Monreale et al. *EPJ Data Science* 2014, **2014**:10
http://www.epjdatascience.com/content/2014/1/10

EPJ Data Science
a SpringerOpen Journal

REGULAR ARTICLE Open Access

Privacy-by-design in big data analytics and social mining

Anna Monreale[1,2*], Salvatore Rinzivillo[2], Francesca Pratesi[1,2], Fosca Giannotti[2] and Dino Pedreschi[1]

*Correspondence:
annam@di.unipi.it
[1]Department of Computer Science, University of Pisa, Largo Pontecorvo, 3, Pisa, Italy
[2]ISTI-CNR, Via G. Moruzzi, 1, Pisa, Italy

Abstract

Privacy is 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. 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. As a result, privacy preservation simply cannot be accomplished by de-identification alone. In this paper, 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.

Keywords: privacy-by-design; bid data analytics; social mining

1 Introduction

The big data originating from the digital breadcrumbs of human activities, sensed as a by-product of the ICT systems that we use everyday, record the multiple dimensions 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 networking services record our connections to friends, colleagues and collaborators; wireless networks and mobile devices record the traces of our movements. These kinds of big data describing human activities are at the heart of the idea of a 'knowledge society', where the understanding of social phenomena is sustained by the knowledge extracted from the miners of big data 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 [1–7], economic and financial crises, the spread of epidemics [8–11], the diffusion of opinions [12] and so on.

The worrying side of this story is that this big data contain personal sensitive information, so that the opportunities of discovering knowledge increase with the risks of privacy violation. When personal sensitive data are published and/or analyzed, one important question to consider is whether this may violate the privacy right of individuals. The human data may potentially reveal many facets of the private life of a person: but a higher level of danger is reached if the various forms of data can be linked together. It is evident

that maintaining control on personal data guaranteeing privacy protection is increasingly difficult and it cannot simply be accomplished by de-identification [13] (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 and in the media, from health records to querylogs to GPS trajectories.

In the past few years, several techniques have been proposed to develop technological frameworks for countering privacy violations, without losing the benefits of big data analytics technology [14–18]. Despite these efforts, no general method exists that is capable of handling both generic personal data and preserving generic analytical results. Anonymity in generic sense is considered a chimera and the concern about intrusion in the private sphere by means of big data is now in news headlines of major media. Nevertheless, big data analytics and privacy are not necessary enemies. The purpose of this paper is precisely to show that many practical and impactful services based on big data analytics can be designed in such a way that the quality of results can coexist with high protection of personal data. The magic word is privacy-by-design. We propose here a methodology for purpose-driven privacy protection, where the purpose is a target knowledge service to be deployed on top of data analysis. The basic observation is that providing a 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., in reference to the kind of the analytical goal desired and the reasonable level of privacy expected.

In this paper we elaborate on the above ideas and instantiate the privacy-by-design paradigm, introduced by Anne Cavoukian [19], in the 1990s, to the designing of big data analytical services. First, we discuss the privacy-by-design principle highlighting how it has been embraced by United States and Europe. Then, we introduce our idea of privacy-by-design in big data analytics domain and show how inscribing privacy 'by design' in four different specific scenarios assuring a good balance between privacy protection and quality of data analysis. To this end, we review a method for a privacy-aware publication of movement data enabling clustering analysis useful for understanding human mobility behavior in specific urban areas [20], a method for a privacy-aware outsourcing of the pattern mining task [21], and a method for a privacy-aware distributed mobility data analytics [22], enabling any company without suitable resources to take advantage from data mining technologies. Finally, we analyze the privacy issues of the socio-meter of urban population presented in [23] and propose a privacy-by-design schema that allows a privacy-aware estimation of the proportion of city users that fall into three categories: residents, commuters, visitors. Especially in this last example, we can see how sometimes it is sufficient to use a bit of smartness in order to have good quality results without compromising individual privacy.

The remaining of the paper is organized as following. In Section 2 we discuss the privacy-by-design paradigm and its articulation in data analytics. Section 3 and Section 4 discuss the application of the privacy-by-design principle in the case of publication of personal mobility trajectories and outsourcing of mining tasks, respectively. In Section 5 we show a possible distributed scenario for privacy preserving mobility analytics, while in Section 6 we present a study of privacy issues of a socio-meter of urban population and propose a schema for guaranteeing user privacy protection. Lastly, Section 7 concludes the paper.

