# Chapter 10
# Privacy by Design for Mobility Data Analytics


**Francesca Pratesi, Anna Monreale, and Dino Pedreschi**

**Abstract** Privacy is an ever-growing concern in our society and is becoming a fundamental aspect to take into account when one wants to use, publish and analyze data involving human personal sensitive information, like data referring to individual mobility. Unfortunately, it is increasingly hard to transform the data in a way that it protects sensitive information: we live in the era of big data characterized by unprecedented opportunities to sense, store and analyze social data describing human activities in great detail and resolution. This is especially true when we work on mobility data, that are characterized by the fact that there is no longer a clear distinction between quasi-identifiers and sensitive attributes. Therefore, protecting privacy in this context is a significant challenge. As a result, privacy preservation simply cannot be accomplished by de-identification alone. In this chapter, we propose the Privacy by Design paradigm to develop technological frameworks for countering the threats of undesirable, unlawful effects of privacy violation, without obstructing the knowledge discovery opportunities of social mining and big data analytical technologies. Our main idea is to inscribe privacy protection into the knowledge discovery technology by design, so that the analysis incorporates the relevant privacy requirements from the start. We show three applications of the Privacy by Design principle on mobility data analytics. First we present a framework based on a data-driven spatial generalization, which is suitable for the privacy-aware publication of movement data in order to enable clustering analysis. Second, we present a method for sanitizing semantic trajectories, using a generalization of visited places based on a taxonomy of locations. The private data then may be used for extracting frequent sequential patterns. Lastly, we show how to apply the idea of Privacy by Design in a distributed setting in which movement data from

F. Pratesi (✉)
ISTI-CNR of Pisa, Pisa, Italy
Computer Science Department of University of Pisa, Pisa, Italy
e-mail: francesca.pratesi@isti.cnr.it

A. Monreale · D. Pedreschi
Computer Science Department of University of Pisa, Pisa, Italy
e-mail: annam@di.unipi.it; pedre@di.unipi.it



A. Gkoulalas-Divanis, C. Bettini (eds.), *Handbook of Mobile Data Privacy*,
https://doi.org/10.1007/978-3-319-98161-1_10

individual vehicles is made private through differential privacy manipulations and then is collected, aggregated and analyzed by a centralized station.

## 10.1 Introduction

The big data originating from the digital breadcrumbs of human activities, generated by ICT systems that we use every day, record the multiple facets of social life: automated payment systems record the tracks of our purchases; search engines record the logs of our queries for finding information on the web; social networks record our connections and communications with friends and colleagues; mobile devices record the traces of our movements. These big data are at the heart of the vision of a "knowledge society", where the understanding of social phenomena is sustained by the knowledge extracted from data describing human activities across the various social dimensions by using social mining technologies. Thus, the analysis of our digital traces can create new opportunities to understand complex aspects, such as mobility behaviors, economic and financial crises, the spread of epidemics, the diffusion of opinions and so on.

The worrying side of the story is that big data contain personal sensitive information, so that the occasions of discovering knowledge increase with the risks of privacy violation. Indeed, when personal sensitive data are published and/or analyzed, it must be checked if this may violate the right of individuals to have full control of their personal sphere. It is clear that maintaining control of personal data is increasingly difficult and it cannot simply be achieved by de-identification (i.e., by removing the direct or explicit identifiers contained in the data, such as name, address and phone number [33]).[1] In the scientific literature and in the media, many examples of re-identification from supposedly anonymous data have been reported, from health records to querylogs to GPS trajectories. In the past years, several techniques have been developed for countering privacy violations, without losing the benefits of big data analytics technology [4, 12, 22, 28, 34]. Despite these results, no general method exists that is able of handling both general personal data and preserving general analytical results. Anonymity in a global sense is believed to be a chimera, and the concern about infringement of the private sphere by means of big data is now in news headlines of major media. Nevertheless, big data analytics and privacy are not necessarily enemies: the goal of this chapter is exactly to show that practical and effective services based on big data analytics can be proposed in such a way that the quality of results can coexist with high protection of personal data. The magic word is Privacy by Design. Here, we review a methodology for purpose-driven privacy protection, where the purpose is a target knowledge service to be deployed on top of data analysis. The key observation is that providing a

[1] This definition of de-identified data is compliant with the General Data Protection Regulation (GDPR) [18], especially referring to Recital 26. Indeed, with the de-identification process we are going to transform identified persons in identifiable persons.

reasonable trade-off between a measurable protection of individual privacy together with a measurable quality of service is unfeasible in general, but it becomes feasible in context, i.e., if we have a previous knowledge of the desired analytical goal and the expected level of privacy.

In this chapter, we elaborate on the above ideas the Privacy by Design paradigm, introduced by Anne Cavoukian, in the 1990s, to deploy big data analytical services. Firstly, we discuss the Privacy by Design principle highlighting and how it has been embraced by the United States and Europe.
Secondly, we introduce the idea of Privacy by Design in mobility data analytics domain and show how inscribing privacy "by design" in three different specific scenarios assuring a good balance between privacy protection and quality of data analysis. As first example, we present a framework based on a data-driven spatial generalization, which is suitable for the privacy-aware publication of movement data in order to enable clustering analysis [23]. Then, we present a method for sanitizing semantic trajectories [25], using a generalization of visited places based on a taxonomy of locations. The private data then may be used for extracting frequent sequential patterns.
Lastly, we show how to apply the idea of Privacy by Design in a distributed setting in which movement data from individual vehicles is made private through differential privacy manipulations and then is collected, aggregated and analyzed by a centralized station [26].

The remaining of the chapter is organized as follows. In Sect. 10.2 we discuss the Privacy by Design paradigm and its articulation in data analytics. Sections 10.3 and 10.4 discuss the application of the Privacy by Design principle in the case of publication of personal mobility trajectories, regarding clustering analyses and semantic trajectories respectively, while in Sect. 10.5 we show a possible distributed scenario for privacy preserving mobility analytics. Lastly, Sect. 10.6 concludes the chapter.

## 10.2 Privacy by Design

Privacy by Design is a paradigm developed by Dr. Ann Cavoukian, the former Ontario's Information and Privacy Commissioner, in the 1990s, to address the emerging and growing threats to online privacy. The key idea of this model is to inscribe the privacy protection into the design of information technologies from the very start. It represents a significant innovation w.r.t. traditional privacy protection approaches since it requires a significant shift from a reactive model to a proactive one. In other words, the idea is preventing privacy issues, instead of remedying to them.

Given the ever growing availability and diffusion of big data and also the impact of big data analytics on both human privacy risks and the possibility of comprehending relevant phenomena, many companies are understanding the necessity to consider privacy at every stage of their business and, thus, to integrate

privacy requirements "by design" into their business model. Unfortunately, in many contexts, it is not completely clear which are the methodologies for incorporating Privacy by Design.

### *10.2.1 Privacy by Design in Law*

The Privacy by Design paradigm has been recognized in legislation, and in the last years, privacy officials in Europe and the United States are embracing this attitude.

In 2010, at the annual conference of "Data Protection and Privacy Commissioners" the International Privacy Commissioners and Data Protection Authorities approved a resolution recognizing Privacy by Design as an *essential component of fundamental privacy protection* [1] and advocates the adoption of this principle as part of an organization's default mode of operation. In 2009, the EU Article 29 Data Protection Working Party and the Working Party on Police and Justice released a joint Opinion, encouraging the incorporation of Privacy by Design principles into a new EU privacy framework [2]. In March 2010, the European Data Protection Supervisor advocated to "include unequivocally and explicitly the principle of Privacy by Design into the existing data protection regulatory framework" [17]. This recommendation was taken into consideration in the reform of Data Protection Rules, entered into force on 5 May 2016. Indeed, in this new European Directive [3], in particular in Article 20, there is an explicit reference to data protection "by design" and "by default".

