10

Mobility Data & Privacy

Fosca Giannotti[1], *Anna Monreale*[2] *and Dino Pedreschi*[2]
[1] *ISTI-CNR, Pisa*
[2] *University of Pisa*

10.1 Introduction

Mobility data represent a very useful source of information and thanks to mobile telecommunications and ubiquitous computing the location of mobile users can be continuously sensed and recorded. The sharing of mobility data raises serious privacy concerns. Mobility data reveal the mobility behavior of the people: where they are going, where they live, where they work, their religion preferences, etc. All this information refers to the private personal sphere of a person and so may potentially reveal many facets of his/her private life. As a consequence, this kind of data has to be considered personal information to be protected against undesirable and unlawful disclosure.

In the case of mobility scenarios, there exist two major different contexts where the location privacy problem has to be taken into consideration: on-line location-based services and off-line data analysis context. In the first case, a user communicates to a service provider his/her location to receive on-the-fly a specific service. An example of LBS is *find the closest Point of Interest (POI)* where a POI could be a restaurant. In the second case, large amounts of mobility data are collected and can be used for off-line data mining analysis able to extract reliable knowledge useful to understand and manage intelligent transportation, urban planning, sustainable mobility, etc.

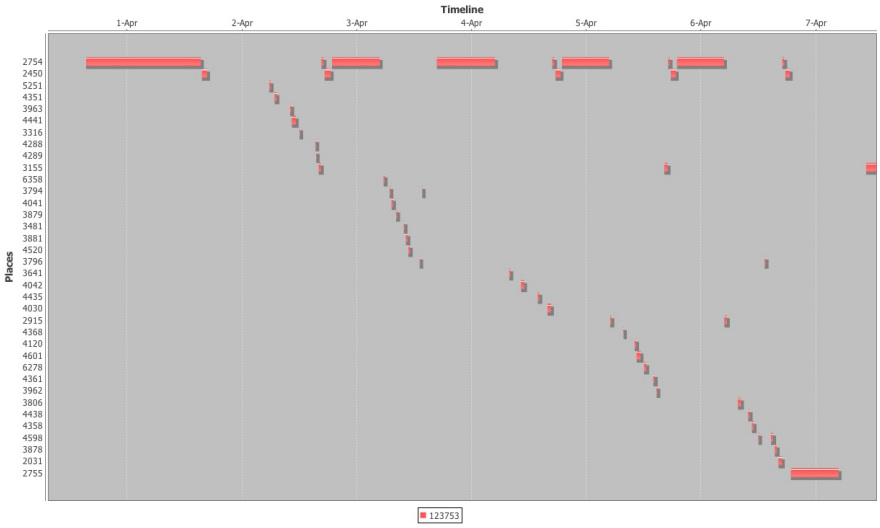
Many PETs (Privacy Enhancing Technologies) for mobility data have been proposed by the scientific community. Section 10.3 reviews the most important methods that have been proposed to address privacy issues in off-line data analysis by highlighting how the privacy models, initially proposed for relational database and presented in Section 10.2, are extended to spatio-temporal data. Privacy issues in the context of on-line location-based services is addressed in Chapter 2.

A common point of view among all these techniques is that, unfortunately, obtaining privacy protection is becoming more and more difficult because of the complex nature of movement data and it cannot simply

be accomplished by de-identification (i.e., by removing the direct identifiers contained in the data). Many examples of re-identification from supposedly anonymous data have been reported in the scientific literature. As an example in the context of GPS trajectories, consider Figure 10.1(a) that shows a de-identified GPS trajectory of a real user driving in Milan city for a period of one week (i.e., the first week of April 2007). Note that, in this figure street names are omitted in order to avoid easy re-identification of the user. Using only simple analytical tools, able to visualize the trajectory with its context, it is possible to show important and sensitive information on the user. For example, from Figure 10.1(a) it is possible to identify the most commonly visited regions; for this specific user there are two. In the figure, the rectangles represent region where the user has spent at least a minimum amount of time (10 minutes in this example) and the color darkness of each polygon is proportional to the number of different visits. While, by Figure 10.1(b) it is possible to infer: a) the region with identifier 2754 is the user's home since he/she usually stays there for the night; b) the region with identifier 2450 (the second most frequent region) is the work place, because he/she usually goes there every day at the same time, stays there for a short time and visits a lot of places during the day. Probably, this person is a sales agent. Clearly, by discovering the group of people living in that home and those working in that company it is possible to identify the user as the person which belongs to both groups.

In general, the data privacy problem requires to find an optimal trade-off between privacy and data utility. From one side, one would like to transform the data in order to avoid the re-identification of individuals and/or locations. Thus, one would like to publish safely the data for mining analysis or to communicate locations for receiving an on-line service without risks (or with negligible risk) for each data subject. From the other side, one would like to minimize the loss of information that can reduce the effectiveness of the underlying data when it is given as input to data mining methods and can cause a bad quality of the received location-based service. Therefore, the goal is to maintain the maximum utility of the data. In order to measure the information loss introduced by the data transformation process it is necessary to define measures of utility; analogously, it is necessary to quantify the risks of privacy violation. Privacy by design, in the research field of privacy-preserving data analysis, is a recent paradigm that promises a quality leap in the conflict between data protection and data utility (Section 10.4). Recent applications of this paradigm for the design of privacy-

(a)



(b)

Figure 10.1 De-identified GPS Trajectory

preserving frameworks for movement data (Section 10.4.1) prove that it is possible to achieve reasonable and measurable privacy guaranties and a good quality of the analytical results.