2 Privacy-by-design

Privacy-by-design is a paradigm developed by Ontario's Information and Privacy Commissioner, Dr. Ann Cavoukian, in the 1990s, to address the emerging and growing threats to online privacy. The main idea is to inscribe the privacy protection into the design of information technologies from the very start. This paradigm represents a significant innovation with respect to the traditional approaches of privacy protection because it requires a significant shift from a reactive model to proactive one. In other words, the idea is preventing privacy issues instead of remedying to them.

Given the ever growing diffusion and availability of big data and given the great impact of the big data analytics on both human privacy risks and the possibility of understanding important phenomena many companies are realizing 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.

2.1 Privacy-by-design in law

The privacy-by-design model has been embraced in Europe and in the United States.

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* [24] and encourages 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, recommending the incorporation of the principles of privacy-by-design into a new EU privacy framework [25]. In March 2010, the European Data Protection Supervisor suggested to 'include unequivocally and explicitly the principle of privacy-by-design into the existing data protection regulatory framework' [26]. This recommendation was taken into consideration in the recent revision of the Data Protection Directive (95/46/EC) currently under discussion at EC. The European Union Data Protection Directive has always included provisions requiring data controllers to implement *technical and organizational measures* in the design and operation of ICT; but this has proven insufficient. Therefore, in the comprehensive reform of the data protection rules proposed on January 25, 2012, the new data protection legal framework introduces, with respect to the Directive 95/46/EC, the reference to *data protection by design and by default* (Article 23 of the Proposal for a Regulation). This article compels the controller to '*implement appropriate technical and organizational measures and procedures in such a way that the processing will meet the requirements of this Regulation and ensure the protection of the rights of the data subject*' and to '*implement mechanisms for ensuring that, by default, only those personal data are processed which are necessary for each specific purpose of the processing…*'.

Privacy-by-design has been embraced also in the United States. In the last years the U.S. Federal Trade Commission hosted a series of public roundtable discussions on privacy issues in the digital age and in a recent staff report [27] it describes a proposed framework with three main recommendations: *privacy-by-design*, *simplified consumer choice*, and *increased transparency of data practices*. Moreover, some pieces of legislation have also been proposed and introduced which include the principles of privacy-by-design, including:

(a) 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, if passed, would require companies that collect, use, store or transfer consumer information to implement a version of privacy-by-design when developing products; (b) the Franken/Blumenthal Location Privacy Protection Act, introduced in 2012, that regulates the transmission and sharing of user location data in USA; and (b) the Wyden/Chaffetz Geolocation and Privacy Surveillance, introduced in 2011, that attempted to limit government surveillance using geolocation information such as signals from mobile phones and GPS devices.

2.2 Privacy-by-design in big data analytics and social mining

As stated above, in many contexts it is not 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 main idea is to inscribe 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 articulation of the general 'by design' principle in the big data analytics domain is that higher protection and quality can be better achieved in a goal-oriented approach. In such an approach, 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 adversary 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 fundamental for the design of a privacy-aware technology. First of all, the techniques for privacy preservation strongly depend on the nature of the data to be protected. For example, methods suitable for social networking data could not be appropriate for trajectory data.

Second, a valid framework has to define the attack model, that could be an honest-but-curious adversary model or a malicious adversary model, and an adequate countermeasure. The two models require different actions due to their characteristics. The first one executes protocols correctly but tries to learn as much as possible. For example, by reading off-line the standard output of the algorithm he can try to deduce information on the other party. This is different from the malicious adversary who since could also deviate arbitrarily from the protocol is harder to be countered. Typically an attack is based on a specific adversary's background knowledge and different assumptions on the background knowledge entail different defense strategies. For example, an attacker could possess an approximated information about the mobility behavior of a person and use it to infer all his movements. In other cases, the adversary could shadow a person and discover some specific places visited by him obtaining an exact information. It is clear that a defense strategy designed for counter attacks with approximate knowledge could be too weak in case of detailed knowledge and vice versa.

