishing is to replace personal identifiers with pseudonyms. However, in [29] authors showed that this simple operation is insufficient to protect privacy. They proposed k-anonymity to make each record indistinguishable from at least $k-1$ other records thus protecting data against the *linking attack*. The k-anonymity framework is the most popular method for data anonymization and, for relational datasets, is based on the attribute distinction among *quasi identifiers* (attributes that could be used for linking with external in-



(a) Disclosure Probability threshold is c=.3 (b) The size of groups is m=10

Figure 16: Execution Time of CAST

formation) and *sensitive* attributes (information to be protected) [31]. Although it has been shown that finding an optimal k-anonymization is NP-hard [19] and that k-anonymity has some limitations [17, 16], this framework is still very relevant and it is often used in the studies on privacy issues in transactional databases [33, 13] and in location-based services (LBSs) [12, 20, 21], as well as on the anonymity of trajectories [2, 25, 35, 22].

In [33] authors study the k-anonymization problem of set-valued data and define the novel concept of k^m-anonymity. They assume an a priori distinction between sensitive items and quasi-identifier and provide a method based on item generalization in a bottom-up enumerative manner. In [13] instead authors present a top-down local recoding method using the same assumptions in [33], but they focus on a stronger privacy guaranty, the complete k-anonymity. These methods present all the limitations deriving from the k-anonymity that are addressed by the concept of l-diversity [17]. The method presented in [9] solve the l-diversity problem efficiently for transactional datasets with a large number of items per transaction. This work distinguishes between non-sensitive and sensitive attributes.

In [2], 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 [25] 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. This two works do not account for any notion of quasi-identifiers. In [35] different moving 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*. This work differs from ours in two aspects. First, in our privacy model the set of quasi-identifiers is global and is chosen by a domain expert who provides the place taxonomy where the places are tagged as "sensitive" or "quasi-identifier". Clearly, the choice and the tagging depends on the specific application domain. In [35] authors assume that in some way for each user the set of quasi-identifier is known. Authors suppose for example that the quasi-identifiers may be chosen by the user while subscribing to a service, or may be part of the users personalized settings, or they may be found by means of data analysis. Second, their algorithm for the data transformation is based on spatial generalization because it takes into account

raw trajectories. In contrast, we apply a generalization of semantic trajectories that uses a place taxonomy. In other words, we do not apply any geometric transformation but we simply generalize a concept describing semantically a place with a more general concept given by the taxonomy.

In [22] authors present a method for the anonymization of movement data combining the notions of spatial generalization and k-anonymity and they show how the results of clustering analysis are faithfully preserved.

It is worth noticing that all these approaches deal with the anonymization of trajectories from the geometric point of view. The main different from geometric-based approaches and the semantics-based approach presented in this paper is that the latter tends to preserve the semantics of the visited places while preserving privacy. On the contrary, geometric-based approaches tend to modify the spatial coordinates of the moving objects.

In [1, 32], 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. The second one 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. The anonymization is obtained only by means of suppression of the dangerous observations from each trajectory.

In the context of LBS the work presented in [7] 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 [7] anonymizes a user's location during the communication with a LBS provider upon a service request.

The frameworks presented in [14, 15] allow to anonymize *location data* for publication. Note that location data is different from trajectory data. Indeed, a location dataset is a collection of pairs composed of user identifier and location while a trajectory dataset is a collection of pairs composed of user identifier and an ordered list of locations visited by the users. In [14], given a location dataset and a sensitive event dataset recording events and their locations, authors study how to anonymize the location dataset w.r.t. the event dataset in such a way that by joining these two datasets through location, every event is covered by at least k user. In [15] Krishnamachari et al. define formally the attack model and propose two algorithms which employ transformations to anonymize the locations of users in the proximity of sensitive sites. Their attack model is not based on traces of locations.

In general, frameworks for sequence data anonymization can be considered very related to our approach since semantic trajectories are a particular type of data with a sequential nature where each place is an item of a sequence. A framework for the k-anonymization of sequences of regions/locations is presented in [28], where the authors also propose an instance of their framework which enables protected datasets to be published while preserving the data utility for sequential pattern mining tasks. In this work does not assume any distinction between quasi-identifier and sensitive item. Indeed, each item can play both the roles: private information to be protected and quasi-identifier if it is known by the attacker. In our framework this classification is possible thanks to the use of the taxonomy and the tagging of a place as sensitive by a domain expert who, given a domain application, can choose in that context which are the sensitive places. Lastly, in [34] Valls et al. consider the problem of preserving privacy for sequence of events, which has similar characteristic to semantic trajectories data. Their approach finds clusters of records and then, for each cluster, it constructs a prototype used to substitute the original values in the masked version of the data of the cluster.