10.2 Basic Concepts for Data Privacy

The analysis and disclosure of personal information to the general public or to third parties such as data miners is subject to the limitations imposed by the regulations for privacy protection. Nevertheless, if this information was rendered anonymous, these limitations would not apply, hence making it possible to share and analyze the information without the explicit user agreement. In the last ten years, different models have been proposed by the scientific community to achieve privacy protection while sharing and analyzing personal sensitive information. The most important privacy models are: k-anonymity, l-diversity, t-closeness, randomization and cryptography-based models.

k-anonymity. The *k-anonymity* model was introduced by in the context of relational database, where data are stored in a table and each row of this table corresponds to one individual. The basic idea of the k-anonymity model is to guarantee that the information of every data subject cannot be distinguished from the information of other k-1 data subjects. This model is based on the assumption of the existence of the following kind of attributes in the user's record: *identifiers*, that explicitly identify data owners, such as name and social security number (SSN); *quasi-identifiers*, that could identify data owners or a small groups of them (e.g., gender and zip code); *sensitive* attributes, that represent sensitive person-specific information (e.g., disease and salary) to be protected. Based on this classification, the privacy requirement defined by k-anonymity is that for each released record (e.g., a record is a row in the table in Figure 10.2) there must be at least other k-1 records with the same quasi-identifier values. A set of records that have the same values for the quasi-identifiers is called *equivalence class*. The techniques adopted in the literature to enforce k-anonymity involve the removal of explicit identifiers and the generalization (e.g., date of birth is changed with the year of birth) or suppression (e.g., removing the date of birth) of quasi-identifiers. It is evident that these techniques reduce the accuracy of the disclosed information.

l-diversity. The k-anonymity model only protects the identity of a user. Indeed, if a group of k records have the same quasi-identifiers values and the same value of the sensitive attribute it is not able to protect the sensitive information. As an example, consider the table in Figure 10.2. Suppose that the adversary knows that Alice was born in 1988,

Quasi-Identifier attributes			Sensitive attribute
Gender	**Date of Birth**	**ZIP Code**	**Disease**
F	1988	561*	Flu
F	1988	561*	Flu
F	1988	561*	Flu
M	1990	910*	Heart Disease
M	1990	910*	Cold
M	1990	910*	Flu

Figure 10.2 A 3-Anonymous Database

lives in the area with ZIP code 56123 and is in the database. He knows that Alice's record is one of the first three in the table. Since all of those patients have the same medical condition (Flu), the adversary can identify Alice's disease.

To overcome this weakness the *l-diversity* model requires of obtaining groups of data subjects with indistinguishable quasi-identifiers and with an acceptable diversity of the sensitive information. In particular, the main idea of this method is that every k-anonymous group should contain at least l different values for the attributes containing personal information.

t-closeness. The problem with l-diversity is that it can be insufficient to prevent the disclosure of private information when the adversary knows the distribution of the private values. Indeed, if the adversary has prior belief about the private information of a data subject, he can compare this knowledge with the probability computed from the observation of the disclosed information. In order to avoid this weakness, the *t-closeness* model requires that, in any group of quasi-identifiers, the distribution of the values of a sensitive attribute is close to the distribution of the attribute values in the overall table. The distance between the two distributions should be no more than a threshold t. Clearly, this limits the information gain of the adversary after an attack.

Randomization. Finally, *randomization* model is based on the idea of perturbing the data to be published by adding a noise quantity. More technically, this method can be described as follows. Denote by $X = \{x_1 \dots x_m\}$ the original dataset. The new distorted dataset, denoted by $Z = \{z_1 \dots z_m\}$, is obtained drawing independently from the probability distribution a noise quantity n_i and adding it to each record $x_i \in X$. The

set of noise components is denoted by $N = \{n_1, \ldots, n_m\}$. The original record values cannot be easily guessed from the distorted data as the variance of the noise is assumed enough large. Instead, the distribution of the dataset can be easily recovered.

Cryptography-based Models. The basic idea of the privacy models based on cryptography techniques is to compute analytical results without sharing the data in such a way that nothing is disclosed except the final result of the analysis. In general, the application of these models allow to compute functions over inputs provided by multiple parties without sharing the inputs. This problem is addressed in cryptography in the field of secure multi-party computation. As an example, consider a function f of n arguments and n different parties. If each party has one of the n arguments it is necessary a *protocol* that allows to exchange information and to compute the function $f(x_1, \ldots, x_n)$, without compromising privacy. There exist some methods that allow transforming data mining problems into secure multi-party computation problems. In literature, many protocols was proposed for the computation of the secure sum, the secure set union, the secure size of set intersection and the scalar product. These protocols can be used as data mining primitives for secure multi-party computation in case of horizontally and vertically partitioned datasets.

10.3 Privacy in Off-line Mobility Data Analysis

In the context of the off-line mobility data analysis large amount of collected mobility data can be used for extracting reliable knowledge useful that allow to understand very complex and interesting phenomena. Indeed, these data can be used for various data analysis that allow to improve systems for city traffic control, mobility management and urban planning. Unfortunately, mobility data provide detailed movement information of individuals and thus this information could be used for their identification and sometimes for inferring personal sensitive information about them. Therefore, when spatio-temporal data has to be analyzed and/or published is fundamental to guarantee individual privacy protection of the respondents represented in the data.

The privacy models described in the previous section have been widely adopted to achieve privacy protection in the context of the off-line analysis of spatio-temporal data. The different and more complex nature of

mobility data with respect to relational tabular data sometimes rendered difficult to apply these privacy models directly and this had led to the definition of some suitable variants. The inadequacy of the above models for trajectory data depends on the fact that these data pose new challenges due to the following characteristics: time-dependency, location-dependency and data sparseness. The location and time component of the mobility data render harder to enforce privacy protection because both information alone or in combination could be used by an attacker to re-identify individuals and discover sensitive information about them. As a consequence, a privacy defense has to take into consideration this fact and to apply a data transformation able to eliminate the privacy threats that derive from the two piece of information. Moreover, the problem is made more difficult by the sparseness of these large amount of data. Indeed, usually an individual visits few locations with respect to the total number of locations available in the territory so, the trajectories are relatively short and it is hard to find overlapping of locations among different trajectories. Additionally, the time component makes the situation more complicated because the same location can be visited by different individuals in different time periods. This makes the mobility data very sparse and in this setting, it is difficult to identify and to group together trajectories for enforcing for example traditional k-anonymity.