Finally, a privacy-aware strategy should find an acceptable trade-off between data privacy and data utility. To this end, it is fundamental to consider the category of analytical queries to be answered for understanding which data properties is necessary to preserve. As an example, the design of a defense strategy for movement data should consider that this data could be used for analyzing collective mobility behavior in a urban area.

Under the above assumptions, we claim that it is conceivable to design a privacy-aware analytical process that can:

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 trade-off between privacy protection and data quality must be the main goal in the design of a privacy-aware technology for big data analytics. If in the designing of a such framework only one of these two aspects is taken into consideration, then the consequence is that either we assure high levels of privacy but the data cannot be used for analytical scopes, or we assure a very good quality of data by putting at risk the individual privacy protection of people in the data. Note that, in big data analytics and social mining typically one is interested into extract collective knowledge and this could not involve the use of personally identifiable information. However, when it does, the *data minimization* principle should be taken into account, since it allows managing data privacy risks, by effectively eliminating risk at the earliest stage of the information life cycle. This principle requires that in the design of big data analytical frameworks we should consider that we need no collection of personally identifiable information, unless a specific purpose is defined. The above privacy-by-design methodology (Point c) can help to understand which is the minimal information that enables a good analysis and protection. As we can see in the scenario presented in Section 6, we are able to find the minimal information for mining data with perfect quality and, we show how the level of data aggregation useful for the analysis already provides very low privacy risks.

In the following, we show how we apply the privacy-by-design paradigm for the design of four analytical frameworks: one for the publication of trajectory data; one for the outsourcing of data mining tasks; one for computing aggregation of movement data in a distributed fashion and one for the quantification of user profiles in GSM data. In the four 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 [14–18]).

3 Privacy-by-design in mobility data publishing

In this section, we discuss 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) [20]. It is suitable for the privacy-aware publication of movement data enabling clustering analysis useful for the understanding of human mobility behavior in specific urban areas. The released trajectories are make anonymous by a suitable process that realizes a generalized version of the original trajectories.

The framework is based on a data-driven spatial generalization of the dataset of trajectories. The results obtained with the application of this framework show how trajectories can be anonymized to a high level of protection against re-identification while preserving the possibility of mining clusters of trajectories, which enables novel powerful analytic

services for info-mobility or location-based services. We highlight that the mobility data published after the privacy transformation strategy, described in the following, is suitable for collective data analyses useful for extracting knowledge describing the collective mobility behavior of a population. Clearly, in cases where for providing a service it is necessary to identify specific, personal trajectories related to a specific user, this framework is not adequate. This because in that context one of the most important aspects is to maintain clear and well-defined the information at *individual* level, which is what we want to obfuscate with our transformation. In other words, the goal here is to *enable collective analytical tool while protecting the individual privacy*.

3.1 State-of-the-art on privacy-preserving mobility data publishing

There have been some works on privacy-preserving publishing of spatio-temporal moving points by using the generalization/suppression techniques. The mostly widely used privacy model of these works is adapted from what so called k-anonymity [28, 29], 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 uniquely identify the individuals. [14] proposes the (k, δ)-anonymity model that exploits the inherent uncertainty of the moving object's whereabouts, where δ represents possible location imprecision. Terrovitis and Mamoulis [30] assume that different adversaries own different, disjoint parts of the trajectories. Their anonymization technique is based on *suppression* of the dangerous observations from each trajectory. Yarovoy et al. [31] consider timestamps as the quasi-identifiers, and define a method based on *k-anonymity* to defend against an attack called *attack graphs*. Nergiz et al. [32] provide privacy protection by: (1) first enforcing k-anonymity, i.e. all released information refers to at least k users/trajectories, (2) randomly reconstructing a representation of the original dataset from the anonymization. Recently, [15] propose a anonymization technique based on microaggregation and perturbation. The advantage of this approach is to obtain anonymous data preserving real locations in the data and t this goal the transformation strategy uses swapping of locations.

All the above anonymization approaches are based on randomization techniques, space translations or swapping of points, and the suppression of various portions of a trajectory. To the best of our knowledge only [20] uses data-driven spatial generalization to achieve anonymity for trajectory datasets; the only work applying spatial generalization is [31], but it uses a fixed grid hierarchy to discretize the spatial dimension. In contrast, the novelty of our approach lies in finding a suitable tessellation of the geographical area into subareas dependent on the input trajectory dataset and in taking into consideration from the start also the analytical properties to be preserved in the data for guaranteeing good performance in terms of clustering analysis.