Privacy by Design has been embraced with the same enthusiasm in the United States. Indeed, the U.S. Federal Trade Commission hosted a series of public discussions on privacy issues in the digital age and in a recent staff report [19] it describes a proposed framework with three main recommendations: *privacy by design*, *simplified consumer choice*, and *increased transparency of data practices*. Moreover, in April 2011, Senators John Kerry (D-MA) and John McCain (R-AZ) proposed their legislation entitled "Commercial Privacy Bill of Rights Act of 2011" that would require companies that collect, use, store or transfer consumer information to implement a version of Privacy by Design when developing products.

### *10.2.2 Privacy by Design in Big Data Analytics and Social Mining*

Unfortunately, it is not always clear what means applying the Privacy by Design principle and which is the best way to apply it for obtaining the desired result. In this section, we discuss the articulation of the general "by design" principle in the big data analytics domain.

Our key idea is to consider privacy protection into any analytical process by design, so that the analysis incorporates the relevant privacy requirements from the very start, evoking the concept of Privacy by Design discussed above.

The application of the general "by design" principle in the big data analytics domain is based on a key concept: higher protection and quality can be better achieved in a goal-oriented approach. Indeed, the data analytical process is designed with assumptions about:

(a) the sensitive personal data 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.

These assumptions are essential for the design of a privacy-aware technology. First of all, privacy preservation techniques strongly depend on the nature of the data to be protected, e.g., an algorithm suitable for social networking data could not be appropriate for trajectory data. Second, a valid framework has to define the attack model based on a specific adversary's background knowledge and correspondent countermeasure: different assumptions on the background knowledge require different defense strategies. For example, an attacker could possess an approximated information about the mobility behavior of an individual and exploit it to infer all his movements. It is worth noting that a defense strategy designed for counter attacks with approximate knowledge could be too weak in case of more detailed knowledge. Finally, a privacy-aware solution should find an acceptable trade-off between data privacy and data utility. Thus, it is fundamental to consider the kind of analytical queries to be answered for understanding which data properties must be preserved. As an example, a defense strategy for spatio-temporal data should consider that these data might be useful for collective mobility analyses in an urban area.

Under the above hypotheses, we claim that it is possible to design a privacy-aware analytical process able to:

1. transform the data into an anonymous version with a quantifiable privacy guarantee, i.e., measuring 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.

We want to point out that different legal frameworks could imply different techniques that are considered to be sufficient for data protection. To define an adequate anonymization level, we mainly rely on the GDPR [18]. Indeed, Privacy by Design is compliant with the GDPR also regarding the principle of reasonableness stated in GDPR (Article 26), where it is stated that "to determine whether a natural person is identifiable, account should be taken of all the means reasonably likely to be used", where the reasonableness should consider some objective factors, such as the costs and the amount of time required for identification, taking into consideration the available technology and technological developments.

In the following, we show three existing ways to apply the Privacy by Design for the design of the same amount of analytical frameworks: one for clustering analysis, one for the publication of trajectory data and one for computing aggregation of movement data in a distributed setting. In the three scenarios we first analyze the privacy issues related to this kind of data; second, we identify the attack model; and third, we provide a method for assuring data privacy taking into consideration the data analysis to be maintained valid. However, these are not the unique privacy-preserving frameworks adopting the Privacy by Design principle, many approaches proposed in the literature can be seen as instances of this promising paradigm (see [4, 12, 27, 28, 34]).

## 10.3 Privacy by Design in Mobility Data Publishing

In this section, we present a framework that offers an instance of the privacy by design paradigm concerning personal mobility trajectories, obtained from GPS devices or cell phones [23]. It is convenient for the privacy-aware publication of movement data, and its focus is on clustering analysis useful for the comprehension of human mobility behavior in specific urban areas. The released trajectories are made anonymous by a suitable process that realizes a generalized version of the original trajectories.

In the following, we consider a mobility dataset as a collection of trajectories $D = \{T_1, T_2, \ldots, T_m\}$ where each $T_i$ is a trajectory represented by a sequence of spatio-temporal points.

**Definition 10.1 (Trajectory)** A Trajectory or spatio-temporal sequence is a sequence of triples $T = \langle x_1, y_1, t_1\rangle, \ldots, \langle x_n, y_n, t_n\rangle$, where $t_i$ $(i = 1 \ldots n)$ denotes a timestamp such that $\forall_{1<i<n}\ t_i < t_{i+1}$ and $(x_i, y_i)$ are points in $\mathbf{R}^2$.

Intuitively, each triple $\langle x_i, y_i, t_i\rangle$ indicates that the object is in the position $(x_i, y_i)$ at time $t_i$.

**Definition 10.2 (Sub-Trajectory)** Let $T = \langle x_1, y_1, t_1\rangle, \ldots, \langle x_n, y_n, t_n\rangle$ be a trajectory. A trajectory $S = \langle x'_1, y'_1, t'_1\rangle, \ldots, \langle x'_m, y'_m, t'_m\rangle$ is a *sub-trajectory* of $T$ or *is contained* in $T$ $(S \preceq T)$ if there exist integers $1 < i_1 < \ldots < i_m <= n$ such that $\forall 1 \leq j \leq m\ \langle x'_j, y'_j, t'_j\rangle = \langle x_{i_j}, y_{i_j}, t_{i_j}\rangle$.

We use $g$ to denote the function that applies the spatial generalization to a trajectory. Given a trajectory $T \in D$, the generalized version of $T$ is generated by a function $g$ that applies the spatial generalization to the trajectory. It is represented by the centroid sequence of areas crossed by $T$. More formally,

**Definition 10.3 (Generalized Trajectory)** Let $T = \langle x_1, y_1, t_1\rangle, \ldots, \langle x_n, y_n, t_n\rangle$ a trajectory. A generalized version of $T$ is a sequence of pairs $T_g = \langle x_{c_1}, y_{c_1}\rangle, \ldots, \langle x_{c_m}, y_{c_m}\rangle$ with $m <= n$ where each $x_{c_i}, y_{c_i}$ is the centroid of an area crossed by $T$.

The privacy by design framework presented in the following is based on a data-driven spatial generalization of the dataset of trajectories and the obtained results put in evidence how trajectories can be anonymized to a high level of protection against re-identification attacks, preserving, at the same time, the possibility of mining clusters of trajectories, which enables powerful analytic services for info-mobility or location based services.

### *10.3.1 Attack and Privacy Model*

Here, it is evaluated the *linkage attack model*, i.e., the ability to link the published data to external information, which enables some respondents associated with the data to be re-identified. 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 [30]. The remaining attributes represent the private respondent's information (PI), sometimes called sensitive attributes (SA), that may be violated by the linkage attack. In privacy-preserving data publishing techniques, such as $k$-anonymity, the goal is to find countermeasures to this particular 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. However, in the case of mobility 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 anymore: here, a (sub)trajectory can play both the role of QI and the role of PI. To understand this point, assume 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 gym, and then at the pub. The adversary could exploit 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 and the known sub-trajectory is compatible with (i.e., is a sub-trajectory of) just one trajectory in the dataset. In this example of a linkage 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, it is rather difficult to distinguish QI and PI: in principle, any specific location can be the theater of a shadowing action by a spy, and therefore any possible sequence of locations can be used as a QI, i.e., as a means for re-identification. As a consequence of this remark, it is reasonable to contemplate 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 linkage attack*, the adversary $M$ knows a sub-trajectory of a respondent $R$ (e.g., a sequence of locations where $R$ has been seen by $M$) and $M$ would try to discover the whole trajectory belonging to $R$ in the data, i.e., learn all places visited by $R$. In particular, we assume the following adversary knowledge.

**Definition 10.4 (Adversary Knowledge)** The attacker has access to the anonymized dataset $D^*$ and knows: (a) the details of the schema used to anonymize the data, (b) the fact that a given user $R$ is in the mobility dataset $D$ and (c) a sub-trajectory $S$ relative to $R$.

This background knowledge is used in the following attack.