The next section shows how the basic data privacy notions presented in Section 10.2 have been adapted to address the new challenges posed by spatio-temporal data in off-line data analysis. We present three categories of PETs: PETs for mobility data publishing, PETs for distributed mobility data mining and PETs for knowledge hiding in mobility data.

10.3.1 PETs for Publishing of Trajectory Data

Mobility data publishing includes sharing the mobility data with specific recipients like data miners and releasing the data for public download. In both cases, the recipients could potentially be adversaries who try to associate sensitive information in the published data with a known person. The privacy-preserving techniques for mobility data publishing have the goal to transform spatio-temporal data to make them anonymous; in other words, they provide suitable formal safeguards against re-identification of individuals represented in the data by their movements.

In literature, most of the proposed PETs for mobility data publishing

use privacy models that are suitable variants of the classical k-anonymity model. They consider adversaries that use location-based knowledge for the re-identification of users. As explained in Section 10.2, an adversary can use *quasi-identifier* attributes (e.g., age, gender and zipcode) that represent public knowledge and to use them as key elements for the re-identification of individuals. Similarly, in spatio-temporal databases the attackers could identify the person corresponding to a given trajectory by using pairs of locations and timestamps that work as quasi-identifiers. In this context the challenge often is the definition of realistic and reasonable quasi-identifiers. Two important questions need to answer when we have to consider quasi-identifiers in spatio-temporal databases: (1) *can we assume the same set of quasi-identifiers for all the individuals in the database?* (2) *where and how should the knowledge of quasi-identifiers be obtained?*

Concerning the first question, in literature some works argue that, unlike in relational microdata, where every tuple has the same set of quasi-identifier attributes, in spatio-temporal data it is very likely that various individuals have different quasi-identifiers and clearly this fact should be taken into consideration in modeling adversary knowledge. Unfortunately, allowing different set of quasi-identifiers for different individuals makes the anonymization problem more challenge because the anonymization groups may not be disjoint.

Concerning the second question typically we have different possibilities: a) the quasi-identifiers may be part of the users personalized settings; b) they may be provided directly by the users when they subscribe to the service; and c) the quasi-identifier may be found by statistical data analysis or data mining.

Given that in the real-world the definition of quasi-identifiers in movements data is not easy some anonymization approaches do not use any information about the quasi-identifiers of trajectories during the anonymization process. In Section 10.3.1 we present the details of a typical technique of this category, while in Section 10.3.1 we explain a typical technique that takes into consideration the quasi-identifiers of trajectories.

Anonymization without Quasi-identifiers

A spatio-temporal technique that does not take into consideration any knowledge about the quasi-identifier of trajectories implicitly assumes that an adversary may identify a user in any location at any time. Clearly, this is a very conservative setting and under this assumption

the anonymized datasets are composed of anonymization groups each one containing at least k identical or very similar trajectories. This typically is achieved by the application of clustering-based approaches.

The application of classical k-anonymity notion in spatio-temporal data is hard because it is necessary to take into account some problems that are specific in this context. As an example, in the definition of the privacy model one should consider the inaccuracy of positioning device that introduces possible location imprecision in the collection of data. This leads to the definition of a variant of the k-anonymity notion called (k, δ)-*anonymity* suitable for moving objects databases, where δ represents the possible location imprecision. This novel concept is based on co-localization that exploits the inherent uncertainty of the moving objects whereabouts. Intuitively, the trajectory is considered as a cylindrical volume with some uncertainty. In other words, the position of a moving object in the cylinder then becomes uncertain. Figure 10.3 illustrates a graphical representation of an uncertain trajectory.

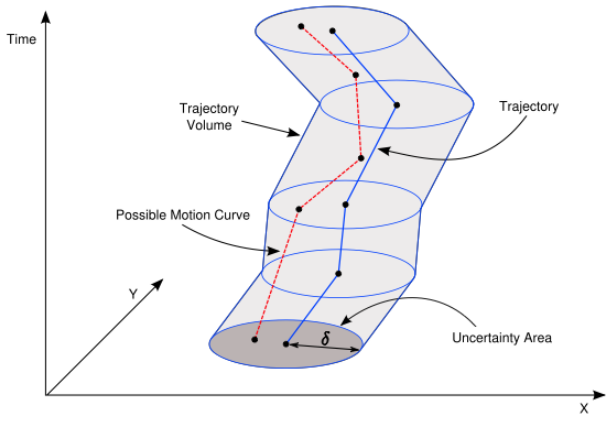


Figure 10.3 Uncertain trajectory: uncertainty area, trajectory volume and possible motion curve.

Two trajectories moving within the same cylinder are indistinguishable; this leads to the definition of (k, δ)-anonymity model:

Definition 10.1 Given an anonymity threshold k and a radius parameter δ, a (k, δ)-anonymity set is a set of at least k trajectories that are co-localized w.r.t. δ.

It was showed that a set of trajectories S, with $|S| \geq k$, is a (k, δ)-anonymity set if and only if there exists a trajectory t_c such that all the trajectories in S are possible motion curves of t_c within an uncertainty radius of $\frac{\delta}{2}$. Given a (k, δ)-anonymity set S, we obtain the trajectory t_c by taking, for each $t \in [t_1, t_n]$, the point (x, y) which represents the center of the minimum bounding circle of all the points at time t of all trajectories in S (Figure 10.4).