3.2 Attack and privacy model

In this framework the *linkage 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 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 [28]. The remaining attributes represent the private respondent's information, that may be violated by the linkage 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 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 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 linkage attack*, the malicious party M knows a sub-trajectory 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.

3.3 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 1(a), 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 time-stamped locations, visited by one of the tracked vehicles. Albeit de-identified, each trajectory is essentially unique - very rarely two different trajectories are exactly the same given the extremely fine spatio-temporal resolution involved. As a consequence, the chances of success for the trajectory linkage

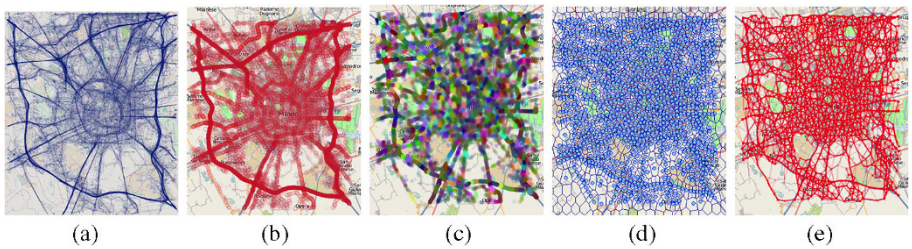

Figure 1 Trajectory Generalization process. (**a**) Original trajectories from Milan GPS dataset; (**b**) simplification of trajectory points by considering only relevant events (start, end, turns); (**c**) relevant points are clustered by spatial proximity to reveal zones with high frequency of movement events; (**d**) from the centroids of each cluster it is derived a voronoi tessellation of the territory; (**e**) original trajectories are generalized as sequence of traversed cells of the previous tessellation.

attack are not low. If the attacker M knows a sufficiently long sub-sequence S of locations visited by the respondent 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 Figure 1(a) 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 sharing common mobility behavior, such as the commuters following similar routes in their home-work and work-home trips. An 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 (Figure 1(b));
2. characteristic points are clustered into small groups by spatial proximity (Figure 1(c));
3. the central points of the groups are used to partition the space by means of Voronoi tessellation (Figure 1(d));
4. each original trajectory is transformed into the sequence of Voronoi cells that it crosses (Figure 1(e)).

As a result 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 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 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 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 appears 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 1(a), 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}.

3.4 Analytics quality

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 2(top) and Figure 2(down) illustrate the most relevant clusters found by mining the original trajectories and the anonymized trajectories, respectively.

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 make 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