8 Conclusions and Future Work

In this paper we have investigated the problem of publishing semantic trajectory datasets while preserving the privacy of the tracked users. Here, the focus is on the semantics of the visited places distinguishing between sensitive and quasi-identifier places. The introduced algorithm, called CAST, exploits a taxonomy to generalize the visited places in order to obtain a *c-safe* dataset. C-safety expresses the upper bound to the probability to infer that a given person has visited a sensitive place. The use of a taxonomy encoding domain knowledge about the places tends to perform a generalization of the visited places that preserves the semantics of the movement. Through a set of experiments on two real-life spatio-temporal datasets, we have shown that CAST, while guaranteeing a good protection through c-safety, also preserves the quality of the sequential pattern analysis.

Further research includes the experimentation of new pattern mining methods on the anonymized trajectories. Another point that needs to be evaluated is to study how to extend CAST to deal with semantic trajectories enriched with temporal information. 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 important point to take into account in a future work is the personalization of the quasi-identifier set for each user. Indeed, it would be interesting to understand how our framework could be adapted in order to allow to each user the specification of the set of quasi-identifier places.

Moreover, we would like to study a way to take into consideration in our framework the correlation among leaf nodes during the generalization. This point is very interesting but in the same time is very challenging because in this context the correlation strongly depends on the semantic of the concepts represented by the leaf nodes. As a consequence, finding an automatic way to compute this correlation could be very hard.

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). The Milano dataset has been donated by Octotelematics (Italy). Chiara Renso acknowledges support from CNR Short Term Mobility program and Dino Pedreschi acknowledges support by Google, under the Google Research Award program. Authors want to thank Prof. Josep Domingo-Ferrer and the anonymous referees for their useful suggestions.

References

[1] O. Abul, F. Bonchi, and F. Giannotti. Hiding Sequential and Spatio-temporal Patterns. *The Transactions on Knowledge and Data Engineering Journal*, 22(12): 1709-1723 (2010).

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

[3] C. C. Aggarwal , P. S. Yu. Privacy-Preserving Data Mining: Models and Algorithms, Springer Publishing Company, Incorporated, 2008

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

[5] 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.

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

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

[8] J. Domingo-Ferrer and A. Solanas A measure of variance for hierarchical nominal attributes In *Information Sciences* 2008. Vol. 178, issue 24, pages 4644-4655.

[9] G. Ghinita, Y. Tao, and P. Kalnis. On the Anonymization of Sparse High-Dimensional Data. In *International Conference on Data Engineering*, 2008.

[10] Google Earth. Google Earth, visited February 24, 2011 http://www.google.com/earth/index.html

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

[12] 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.

[13] Y. He and J. F. Naughton. Anonymization of Set-Valued Data via Top-Down, Local Generalization. In *Proceedings of the Very Large Data Base Endowment Volume 2 Issue 1, August 2009.*

[14] H. Hu, J. Xu, S. Tung On, J. Du, J. Kee-Yin Ng. Privacy-aware location data publishing. In *ACM Trans. Database Syst.* 35:3, 2010

[15] B. Krishnamachari, G. Ghinita, P. Kalnis. Privacy-Preserving Publication of User Locations in the Proximity of Sensitive Sites. In *roceedings of the 20th international conference on Scientific and Statistical Database Management, SSDBM* 2008, pages 95-113.

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

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

[18] J. A. Manso, V.C. Times, G. Oliveira, L.O. Alvares, V. Bogorny DB-SMoT: a Direction-based spatio-temporal clustering method. In *Fifth IEEE International Conference on Intelligent Systems (IEEE IS 2010)*, 2010.

[19] A. Meyerson and R. Williams. On the complexity of optimal k-anonymity. In *Symposium on Principles of Database Systems (PODS)*, 2004.

[20] M. F. Mokbel, C. Chow, and W. G. Aref. The new casper: Query processing for location services without compromising privacy. In *Proceedings of Very Large Data Base*, 2006.

[21] M. F. Mokbel, C. Chow, and W. G. Aref. The new casper: A privacy-aware location-based database server. In *IEEE 23rd International Conference on Data Engineering 2007 Publisher: IEEE, Pages: 1499-1500*.

[22] 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,

[23] A. Monreale, D. Pedreschi, R. G. Pensa Anonymity Technologies for Privacy-Preserving Data Publishing and Mining. In *Privacy-Aware Knowledge Discovery: Novel Applications and New Techniques*, F. Bonchi, and E. Ferrari (Eds). Chapman & Hall/CRC Data Mining and Knowledge Discovery Series. 2010.

[24] M. Nanni and R. Trasarti and C. Renso and F. Giannotti and D. Pedreschi Advanced knowledge discovery on movement data with the GeoPKDD system. In *International Conference on Extending Data Base Technology* 2010, pages 693-696.

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

[26] Octotelematics. http://www.octotelematics.com/.

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

[28] 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.

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

[30] 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).

[31] 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.

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

[33] M. Terrovitis, N. Mamoulis, P. Kalnis. Privacy-preserving anonymization of set-valued data. In *Proceedings of the Very Large Data Base Endowment* 1(1): 115-125 (2008)

[34] 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.

[35] R. Yarovoy, F. Bonchi, L. V. S. Lakshmanan, and W. H. Wang. Anonymizing moving objects: how to hide a mob in a crowd? In *International Conference on Extending Data Base Technologies*, 2009.