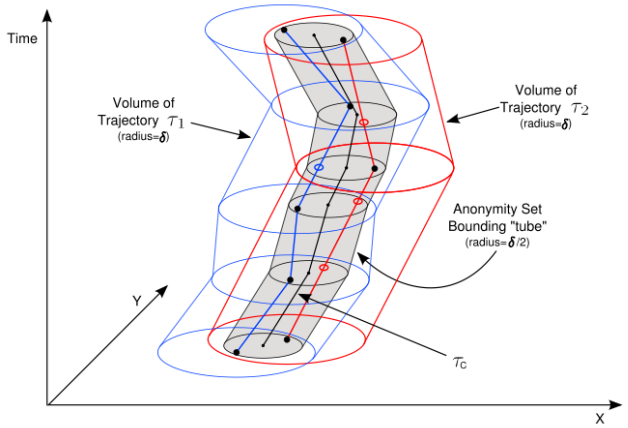


Figure 10.4 A $(2, \delta)$-anonymity set formed by two co-localized trajectories, their respective uncertainty volumes, and the central cylindrical volume of radius $\frac{\delta}{2}$ that contains both trajectories.

The (k, δ)-anonymity framework requires to transform a trajectory database D in D' in such a way that for each trajectory $t \in D'$ it exists a (k, δ)-anonymity set $S \subset D'$, $t \in S$ and the distortion between D and D' is minimized. To achieve (k, δ)-anonymous datasets we can apply a method based on trajectory clustering and spatial translation, that is a form of perturbation. In particular, it consists of three main steps:

1. **Pre-processing step.** The goal of this phase is to find a partition of the original database in equivalence classes w.r.t. the time span. In other words, each equivalence class contains trajectories with the same starting time and ending time. This step it is necessary because the algorithm has to compute the Euclidean distance between trajectories and if it is computed on the input raw data could lead to the generation of very small equivalence classes.
2. **Clustering step.** In this phase the trajectories are clustered by using

a greedy approach. It iteratively selects a pivot trajectory as cluster center and assigns its nearest $k-1$ trajectories to the cluster. The clusters must have a radius not larger than a given threshold to guarantee a certain compactness of the groups of trajectories. So, if this criterium of compactness is not satisfied then the process is repeated selecting a different pivot trajectory. Clearly, if a remaining trajectory cannot be added to any cluster without violating the compactness constraint, then it is trashed because it is considered as an outlier.

3. **Space transformation step.** The aim of this step is to transform each cluster in a (k, δ)-anonymity set. This is achieved perturbing each trajectory by the spatial translation that allows to put all the trajectories within a common uncertainty cylinder.

Anonymization based on Quasi-identifiers

A privacy-preserving technique for publication of spatio-temporal data, which considers quasi-identifiers in its model, has to address some issues that are not very easy. The most important is due to the fact that in this particular setting a specific set of locations, or of timestamps, cannot be a quasi-identifier for all the individuals in the database. Therefore, in the adversary model one should take into account that various moving objects have different quasi-identifiers. More formally, given a spatio-temporal database $D = \{O_1, \ldots, O_n\}$ corresponding to n individuals, and a set of m discrete time points $T = \{t_1, \ldots, t_m\}$, the quasi-identifier is defined as a function: $QID : \{O_1, \ldots, O_n\} \rightarrow 2^{\{t_1, \ldots, t_n\}}$. Since different moving objects may have different QID, the anonymization groups associated with different objects may not be disjoint. As an example consider the trajectories in Figure 10.5(a) and illustrated in Figure 10.5(c). Let $k = 2$ and $QID(O_1) = \{t_1\}$, $QID(O_2) = QID(O_3) = \{t_2\}$. Assume that the anonymization group for O_1 w.r.t. its QID $\{t_1\}$ is $\{O_1, O_2\}$ (dark rectangle in Figure 10.5(c)). This means in the anonymized database the region $[(1, 2), (2, 3)]$ is assigned to O_1 and O_2 at time t_1. Then, consider the anonymization group for O_2 as well as for O_3 w.r.t. their QID $\{t_2\}$ is $\{O_2, O_3\}$. Thus, in the anonymized database, O_2 and O_3 will both be assigned to the common region $[(2, 6), (3, 7)]$ (the second dark rectangle) at time t_2. Clearly, the anonymization groups of O_1 and O_2 overlap.

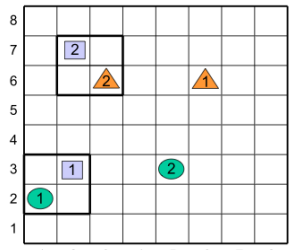
Clearly, it is possible that by combining overlapping anonymization groups, some moving objects may be uniquely identified, therefore a suitable privacy model has to be defined. Before describing the privacy model we have to introduce the attack model, called *attack graph*. An attack graph associated with a trajectory database D and its distorted

MOB	t_1	t_2
O_1	$(1,2)$	$(5,3)$
O_2	$(2,3)$	$(2,7)$
O_3	$(6,6)$	$(3,6)$

(a)

MOB	t_1	t_2
O_1	$[(1,2),(2,3)]$	$(5,3)$
O_2	$[(1,2),(2,3)]$	$[(2,6),(3,7)]$
O_3	$(6,6)$	$[(2,6),(3,7)]$

(b)



(c)

Figure 10.5 (a) original database; (b) a 2-anonymity scheme that is not safe, and (c) its graphical representation

version D' is defined as the bipartite graph G composed of nodes for every individual I in D (called I-nodes) and nodes for every moving object id O (called O-nodes) in the transformed database D'. G has an edge (I, O) iff for each $t \in QID(I)$, $D(O,t) \sqsubseteq D'(O,t)$, where $\sqsubseteq$ denotes spatial containment between two regions.

The attacker for conducting an attack can construct this attack graph and checks the degree of each I-node. There is no privacy breach if each I-node has degree k or more.

According to this attack the privacy model is formally defined as follows:

Definition 10.2 Let D be a trajectory database and D' its distorted version. Let G be the attack graph w.r.t. D and D'. Then D' is k-anonymous provided that (i) every I-node in G has degree k or more; and (ii) G is symmetric, i.e., whenever G contains an edge (I_i, O_j), it also contains the edge (I_j, O_i).