**Definition 10.5 (Attack Model)** Given the anonymized dataset $D^*$ and a sub-trajectory $S$ relative to a user $R$, the attacker: *(i)* generates the partition of the territory starting from the trajectories in $D^*$; *(ii)* computes $g(S)$ generating the sequence of centroids of the areas containing the points of $S$; *(iii)* constructs a set of candidate trajectories in $D^*$ containing the generalized sub-trajectory $g(S)$ and tries to identify the whole trajectory relative to $R$. The probability of identifying the whole trajectory by a sub-trajectory $S$ is denoted by $prob(S)$.

### *10.3.2 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 Fig. 10.1a, obtained from a massive dataset

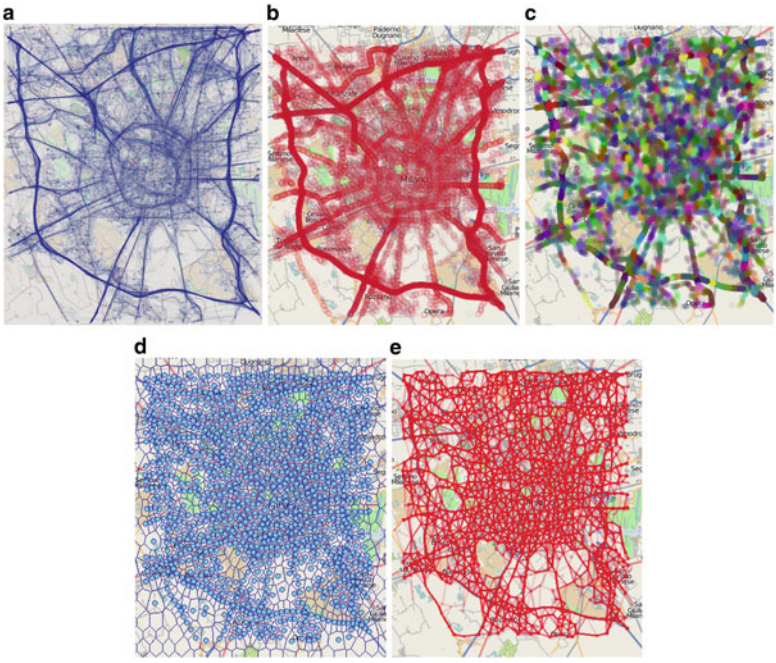


**Fig. 10.1** (**a**) Milan GPS Trajectories, (**b**) characteristic points, (**c**) spatial clusters, (**d**) tessellation of the territory, and (**e**) generalized trajectories

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 time-stamped locations, visited by one of the tracked individuals or vehicles. Although de-identified, each trajectory is essentially unique—two different trajectories seldom are exactly the same, due to the extremely fine spatio-temporal resolution involved. Therefore, the chances of success for this attack are not low. If the attacker $M$ has access to a sufficiently long sub-sequence $S$ of locations visited by the individual $R$, it is possible that only a few trajectories in the dataset match with $S$, possibly just one. Indeed, publishing raw trajectory data such as those depicted in Fig. 10.1a is an unsafe practice, which leads to a high risk of infringement on 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 sharing common mobility behavior, such as the commuters following similar routes in their home-work and work-home trips. A privacy 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 (Fig. 10.1b);
2. characteristic points are clustered into small groups by spatial proximity (Fig. 10.1c);
3. the central points of the groups are used to partition the space by means of Voronoi tessellation (Fig. 10.1d);
4. each original trajectory is transformed into the sequence of Voronoi cells that it crosses (Fig. 10.1e).

As a consequence of this data-driven transformation, where trajectories are generalized from sequences of points to sequences of cells, the re-identification probability already drops significantly. Further transformation can be applied to lower this probability even more, obtaining a safe theoretical upper bound for the worst case (i.e., the maximal probability that the linkage attack succeeds), and an extremely low average probability. A possible technique is to ensure that for any sub-trajectory used by the attacker, the re-identification probability is always controlled below a given threshold $\frac{1}{k}$; in other words, ensuring the $k$-anonymity property in the released dataset. Here, the notion of $k$-anonymity is based on the definition of *$k$-harmful trajectory,* i.e., a trajectory occurring in the database with a frequency less than $k$. Thus, a trajectory database $D^*$ is considered a $k$-anonymous version of a database $D$ if: each $k$-harmful trajectory in $D$ appears at least $k$ times in $D^*$ or if it does not appear in $D^*$ anymore. To obtain this $k$-anonymous database, the generalized trajectories, produced 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 shown in Fig. 10.1a, the probability of success is theoretically bounded by $\frac{1}{20}$ (i.e., 20-anonymity is achieved), but the real upper bound for 95% of attacks is below $10^{-3}$.

### *10.3.3 Analytics Quality*

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

Figure 10.2(top) and (down) listed the most relevant clusters found by mining the original trajectories and the anonymized trajectories, respectively.

A direct consequence of the anonymization process is an increase in the concentration of trajectories, i.e., many original trajectories are bundled on the

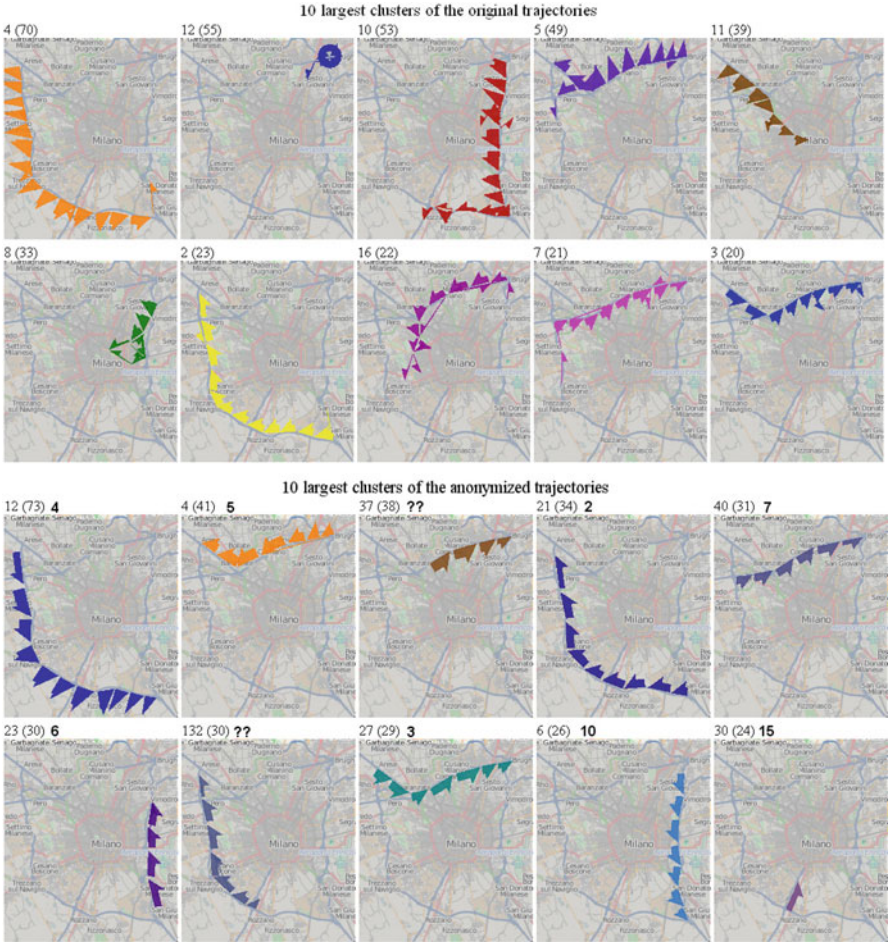


**Fig. 10.2** 10 largest clusters of the original trajectories (top) and of the anonymized trajectories (down)

same route; the clustering method will be influenced by the variation in the density distribution. This change is mainly caused by the reduction of noisy data. In fact, the anonymization procedure 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. F-measure is adopted to evaluate quantitively the clustering preservation. This measure is usually used to express the combined values of precision and recall and is defined as the harmonic mean of the two measures. Here, the recall measures how the cohesion of a cluster is preserved: if the whole original cluster is mapped into a single anonymized cluster its value is 1; otherwise, the value tends to zero if the original elements are scattered among several anonymized clusters. The precision indicates how the singularity of a cluster is mapped into the anonymized version: it is 1 if the anonymized cluster contains only elements corresponding to the original cluster, it tends to zero if there are other elements corresponding to other clusters. The contamination of an anonymized cluster may depend on two factors: (1) there are elements corresponding to other original clusters or (2) 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 Fig. 10.2(top) and (down) is also confirmed by various cluster comparisons by F-measure, re-defined for clustering comparison (Fig. 10.3).