In order to achieve a k-anonymous database with respect to the above definition an algorithm has to find a k-anonymization groups, i.e., given a moving object O it has to find a set of $k-1$ moving objects which have the minimum aggregate distance from O over the entire set of times in $QID(O)$. This goal can be obtained with two steps:

1. Producing the $top-k$ candidates for forming the anonymization group of a specific object by using a method based on Hilbert index of spatial objects for efficient indexing of trajectories.
2. Generalizing the positions of the trajectories in the same anonymization group in regions with regard to the quasi-identifiers of all moving

objects in the group. Given that the anonymization groups are overlapping, this must be done with care in order to avoid backtracking and revisiting of previously computed generalizations.

10.3.2 Other PETs for Off-line Mobility Data Analysis

PETs for mobility data publishing represent an important part of the literature in privacy in mobility data analysis but there are other interesting techniques that consider different scenarios, different settings and apply different privacy models. In the following we briefly review these techniques.

Distributed Privacy-preserving Mobility Data Mining

The methods belonging to this group aim at the analysis of datasets that are partitioned and distributed among several parties that do not want to (or cannot) share the data or certain corporate information that is represented in the data, but are interested in developing global models of common interest. Therefore, the main assumption in this scenario is that multiple data holders want to collaboratively perform data mining on the union of their data without revealing their sensitive information. The question addressed in these cases is how to compute the results without sharing the data in such a way that nothing is disclosed except the final result of the data mining result. This problem is addressed in cryptography in the field of *secure multi-party computation*. An example of problem tackled by this kind of approach is the privacy-preserving clustering in horizontally partitioned spatio-temporal data. Here, each horizontal partition contains trajectories of distinct moving objects collected by separate site and wants to cluster these trajectories without publishing sensitive location information to the other data holders. At the end of the protocol the global clustering results will be public to each data holder. The method used to achieve this goal is to construct the dissimilarity matrix of the trajectories in a privacy preserving manner which can be the input of any hierarchical clustering algorithm. In this setting there is a third party that has the following tasks: a) managing the communication among data holders; b) construct a global dissimilarity matrix; c) clustering the trajectories by using the dissimilarity matrix; and d) publishing the final result to the data holders. Each party involved is considered semi-trusted, in the sense that they follow the protocol as expected to, but cannot store any information

to infer sensitive data. Moreover, all parties do not share any sensitive information with each other.

As an example of application of this technique consider the case of a traffic control office that wants to solve the traffic congestion. To this aim, it has to cluster users' trajectories. Now, suppose that these spatio-temporal data are collected by GSM operators that cannot share these data due to privacy issues. The best solution in this context is to apply a privacy-preserving clustering algorithm for horizontally partitioned data that avoids the sharing of the spatio-temporal data.

Knowledge Hiding in Mobility Data

Knowledge hiding refers to the activity of hiding patterns, considered sensitive, in a database to be published. If the data is published as it is, the sensitive patterns may be surfaced by means of data mining techniques. Knowledge hiding is usually obtained by the *sanitization* of the database in such a way that the sensitive knowledge can no longer be inferred, while the original database is changed as less as possible. This problem is more interesting in the context of spatio-temporal patterns in a database of trajectories. Spatio-temporal geo-referenced traces (i.e., sequences of locations) left by mobile phones and other location-aware devices contain detailed information about personal and vehicular mobile behavior, and therefore offer interesting practical opportunities to find behavioral patterns, to be used for instance in traffic and sustainable mobility management, e.g., to study the accessibility to services. The collected mobility data contain also some typical mobile behaviors (i.e., frequent patterns), that are considered sensitive for political or security reasons, so it is necessary a method able to hide such sensitive patterns before the disclosure of the database. A valid hiding technique in this context has to take into consideration the road network, and therefore considering trajectories of objects moving over a background road network. This network is modeled as a directed graph. A privacy solution has to sanitizes the input trajectory database D in such a way that a set of sensitive spatio-temporal patterns P is hidden while the most information in D is maintained. The resulting database D', that is the released one, is consistent with the background road network. The privacy solution avoids creating unreal trajectories in the sanitization process, since the road network is a publicly available knowledge and thus unreal trajectories could be easily identified. The second requirement satisfied is that all sensitive patterns are hidden in D', i.e., they

have a support not more than the given disclosure threshold ψ. Finally, the third requirement is that D' is kept as similar as possible to D.

10.4 Privacy by Design in Data Mining

How showed in the previous sections, several techniques have been proposed by the scientific community to develop technological frameworks for countering the threats of undesirable and unlawful effects of privacy violation, without obstructing the knowledge discovery opportunities of data mining technologies. The common result obtained is that no general method exists which is capable of both dealing with "generic personal data" and preserving "generic analytical results". The ideal solution would be to inscribe privacy protection into the knowledge discovery technology by design, so that the analysis incorporates the relevant privacy requirements from the very start. Here, it is evoked the concept of Privacy by Design coined in the '90s by Ann Cavoukian, the Information and Privacy Commissioner of Ontario, Canada. In brief, Privacy by Design refers to the philosophy and approach of embedding privacy into the design, operation and management of information processing technologies and systems.

The articulation of the general "by design" principle in the data mining domain is that higher protection and quality can be better achieved in a goal-oriented approach. In such an approach, the data mining process is designed with assumptions about:

(a) the sensitive personal data that are the subject of the analysis;
(b) the attack model, i.e., the knowledge and purpose of a malicious party that has an interest in discovering the sensitive data of certain individuals;
(c) the category of analytical queries that are to be answered with the data.

Under the above assumptions, it is conceivable to design a privacy-preserving analytical process able to:

1. transform the data into an anonymous version with a quantifiable privacy guarantee - i.e., the probability that the malicious attack fails;
2. guarantee that a category of analytical queries can be answered correctly, within a quantifiable approximation that specifies the data utility, using the transformed data instead of the original ones.

The next section will provide an example of application of this methodology. Specifically, Section 10.4.1 shows the design of a privacy-preserving framework for the publication of raw movement data, while preserving clustering analysis.

10.4.1 Privacy by design for trajectory anonymization

In this section we present a framework that offers an instance of the privacy by design paradigm in the case of personal mobility trajectories (obtained from GPS devices or cell phones). The results show how such trajectories can be anonymized to a high level of protection against reidentification while preserving the possibility of mining clusters of trajectories, which enables novel powerful analytic services for infomobility or location-based services.

The application of the above methodology requires to understand: the specific properties of the trajectories to be protected; which characteristics it is necessary to preserve for guaranteeing a good quality of the clustering analysis that have to be performed on these data; and which adversary's knowledge the attacker may use for the user re-identification. Clearly, this information it is fundamental for the design of the data transformation technique.

Attack Model. In this framework the *linking attack model* is considered, i.e., the ability to link the published data to external information, which enables some respondents associated with the data to be reidentified. In relational data, linking is made possible by *quasi-identifiers*, i.e., attributes that, in combination, can uniquely identify individuals, such as birth date and gender (see Section 10.2). The remaining attributes represent the private respondent's information, that may be violated by the linking attack. In privacy-preserving data publishing techniques, such as k-anonymity, the goal is precisely to find countermeasures to this attack, and to release person-specific data in such a way that the ability to link to other information using the quasi-identifier(s) is limited. In the case of spatio-temporal data, where each record is a temporal sequence of locations visited by a specific person, the above dichotomy of attributes into quasi-identifiers (QI) and private information (PI) does not hold any longer: here, a (sub)trajectory can play both the role of QI and the role of PI. To see this point, consider the attacker may know a sequence of places visited by some specific person P: e.g., by shadowing P for some time, the attacker may learn that P was in the

shopping mall, then in the park, and then at the train station. The attacker could employ such knowledge to retrieve the complete trajectory of P in the released dataset: this attempt would succeed, provided that the attacker knows that P's trajectory is actually present in the dataset, if the known trajectory is compatible with (i.e., is a sub-trajectory of) just one trajectory in the dataset. In this example of a linking attack in the movement data domain, the sub-trajectory known by the attacker serves as QI, while the entire trajectory is the PI that is disclosed after the re-identification of the respondent. Clearly, as the example suggests, is rather difficult to distinguish QI and PI: in principle, any specific location can be the theater of a shadowing actions by a spy, and therefore any possible sequence of locations can be used as a QI, i.e., as a means for re-identification. Put another way, distinguishing between QI and PI among the locations means putting artificial limits on the attacker's background knowledge; on the contrary, it is required in privacy and security research to have assumptions on the attacker's knowledge that are as liberal as possible, in order to achieve maximal protection.

As a consequence of this discussion, it is reasonable to consider the radical assumption that any (sub)trajectory that can be linked to a small number of individuals is a potentially dangerous QI and a potentially sensitive PI. Therefore, in the *trajectory linking attack*, the malicious party M knows a subtrajectory of a respondent R (e.g., a sequence of locations where R has been spied on by M) and M would like to identify in the data the whole trajectory belonging to R, i.e., learn all places visited by R.

Privacy-preserving Technique. *How is it possible to guarantee that the probability of success of the above attack is very low while preserving the utility of the data for meaningful analyses?* Consider the source trajectories represented in Figure 10.6, obtained from a massive dataset of GPS traces (17,000 private vehicles tracked in the city of Milan, Italy during a week).

Each trajectory is a de-identified sequence of timestamped locations, visited by one of the tracked vehicles. Albeit de-identified, each trajectory is essentially unique – very rarely are two different trajectories exactly the same given the extremely fine spatio-temporal resolution involved. As a consequence, the chances of success for the trajectory linking attack are not low. If the attacker M knows a sufficiently long subsequence S of locations visited by the respondent R, it is possible that only a few trajectories in the dataset match with S, possibly just

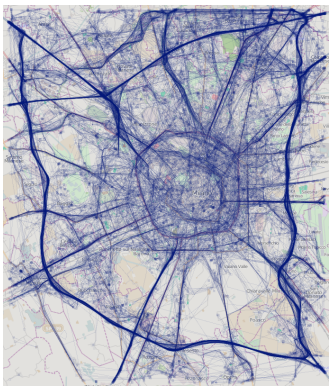


Figure 10.6 Milan GPS Trajectories

one. Indeed, publishing raw trajectory data such as those depicted in Figure 10.6 is an unsafe practice, which runs a high risk of violating the private sphere of the tracked drivers (e.g., guessing the home place and the work place of most respondents is very easy). Now, assume that one wants to discover the trajectory clusters emerging from the data through data mining, i.e., the groups of trajectories that share common mobility behavior, such as the commuters that follow similar routes in their homework and workhome trips. An anonymizing transformation of the trajectories consists of the following steps:

1. characteristic points are extracted from the original trajectories: starting points, ending points, points of significant turn, points of significant stop (Figure 10.7(a));
2. characteristic points are clustered into small groups by spatial proximity (Figure 10.7(b));
3. the central points of the groups are used to partition the space by means of Voronoi tessellation (Figure 10.7(c));
4. each original trajectory is transformed into the sequence of Voronoi cells that it crosses (Figure 10.7(d)).