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 guarantees and the protection w.r.t. the linkage attack can be quantified.

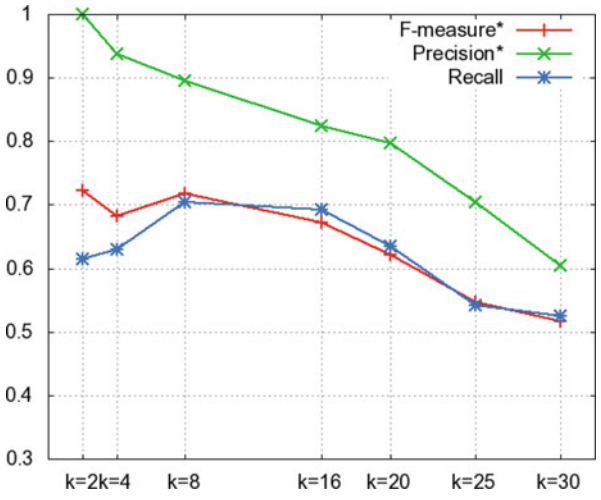


**Fig. 10.3** F-measure for comparison of the clusterings of the anonymized dataset versus the clustering of the original trajectories

## 10.4 Privacy by Design in Semantic Trajectories Anonymization

In this section, we present a framework that offers an instance of the Privacy by Design paradigm concerning mobility trajectories enriched with semantic information, i.e., *semantic trajectories* introduced in [31] for reasoning over trajectories from a semantic point of view.

In detail, a semantic trajectory is a sequence of stops and moves of an individual during her movement. 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. Each location of the 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*.

Important parts of a trajectory, i.e., stops, correspond to the set of $x$, $y$, $t$ points of a trajectory that are important from an application point of view. A set of important places characterizes a semantic trajectory.

**Definition 10.6 (Semantic Trajectory)** Given a set of important places $\mathscr{I}$, a *semantic trajectory* $T = p_1, p_2, \ldots, p_n$ with $p_i \in \mathscr{I}$ is a temporally ordered sequence of important places, that the moving object has visited.

The Privacy by Design framework presented in this section (introduced in [25]) enables sophisticated reasoning on the scope of people's movements by maintaining under control the individual privacy. In particular, the released semantic trajectories are made *safe* concerning the inference of sensitive information derived from the knowledge of the reason of the individual's movement and from the knowledge of the place that the individual visited. The framework is based on a data transformation that generalizes places driven by a place taxonomy, thus providing a way to preserve the semantics of the generalized trajectories.

The results obtained with the application of this framework show how it possible to preserve the semantics of trajectories making them useful for extracting valid mobility semantic patterns while guaranteeing the limitation of sensitive information inferences from the individual visits.

### *10.4.1 Attack and Privacy Model*

The use of a domain taxonomy for generalizing places enables the identification of *sensitive* and *non-sensitive* places. A place is considered sensitive when it allows inferring personal information about the individual who has stopped there. For example, a stop at an oncology clinic may indicate that the user has some health problem. Other places (such as parks, restaurants, cinemas, etc.) are considered as

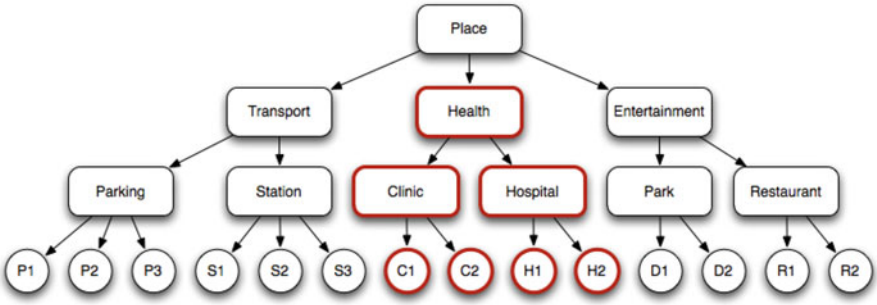


**Fig. 10.4** The places taxonomy

*quasi-identifiers*. The labeled taxonomy is given by the domain expert who tags each concept with the corresponding "sensitivity" label.

In this context, the attack model considers an attacker with the following adversary knowledge:

**Definition 10.7 (Adversary Knowledge)** The attacker has access to the generalized dataset $D^*$ 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 $R$.

In this model, the idea is to keep private all the sensitive places visited by a given user. As a consequence, the attack model considers the ability to link the released data to other external information enabling the inference of visited sensitive places.

In practice, given the quasi-identifier sequence $S_Q$, the attacker constructs a set of candidate semantic trajectories in $D^*$ containing $S_Q$ and tries to infer the sensitive leaf places related to $R$. $Prob(S_Q, S)$ denotes the probability that, given a quasi-identifier place sequence $S_Q$ related to a user $R$, the attacker infers his/her set of sensitive places $S$ which are the leaves of the taxonomy $PTax$. An example of labelled taxonomy is depicted in Fig. 10.4.

### *10.4.2 Privacy-Preserving Technique*

*How to guarantee that the probability of success of the above attack is very low while preserving the utility of the data for meaningful analyses?* From a data protection perspective it is necessary to control the probability $Prob(S_Q, S)$ and a solution is to release a $c$-safe dataset, i.e., a dataset where 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 $c \in [0, 1]$. On the contrary, for a data utility point of view, a data analyst might use the semantic trajectories to extract common and frequent

human behaviors by sequential pattern mining analyses, having in this way the possibility to reason on the semantic of the human movements. Therefore, we need a privacy transformation that tries to minimize the cost of a trajectory generalization. A privacy transformation of semantic trajectories consists of the following steps:

1. suppressing from the original semantic trajectories each *sensitive place* when, for that given user, that place is a *quasi-identifier*;
2. grouping semantic trajectories in groups of a predefined size, $m$;
3. 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. The generalization is performed with the support of the taxonomy $PTax$.

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. In order to consider the generalization cost it is possible to use distance functions that measure the cost to transform an original semantic trajectory into a generalized one, based on the taxonomy. A measure might be the distance in steps from two places in the taxonomy tree, the so called *Hops-based distance*.

If we consider the dataset in Fig. 10.1a, after the privacy transformation where the probability of success is theoretically bounded by 0.3 we have an empirical upper bound of 0.07 in average on 10,000 attacks using as background knowledge 5 places (see Fig. 10.5).

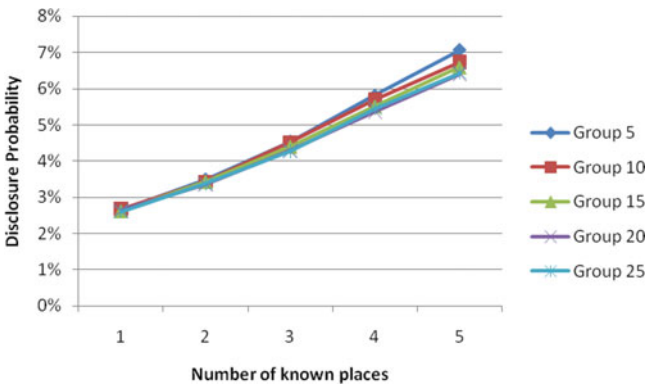


**Fig. 10.5** The empirical disclosure probability on Milano dataset

### *10.4.3 Analytics Quality*

Now, the point is to understand if the guaranteed privacy protection also allows the possibility to perform some analysis based on the sequential pattern mining extractable from the *c*-safe data. To evaluate this, it is necessary to measure the quality of the sequential patterns. Figure 10.6a shows the effect of the privacy transformation on the number of patterns extractable from the dataset after the sanitization.The figure highlights the fact that the generalization has a double effect on the patterns: (1) the frequency of generalized places increases, (2) 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. Figure 10.6b depicts instead the trend of the *coverage coefficient*. This index measures how many