As a result of this data-driven transformation, where trajectories are generalized from sequences of points to sequences of cells, the probability of re-identification already drops significantly. Further techniques can be adopted to lower it even more, obtaining a safe theoretical upper bound for the worst case (i.e., the maximal probability that the linking attack succeeds), and an extremely low average probability. A possible tech-

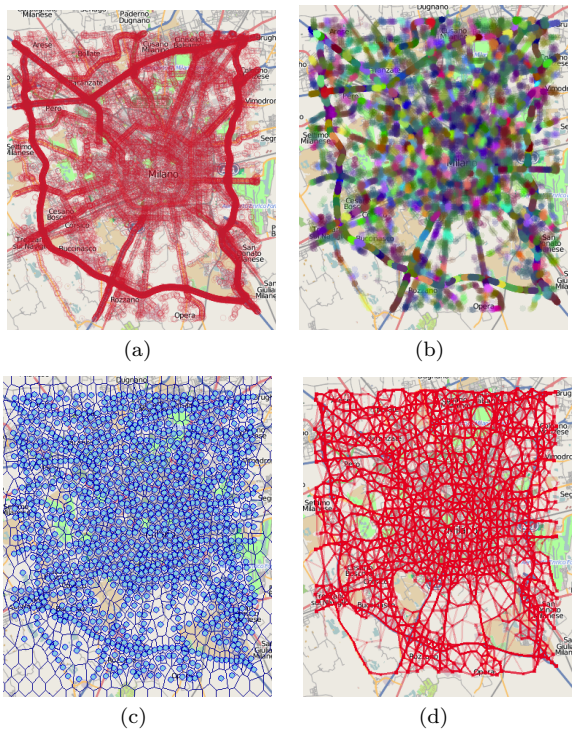


Figure 10.7 (a) characteristic points (b) spatial clusters (c) tessellation of the territory (d) generalized trajectories

nique is to ensure that for any sub-trajectory used by the attacker, the re-identification probability is always controlled below a given threshold $1/k$; in other words, ensuring the k-anonymity property in the released dataset. Here, the notion of k-anonymity proposed is based on the definition of *k-harmful trajectory,* i.e., a trajectory occurring in the database with a frequency less than k. Therefore, a trajectory database D^* is considered a k-anonymous version of a database D if: each k-harmful trajectory in D is frequent at least k times in D^* or if it does not appear in D^* anymore. To achieve this k-anonymous database the generalized trajectories, obtained after the data-driven transformation, are transformed in such a way that all the k-harmful sub-trajectories in D are not k-harmful in D^*.

In the example in Figure 10.6 the probability of success is theoretically

bounded by 1/20 (i.e., it is achieved 20-anonymity), but the real upper bound for 95% of the attacks is below 10^{-3}.

Clustering Analysis. The above results indicate that the transformed trajectories are orders of magnitude safer than the original data in a measurable sense: *but are they still useful to achieve the desired result, i.e., discovering trajectory clusters?*

Figure 10.8 and Figure 10.9 illustrate the most relevant clusters found by mining the original trajectories and the anonymized trajectories, respectively.

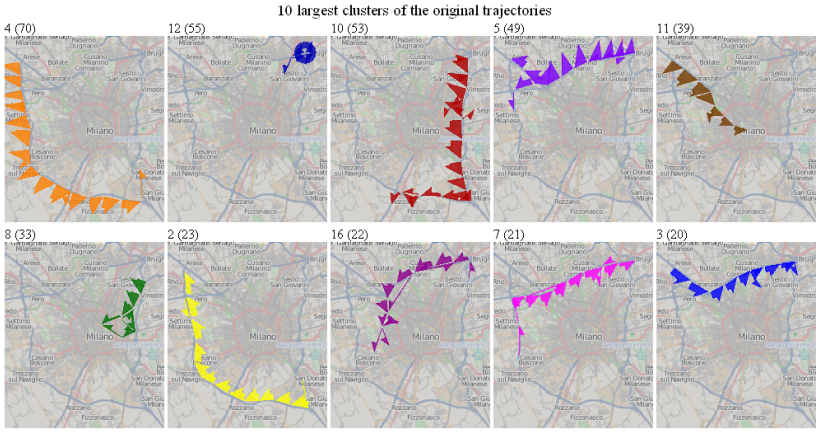


Figure 10.8 10 largest clusters of the original trajectories.

,

A direct effect of the anonymization process is an increase in the concentration of trajectories (i.e. several original trajectories are bundled on the same route); the clustering method will thus be influenced by the variation in the density distribution. The increase in the concentration of trajectories is mainly caused by the reduction of noisy data. In fact, the anonymization process tends to render each trajectory similar to the neighboring ones. This means that the original trajectories, initially classified as noise, can now be "promoted" as members of a cluster. This phenomenon may produce an enlarged version of the original clusters. To evaluate the clustering preservation quantitively the F-measure is adopted. The F-measure is usually adopted to express the combined values of precision and recall and is defined as the harmonic mean of

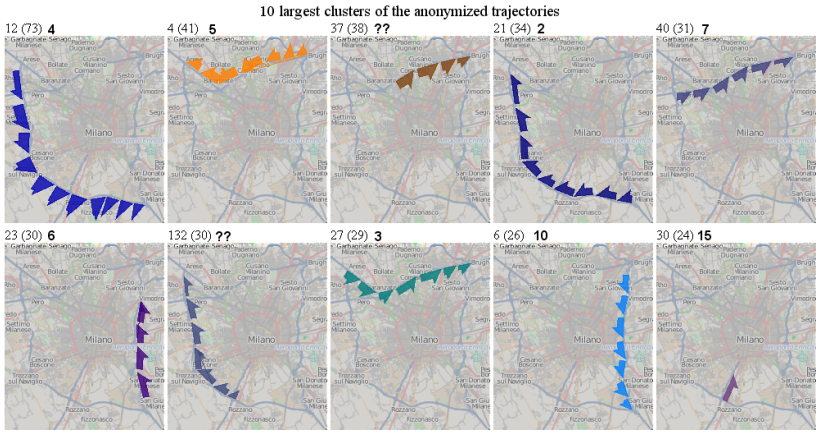


Figure 10.9 10 largest clusters of the anonymized trajectories.

the two measures. Here, the recall measures how the cohesion of a cluster is preserved: it is 1 if the whole original cluster is mapped into a single anonymized cluster, it tends to zero if the original elements are scattered among several anonymized clusters. The precision measures how the singularity of a cluster is mapped into the anonymized version: if the anonymized cluster contains only elements corresponding to the original cluster its value is 1, otherwise the value tends to zero if there are other elements corresponding to other clusters. The contamination of an anonymized cluster may depend on two factors: (i) there are elements corresponding to other original clusters or (ii) there are elements that were formerly noise and have been promoted to members of an anonymized cluster.