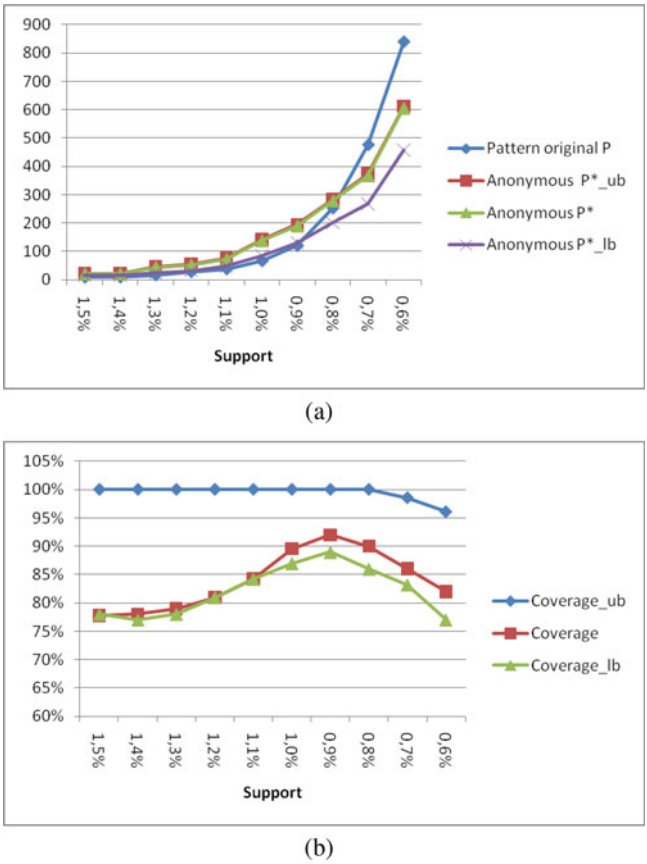


**Fig. 10.6** (**a**) Number of patterns extracted from Milan data and (**b**) coverage of the patterns varying the support threshold

patterns extracted from the original dataset are covered at least by the patterns extracted from the anonymized dataset with a certain level of generalization. It is important to notice that the coverage does not measure *how much* the patterns are generalized, but only if they are covered by a pattern obtained from the anonymized dataset or not. The results highlight that the coverage guaranteed by the patterns after the privacy transformation is not 100% but the levels are high enough to enable analyses; in fact, by changing the support (i.e., the minimum frequency used for the pattern extraction) the coverage is always greater than 75%.

## 10.5 Privacy by Design in Distributed Systems

The previous scenarios (Sects. 10.3 and 10.4) are based on centralized environments, where the privacy preservation step is performed by a central entity; in fact, we showed two variants of $k$-anonymity which can be used only by a trusted aggregation center. However, Privacy by Design paradigm can also be applied with success to distributed systems. In this section, we discuss an instance of this case [26]; in particular, we analyze the handling of personal mobility trajectories, generated by several vehicles distributed in a territory and collected by a central entity, called *coordinator*. Streams of data updates arrive continuously at remote sites (i.e., vehicles), while the coordinator is responsible for computing the aggregation of movement data on a territory by combining the information received by each node.

We show how privacy can be obtained before data leave users, ensuring the utility of some data analysis performed at the collective level, also after the transformation. This example brings evidence to the fact that the Privacy by Design model has the potential of delivering high data protection combined with high quality even in massively distributed techno-social systems.

### *10.5.1 Attack and Privacy Model*

As in the case analyzed in Sect. 10.3, any data from which the typical mobility behavior of a user may be inferred is assumed as sensitive information. This information is considered sensitive for two main reasons: (1) typical movements can be used to identify drivers even when a simple de-identification of the individual in the system is applied; and (2) the places visited could identify distinguishing sensitive areas such as clinics, hospitals and routine locations such as the user's home and workplace.

The assumption is that each node in the system is honest; in other words, attacks at the node level are not considered. Instead, potential attacks are from any adversary between the node and the coordinator (i.e., attacks during the communications), and from any adversary at coordinator site, so this privacy preserving technique has to guarantee privacy even against a malicious behavior of the coordinator. For example,

the coordinator may be able to obtain real mobility statistic information from other sources, such as from public datasets on the web, or through personal knowledge about a specific individual, like in the previously (and diffusely) discussed linking attack.

The solution proposed in [26] is based on *Differential Privacy*, a recent paradigm of randomization presented in [14] by Dwork. The general idea of this model is that the privacy risks should not increase for a respondent as a result of occurring in a statistical database; differential privacy ensures, in fact, that the ability of an adversary to inflict harm should be essentially the same, independently of whether any individual opts in to, or opts out of, the dataset. This privacy model is called $\epsilon$-differential privacy, due to the level of privacy guaranteed $\epsilon$, also called *privacy budget*. Note that when $\epsilon$ grows very little perturbation is introduced and this yields a low privacy protection; on the contrary, better privacy guarantees are obtained when $\epsilon$ tends to zero. Differential privacy guarantees a record owner that any privacy breach will not be a result of participating in the database since nothing, or almost nothing, that can be learned from the database with his record can be learned from the database without his data. Moreover, in [14] is formally proved that $\epsilon$-differential privacy can provide a guarantee against adversaries with arbitrary background knowledge, thus, in this case, we do not need to define any explicit background knowledge for attackers.

In a nutshell, the differential privacy mechanism works by adding appropriately chosen random noise (from a specific distribution) to the true answer, then returning the perturbed answer. The formal definition of differential privacy [14] is the following. Here the parameter, $\epsilon$, specifies the level of guaranteed privacy.

**Definition 10.8 ($\epsilon$-Differential Privacy)** [14] A privacy mechanism $A$ gives $\epsilon$-differential privacy if for any dataset $D_1$ and $D_2$ differing on at most one record, and for any possible output $D'$ of $A$ we have $Pr[A(D_1) = D'] \leq e^{\epsilon} \times Pr[A(D_2) = D']$ where the probability is taken over the randomness of $A$.

A basic notion used by differential privacy mechanisms is the *sensitivity* of a query, which provides a way to set the noise distribution in order to calibrate the noise magnitude on the basis of the type of query.

**Definition 10.9 (Global Sensitivity)** [13] For any function $f : D \rightarrow R^d$, its sensitivity is $\Delta f = max_{D_1,D_2} ||f(D_1) - f(D_2)||_1$ for all $D_1$, $D_2$ differing in at most one record.

Intuitively, the sensitivity measures the maximum distance between the same query executed on two close datasets, i.e., datasets differing on one single element (either a user or an event). As an example, consider a count query on a medical dataset, which returns the number of patients having a particular disease. The result of the query performed on two close datasets, i.e., differing exactly on one patient, can change at most by 1; thus, in this case (or, more generally, in count query cases), the sensitivity is 1.

A little variant of this model is the $(\epsilon, \delta)$-differential privacy [16], where the noise is bounded at the cost of introducing a privacy loss, $\delta$. $(\epsilon, \delta)$-differential privacy

allows a small amount of privacy loss due to a variation in the output distribution for the privacy mechanism $A$.

**Definition 10.10 ($(\epsilon, \delta)$-Differential Privacy)** [16] A privacy mechanism $A$ gives $(\epsilon, \delta)$-differential privacy if for any dataset $D_1$ and $D_2$ differing on at most one record, and for any possible output $D'$ of $A$ we have $Pr[A(D_1) = D'] \leq e^{\epsilon} \times Pr[A(D_2) = D'] + \delta$ where the probability is taken over the randomness of $A$.

The questions are: *How can we hide the event that the user moved from a location* a *to a location* b *in a time interval $\tau$? And how can we hide the real count of moves in that time window?* In other words, *How can we enable collective movement data aggregation for mobility analysis while guaranteeing individual privacy protection?* The solution that we report is based on ($\epsilon$,$\delta$)-differential privacy, and provides a good balance between privacy and data utility.

### *10.5.2 Privacy-Preserving Technique*

First of all, each participant must share a common partitioning of the considered area; for this purpose, it is possible to use an existing division of the territory (e.g., census sectors, road segments, etc.) or to determine a data-driven partition as the Voronoi tessellation introduced in Sect. 10.3.2. Once this is accomplished, each trajectory is generalized as a sequence of crossed areas (i.e., a sequence of movements). For the sake of convenience, this information is mapped onto a *frequency vector*, linked to the partition.

In order to perform this mapping task, we firstly need a function *Move Frequency* ($MF$) to compute how many times the move appears in a generalized trajectory $T_g$ within a given time interval.

**Definition 10.11 (Move Frequency)** Let $T_g$ be a generalized trajectory and let $(l_{c_i}, l_{c_j})$ be a move. Given the temporal interval $\tau$ the move frequency function is defined as:

$$MF(T_g, (l_{c_i}, l_{c_j}), \tau) = |\{(l_{c_i}, l_{c_j}, t_i, t_j) \in T_g : t_i \in \tau \wedge t_j \in \tau\}|.$$

This function can be easily extended to take into consideration a set of generalized trajectories $\mathscr{T}^{\mathscr{G}}$. In this case, computed information represents the total number of movements from the cell $c_i$ to the cell $c_j$ in a time interval in the set of trajectories.

**Definition 10.12 (Global Move Frequency)** Let $\mathscr{T}^{\mathscr{G}}$ be a set of generalized trajectories and let $(l_{c_i}, l_{c_j})$ be a move. Let $\tau$ be a time interval. The global move frequency function is defined as:

$$GMF(\mathscr{T}^{\mathscr{G}}, (l_{c_i}, l_{c_j}), \tau) = \sum_{\forall T_g \in \mathscr{T}^{\mathscr{G}}} MF(T_g, (l_{c_i}, l_{c_j}), \tau).$$

The number of movements between two cells computed by either the function $MF$ or $GMF$ describes the amount of traffic flow between the two cells in a specific time interval $\tau$. This information can be represented by a frequency vector. To define the frequency vector, we first define *vector of moves*.

**Definition 10.13 (Vector of Moves)** Let $C = \{c_1, c_2, \ldots, c_p\}$ be the set of the cells composing the territory partition. The vector of moves $M$ is a vector of size $s = |\{(c_i, c_j)|c_i \ is \ adjacent \ to \ c_j\}|$, in which each element $M[k] = (l_{c_i}, l_{c_j})$, where $1 \leq k \leq s$, is the move from the cell $c_i$ to the adjacent cell $c_j$.

At this point, we can define the frequency vector.

**Definition 10.14 (Frequency Vector)** Let $C = \{c_1, c_2, \ldots, c_p\}$ be the cells that compose the territory partition and let $M$ be its vector of moves. Given a set of generalized trajectories $\mathscr{G}$ in a time interval $\tau$, its *frequency vector* $f$ is a vector of size $s = |\{(c_i, c_j)|c_i \ is \ adjacent \ to \ c_j\}|$, in which each element $f[k] = GMF(\mathscr{G}, M[k], \tau)$.

Unfortunately, releasing frequency of moves instead of raw trajectory data to the coordinator is still not privacy-preserving, as the intruder may still infer the sensitive typical movement information of the driver. As an example, the attacker could discover the driver's most frequent move; this information can be very sensitive because it usually corresponds to a user's transportation between home and workplace. Thus, the proposed solution is based on the differential privacy model, relying on a Laplace distribution [15]. At the end of a preset time interval $\tau$, each node, before sending the frequency vector to the coordinator and for each element in the vector, extracts the noise from the Laplace distribution and adds it to the original value in that position of the vector. At the end of this operation, the node $V_j$ converted its frequency vector $f_{V_j}$ into its private version $\tilde{f_{V_j}}$. This ensures the respect of the $\epsilon$-differential privacy. This simple general strategy has some inconveniences: first, it could lead to a large amount of noise that, although with small probability, can be arbitrarily large; second, adding noise drawn from the Laplace distribution could produce negative frequency counts of moves, which does not make sense in mobility scenarios. In order to fix these two problems, it is possible to bound the noise drawn from the Laplace distribution, reducing to an ($\epsilon$,$\delta$) differential privacy schema. In particular, for each value $x$ of the vector $f_{V_j}$, it is possible to draw the noise bounding it in the interval $[-x, x]$. In other words, for any original frequency $f_{V_j}[i] = x$, its perturbed version after adding noise falls in the interval $[0, 2x]$. This approach satisfies ($\epsilon$,$\delta$)-differential privacy, where $\delta$ measures the privacy loss. Note that, since in a distributed environment communications need to be quite limited, it is possible to reduce the amount of transmitted information, i.e., the size of frequency vectors. A possible solution to this problem is reported in [26], but this discussion is omitted here because is beyond the purpose of our review.

### *10.5.3 Analytical Quality*

So far we reported the formal guarantees to individual privacy preservation, now we have to show how individually transformed values are still useful once they are collected and aggregated by the coordinator, i.e., they are still suitable at a collective level for analysis. In the proposed framework, the coordinator gathers the perturbed frequency vectors from all the vehicles in the time interval $\tau$ and sums them movement by movement. This achieves to obtain the resulting global frequency vector, which indicates the flow values for each possible link of the spatial tessellation. Since the privacy transformation operates on the entries of the frequency vectors, and hence on the flows, we report the comparison (before and after the transformation) of two measures: (1) the *Flow per Link*, i.e., the directed volume of traffic between two adjacent zones; (2) the *Flow per Zone*, i.e., the sum of the incoming and outgoing flows in a zone. The following results refer to the application of this technique on a large dataset of GPS vehicles traces, collected in a period from 1st May to 31st May 2011, in the geographical areas around Pisa, in central Italy. It counts for around 4200 vehicles, generating around 15,700 trips in total. The $\tau$ interval is 1 day, so the global frequency vector represents the sum all the trajectories crossing any link, at the end of each day. The reported results are relative to 25th May 2011, but they are similar to ones obtained in the other working days.

Figure 10.7 shows the resulting Complementary Cumulative Distribution Functions (CCDFs) of different privacy transformation varying $\epsilon$ from 0.9 to 0.01. Figure 10.7-Left shows the global (approximated) *Flow per Link* distribution: fixed a value of flow (x) is counted the number of links (y) that have that flow. Figure 10.7-Right shows the distribution of sum of flows passing for each zone, i.e., *Flow per Zone*: given a flow value (x) it shows how many zones (y) present that total flow. From the distributions, we can observe that the privacy transformation preserves