The immediate visual perception that the resulting clusters are very similar in the two cases in Figures 10.8 & 10.9 is also confirmed by various cluster comparison by Fmeasure, re-defined for clustering comparison (Figure 10.10).

The conclusion is that in the illustrated process the desired quality of the analytical results can be achieved in a privacy-preserving setting with concrete formal safeguards and the protection with respect to the linking attack can be measured.

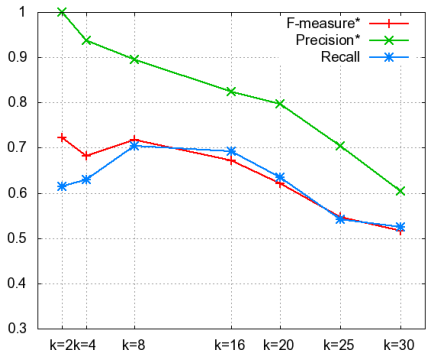


Figure 10.10 10 largest clusters of the anonymized trajectories.

10.5 Conclusion

Mobility data represent an important source of knowledge but the sharing of these data can raise serious privacy concerns: mobility data may potentially reveal many facets of private life person. Mobility data privacy problems have to be addressed in two different scenarios: on-line location-based services and off-line data analysis context. Many recent research works have focused on the study of privacy protection in spatio-temporal data and many privacy enhancing technologies have been proposed, which essentially aim at finding an acceptable trade-off between data privacy on the one hand and data utility on the other. So far, the common result obtained is that no general method exists which is capable of both dealing with "generic personal data" and preserving"generic analytical results". A recent paradigm, called *privacy by design*, promises a quality leap in the conflict between data protection and data utility. The application of this paradigm in mobility data mining showed that the desired quality of the analytical results can be achieved in a privacy-preserving setting with concrete formal safeguards and the protection with respect to the linking attack can be measured. The implication of this finding is far reaching; once an analytical process has been found and specified, it can be deployed and replicated with the mentioned privacy-preserving safeguards in order to perform mobility data analysis in different periods of time, in different cities, in different contexts: once deployed, it is a safe service that generates knowledge of the expected quality starting from truly anonymous data.

10.6 Bibliographic Notes

The literature on privacy in mobility data is becoming extensive. In the following, we will provide an essential list of bibliographic references for the reader, including those describing the problems and the solutions discussed in the chapter.

Privacy issues in mobility data mining were deeply discussed in the the book by Giannotti and Pedreschi (2008). Anna Monreale and G.Pensa (2010) proposes an overview on the main privacy-preserving data publishing and mining techniques proposed by the data mining community and by the statistical disclosure control community. This contribution also discusses the privacy issues in complex domains, focusing the attention on the context of spatio-temporal data and describes some approaches proposed for anonymity of this type of data.

The k-anonymity model was introduces by Samarati and Sweeney (1998) and then,Machanavajjhala et al. (2007) and Li et al. (2007) proposed l-diversity and t-closeness to overcome the weaknesses of the k-anonymity. This privacy model and its variants have been widely adopted to achieve privacy in mobility data, especially in privacy-preserving publishing of trajectories. A recent survey on trajectory anonymity publishing is presented by Bonchi et al. (2011).

The problem of hiding sensitive spatio-temporal patterns in a trajectory data was studied in Abul et al. (2010), while a privacy-preserving clustering method in horizontally partitioned spatio-temporal data was addressed by Inan and Saygin (2006).

The *privacy by design* paradigm in data mining was introduced by Monreale (2011). This PhD thesis proposed this novel methodology to address the privacy issues in complex data with a particular focus on data with a sequential nature such as trajectory data.

References

Abul, Osman, Bonchi, Francesco, and Giannotti, Fosca. 2010. Hiding Sequential and Spatiotemporal Patterns. *IEEE Trans. Knowl. Data Eng.*, **22**(12), 1709–1723.

Anna Monreale, Dino Pedreschi, and G.Pensa, Ruggero. 2010. Anonymity Technologies for Privacy-Preserving Data Publishing and Mining. Pages 3–33 of: *Privacy-Aware Knowledge Discovery: Novel Applications and New Techniques, Chapman Hall/CRC Press.*

Bonchi, Francesco, Lakshmanan, Laks V. S., and Wang, Wendy Hui. 2011. Trajectory anonymity in publishing personal mobility data. *SIGKDD Explorations*, **13**(1), 30–42.

Giannotti, Fosca, and Pedreschi, Dino (eds). 2008. *Mobility, Data Mining and Privacy - Geographic Knowledge Discovery.* Springer.

Inan, Ali, and Saygin, Yücel. 2006. Privacy Preserving Spatio-Temporal Clustering on Horizontally Partitioned Data. Pages 459–468 of: *DaWaK.*

Li, Ninghui, Li, Tiancheng, and Venkatasubramanian, Suresh. 2007. t-Closeness: Privacy Beyond k-Anonymity and l-Diversity. Pages 106–115 of: *ICDE.*

Machanavajjhala, Ashwin, Kifer, Daniel, Gehrke, Johannes, and Venkitasubramaniam, Muthuramakrishnan. 2007. *L*-diversity: Privacy beyond *k*-anonymity. *TKDD*, **1**(1).

Monreale, Anna. 2011. *Privacy by Design in Data Mining.* Ph.D. thesis, Department of Computer Science, University of Pisa, Italy.

Samarati, Pierangela, and Sweeney, Latanya. 1998. Generalizing Data to Provide Anonymity when Disclosing Information (Abstract). Page 188 of: *PODS.*