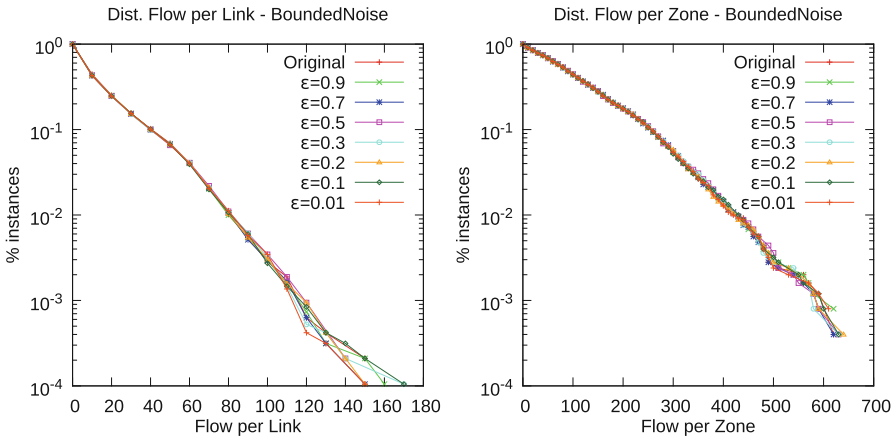


**Fig. 10.7** CCDFs of *Flow per Link* (Left); CCDFs of *Flow per Zone* (Right)

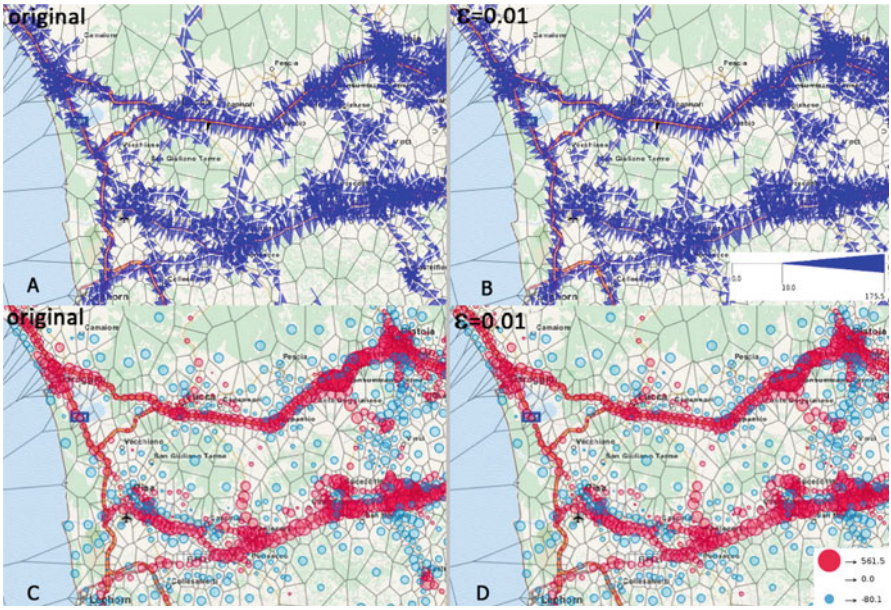


**Fig. 10.8** Visualization of *Flow per Link* (A-B) and *Flow per Zone* (C-D)

very well the distribution of the original flows, even for more restrictive values of the parameter $\epsilon$. Also considering several flows together, like those ones that are incident to a given zone (Fig. 10.7-Right), the distributions are well preserved for all the privacy transformations. These results reveal how a method which *locally* perturbs values, at a *collective* level permits to obtain a very good utility.

Qualitatively, Fig. 10.8 shows a visual comparison of results of the privacy transformation with the original ones. This is an example of analyses that can be carried out with mobility data. Since the global complementary cumulative distribution functions are comparable, it is possible to choose a very low epsilon ($\epsilon = 0.01$) with the aim to emphasize the very good quality of mobility analysis that an analyst can obtain even if the data are transformed by using this low $\epsilon$ value, i.e. obtaining a better privacy protection. In Fig. 10.8a, b each flow is drawn with arrows with the thickness proportional to the volume of trajectories observed on a link. From the figure it is evident how the relevant flows are maintained in the transformed global frequency vector, highlighting the major highways and urban centers. The *Flow per Zone* is also preserved, as it is shown in Fig. 10.8c, d, where the flow per each cell is rendered with a circle of radius proportional to the difference from the median value of each global frequency vector. The maps allow us to recognize the dense areas (red circles, above the median) separated by sparse areas (blue circles, below the median). The high density traffic zones follow the highways and the major city centers along their routes. These two comparisons confirm the intuition that, while the transformations protect individual sensitive information, the utility of data is preserved.

## 10.6 Conclusion

The potential impact of the big data analytics and social mining is high because it could generate enormous value to society. Unfortunately, often big data describe sensitive human activities and the privacy of people is always more at risk. The danger is also increasing thanks to the emerging capability to integrate diversified data. In this chapter, we have introduced the articulation of the Privacy by Design in big data analytics and social mining for enabling the design of analytical processes that minimize, or even prevent, the privacy harm. We have discussed how applying the Privacy by Design principle to three different scenarios showing that under suitable conditions is feasible to reach a good trade-off between data privacy and good quality of the data. We believe with the Privacy by Design principle social mining has the potential to provide a privacy-respectful social microscope, or socioscope, needed to observe the hidden mechanisms of socio-economic complexity.

## 10.7 Bibliographic Notes

In the following, we provide a quick overview of some techniques and solutions adopted in privacy-preserving data mining for mobility data. The Privacy by Design model was applied in data mining in several contexts [24, 27], with special treatment to mobility data, due to their complex nature, their sensitivity and their importance for understanding human behaviors. Privacy issue in mobility data mining and sharing have been intensively studied in literature [8, 20, 22], and the existing methods of privacy-aware releasing and sharing of (trajectory) data can be classified into two main classes: (1) generalization/suppression based data perturbation, and (2) randomization/differential privacy perturbation.

The most widely used privacy model for generalization and suppression perturbation is adapted from what so called $k$-anonymity [30, 32], which requires that an individual should not be identifiable from a group of size smaller than $k$ based on their quasi-identifies (QIDs), i.e., a set of attributes that can be used to identify uniquely the individuals. Unfortunately, in trajectory data, it often impossible to distinguish clearly between quasi-identifiers and sensitive attribute. In [36], Yarovoy et al. deeply analyze the problem of quasi-identifiers in mobility data: they show that the anonymization groups may not be disjoint. Thus there may exist objects that can be identified explicitly by combining different anonymization groups. They suggest that QIs may be provided directly by personal settings or found by means of statistical data analysis. In [4], Abul et al. propose the notion of $(k,\delta)$-anonymity for moving object databases, where $\delta$ represents the possible location imprecision. This is an innovative concept of $k$-anonymity based on co-localization, which takes advantage of the inherent uncertainty of the whereabouts of the moving objects. The authors also proposed an approach, called Never Walk Alone, based on trajectory

clustering and spatial translation, and they present its improvement, Wait for Me, in [5]. This method is very similar to the previous one, but it is based on EDR distance (instead of Euclidean distance), which is time-tolerant, so Wait for Me can recognize similar trajectories even if they are (slightly) shifted in time. Finally, Domingo Ferrer and Trujillo-Rasua [12] show a solution based on perturbation and micro-aggregation: this method $k$-anonymizes each location independently, using the whole set of trajectories. Particularly, the algorithm creates clusters of locations (close in time and space) in such a way that the locations in each group belong to k different trajectories. The result of this transformation is that the probability that a location of a true trajectory appears in its anonymized version is at most $\frac{1}{k}$ while guaranteeing that the anonymized trajectories are suitable for range query for every value of $k$.

Regarding the application of Differential Privacy mechanisms to mobility data, many works have been proposed in last years. In [35] authors provide an algorithm, based on Markov Chain and Differential Privacy, which aims to protect the continual location sharing of perturbed locations in the context of Location Based Services. In particular, they select a set of locations that are highly probable for a user, guaranteeing that the probability of these locations is similar to the other, and chooses one of these locations to be released outside. In this case, the event protected by Differential Privacy is a specific request to a service, instead of a specific move. However, they do not provide guarantees if the attacker has a stronger external knowledge w.r.t. the history of the released locations. This additional constraint is analyzed in [7], where Andrés et al. show a technique for Location Based Services independent from the side information of users. They use an extension of Laplace distribution for the continuous plane and promise a privacy level which is distance-dependent, i.e., guarantees are stronger if you get closer to the real location of the user. A very promising research line about Differential Privacy on spatio-temporal data is the one related to space partitioning. Ho and Ruan [21] apply Differential Privacy to interesting locations to perform location pattern discovery, granting protection at the user-level also when a user contributes to more than one record. They partition the space of the data into smaller ones, in order to limit the total number of events and, consequently, the events connected with each individual in each dataset, in order to overcome the problem of the presence of a clear upper-bound to the events related to a single user. In [10], Cormode et al. describe a solution to publish differentially private spatial index (e.g., quadtrees and kd-trees) to provide a private description of the data distribution [10]. Its main utility concern is the accuracy of multi-dimensional range queries (e.g., how many individuals fall within a given region). Therefore, the spatial index only stores the counts of a specific spatial decomposition, even their solution does not store the movement information (e.g., how many individuals move from location $i$ to location $j$). In [9], authors rely on a prefix tree of trajectories with injected Laplace noise; the prefix tree is data-dependent, i.e., it should have a different structure when the underlying database changes. Qardaji et al. [29] provide an adaptive uniform partition method, considering different density-regions, i.e., depending on the total number of points in the dataset. In Acs et al. [6], authors apply Geometrical mechanism to a partition

of a territory, taking advantage of a Voronoi tessellation to keep track of the presence of individuals and use clustering and sampling with Fourier-based perturbation Finally, Cormode et al. [11] propose to publish a contingency table of trajectory data, that can be indexed by specific locations so that each cell in the table contains the number of people who commute from the given source to the given destination. The purpose of this work is to address the sparsity issue of the contingency table and presents a method of releasing a compact summary of the contingency table with Laplace noise.

## References


1. *Privacy by Design Resolution*, International Conference of Data Protection and Privacy Commissioners (Jerusalem, Israel, October 27–29, 2010)
2. *Article 29 Data Protection Working Party and Working Party on Police and Justice, The Future of Privacy: Joint Contribution to the Consultation of the European Commission on the Legal Framework for the Fundamental Right to Protection of Personal Data*, 02356/09/EN, WP 168 (Dec. 1, 2009)
3. *Directive (EU) 2016/680 of the European Parliament and of the Council of 27 April 2016*, Official Journal of the European Union (2016)
4. O. Abul and F. Bonchi and M. Nanni, in *Never Walk Alone: Uncertainty for Anonymity in Moving Objects Databases*, ICDE 2008, pp. 376–385
5. O. Abul and F. Bonchi and M. Nanni, in *Anonymization of Moving Objects Databases by Clustering and Perturbation*, Inf. Syst., vol 35, num 8, pp. 884–910 (Elsevier Science Ltd., December, 2010), doi: 10.1016/j.is.2010.05.003
6. G. Ács and C. Castelluccia, *A case study: privacy preserving release of spatio-temporal density in Paris*, KDD 2014, pp. 1679–1688
7. M. E. Andrés and N.E. Bordenabe and K. Chatzikokolakis and C. Palamidessi, *Geo-indistinguishability: differential privacy for location-based systems*, ACM Conference on Computer and Communications Security, p. 901–914, 2013
8. F. Bonchi and L.V.S. Lakshmanan and H. W. Wang, in *Trajectory anonymity in publishing personal mobility data*, SIGKDD Explor. Newsl. 2011, pp. 30–42, https://doi.org/10.1145/2031331.2031336
9. R. Chen and B.C.M. Fung and B.C. Desai and N.M. Sossou, *Differentially private transit data publication: a case study on the Montreal transportation system*, KDD 2012, pp. 213–221
10. G. Cormode and C. Procopiuc and D. Srivastava and E. Shen and T. Yu, *Differentially private spatial decompositions*, ICDE 2012, pp. 20–31
11. G. Cormode and C. Procopiuc and D. Srivastava and T. Tran, *Differentially private summaries for sparse data*, ICDT 2012, pp. 299–311
12. J. Domingo-Ferrer and R. Trujillo-Rasua, in *Microaggregation- and permutation-based anonymization of movement data*, Inf. Sci. 2012, volume 208 pp. 55–80
13. C. Dwork, in *Differential privacy: A survey of results*, International Conference on Theory and Applications of Models of Computation, pages 1–19. Springer, 2008
14. C. Dwork, in *Differential Privacy*, ICALP 2006, Lecture Notes in Computer Science, vol 4052, https://doi.org/10.1007/11787006_1
15. C. Dwork and F. Mcsherry and K. Nissim and A. Smith, in *Calibrating noise to sensitivity in private data analysis*, Proceedings of the 3rd Theory of Cryptography Conference (Springer, 2006), pp. 265–284

16. C. Dwork, K. Kenthapadi, F. McSherry, I. Mironov, M. Naor, in *Our Data, Ourselves: Privacy Via Distributed Noise Generation*, Advances in Cryptology-EUROCRYPT 2006, pp. 486–503. Springer Berlin Heidelberg, 2006
17. European Data Protection Supervisor in *Opinion of the European Data Protection Supervisor on Promoting Trust in the Information Society by Fostering Data Protection and Privacy* (Mar. 18, 2010)
18. European Parliament & Council. General data protection regulation, 2016. L119, 4/5/2016
19. Federal Trade Commission (Bureau of Consumer Protection, in *Preliminary Staff Report, Protecting Consumer Privacy in an Era of Rapid Change: A Proposed Framework for Business and Policy Makers* (Dec. 2010)
20. F. Giannotti and D. Pedreschi, in *Mobility, Data Mining and Privacy: A Vision of Convergence*, Mobility, Data Mining and Privacy - Geographic Knowledge Discovery (2008), https://doi.org/10.1007/978-3-540-75177-9_1
21. S. Ho and S. Ruan, *Preserving Privacy for Interesting Location Pattern Mining from Trajectory Data*, Transactions Data Privacy (2013) 6(1): 87–106, 2013
22. A. Monreale and D. Pedreschi and R. G. Pensa, in *Anonymity Technologies for Privacy-Preserving Data Publishing and Mining*, Privacy-Aware Knowledge Discovery: Novel Applications and New Techniques (2010), pp. 3–33
23. A. Monreale and G. L. Andrienko and N. V. Andrienko and F. Giannotti and D. Pedreschi and S. Rinzivillo and S. Wrobel, in *Movement Data Anonymity through Generalization*, Transactions on Data Privacy (2010), volume 3 number 2, pp. 91–121
24. A. Monreale, in *Privacy by Design in Data Mining*, PhD Thesis, Dept. of Computer Science (University of Pisa, 2011)
25. A. Monreale and R. Trasarti and D. Pedreschi and C. Renso and V. Bogorny in *C-safety: a framework for the anonymization of semantic trajectories*, Transactions on Data Privacy (2011), volume 4 number 2 pp. 73–101
26. A. Monreale and W. Hui Wang and F. Pratesi and S. Rinzivillo and D. Pedreschi and G. Andrienko and N. Andrienko, in *Privacy-preserving Distributed Movement Data Aggregation*, AGILE (Springer 2013), https://doi.org/10.1007/978-3-319-00615-4_13
27. A. Monreale and S. Rinzivillo and F. Pratesi and F. Giannotti and D. Pedreschi, in *Privacy-by-design in big data analytics and social mining*, EPJ Data Science (2014)
28. R. G. Pensa and A. Monreale and F. Pinelli and D. Pedreschi, in *Pattern-Preserving k-Anonymization of Sequences and its Application to Mobility Data Mining* PiLBA 2008
29. W. H. Qardaji and W. Yang and N. Li, in *Differentially private grids for geospatial data*, ICDE 2013, pp 757–768
30. P. Samarati and L. Sweeney, in *Protecting privacy when disclosing information: k-anonymity and its enforcement through generalization and suppression* (SRI International, 1998)
31. 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).
32. L. Sweeney, in *Computational disclosure control: a primer*, Ph.D. thesis, Dept. of Electrical Eng. and Computer Science (MIT, 2001)
33. L. Sweeney, in *Simple Demographics Often Identify People Uniquely*, Carnegie Mellon University, Data Privacy Working Paper 3. Pittsburgh 2000
34. W.K. Wong and D.W. Cheung and E. Hung and B. Kao and Nikos Mamoulis, in *Security in Outsourcing of Association Rule Mining*, VLDB 2007, pp. 111–122
35. Y. Xiao and L. Xiong, in *Protecting Locations with Differential Privacy under Temporal Correlations*, ACM Conference on Computer and Communications Security, p.298–1309, 2015
36. R. Yarovoy and F. Bonchi and L.V. S. Lakshmanan and W. H. Wang, in *Anonymizing moving objects: how to hide a mob in a crowd?*, International Conference on *Extending DataBase Technology* (2009), pp. 72–83