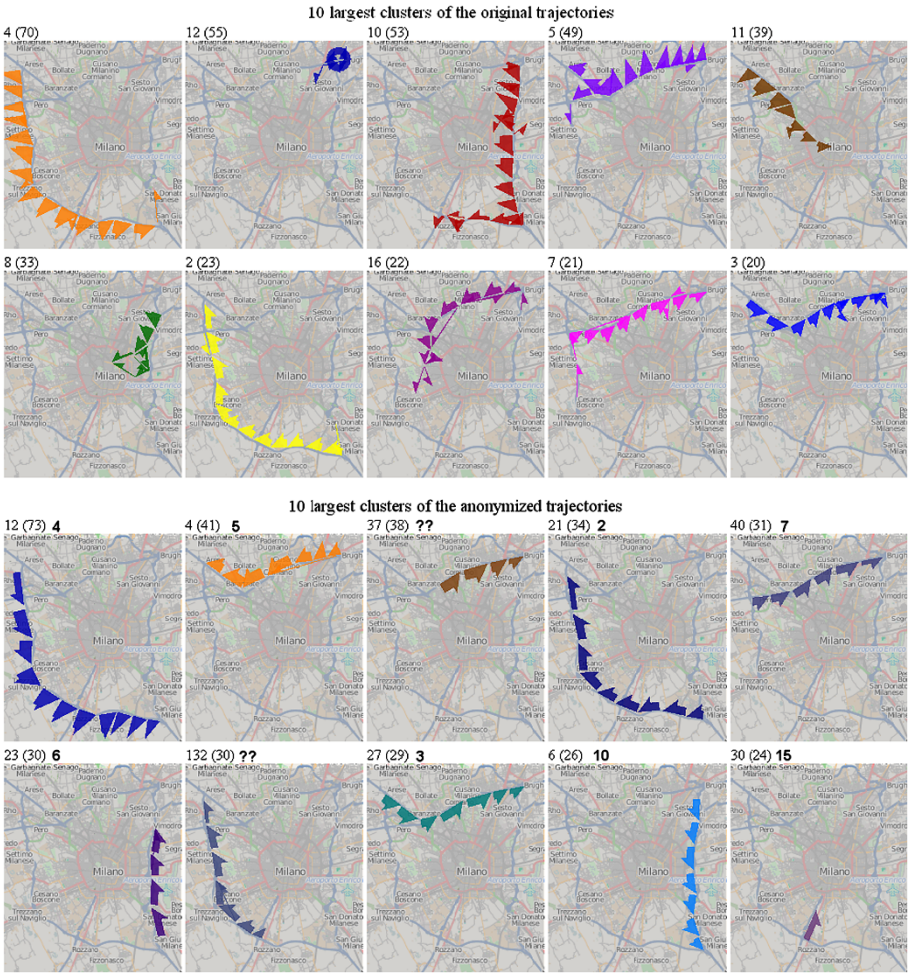

Figure 2 Comparison of the 10 largest clusters of the original trajectories (top) and of the anonymized trajectories (down). Each map summaries the trajectories within a single cluster by directed arrows with thickness proportional to flow. Bottom maps have a reference to the cluster id (in bold) of original trajectories with a similar shape.

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 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 2(top) and 2(down) is also confirmed by various cluster comparisons by Precision, Recall and F-measure, re-defined for clustering comparison (Figure 3). Here,

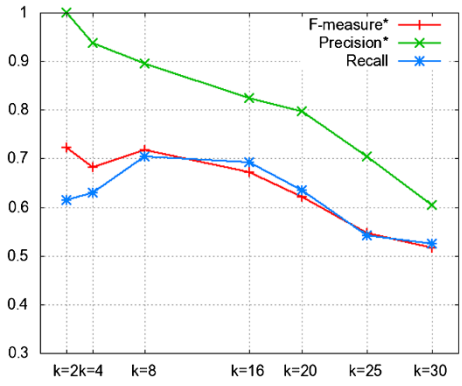

Figure 3 Quantitative comparison of clustering from original and anonymized trajectories by means of *F*-measure, Precision and Recall. When increasing k the outliers tend to be suppressed hence the precision tend to decrease. Recall on the contrary is preserved since it is influenced by denser clusters.

precision measures the percentage of objects that are preserved within the same transformed cluster; *recall* measures the percentage of objects of a transformed cluster that were into the same original cluster; F-measure is the harmonic mean of the previous measures.

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 w.r.t. the linkage attack can be measured.

4 Privacy-by-design in data mining outsourcing

In this section, we discuss an instance of the privacy by design paradigm, in the case of outsourcing of the pattern mining task [21]; in particular, the results show how a company can outsource the transaction data to a third party and obtain a data mining service in a privacy-preserving manner.

The particular problem of *outsourcing mining tasks within a privacy-preserving framework* is very interesting and is acquiring novel relevance with the advent of cloud computing and its model for IT services based on the Internet and big data centers. Business intelligence and knowledge discovery services, such as advanced analytics based on data mining technologies, are expected to be among the services amenable to be externalized on the cloud, due to their data intensive nature, as well as the complexity of data mining algorithms.

However, the key business analysis functions are unlikely to be outsourced, as they represent a strategic asset of a company: what is instead appealing for a company is to rely on external expertise and infrastructure for the data mining task, i.e., how to compute the analytical results and models which are required by the business analysts for understanding the business phenomena under observation. As an example, the operational transactional data from various stores of a supermarket chain can be shipped to a third party which provides mining services. The supermarket management need not employ an in-house team of data mining experts. Besides, they can cut down their local data management requirements because periodically data is shipped to the service provider who is in charge of maintaining it and conducting mining on it in response to requests from business analysts of the supermarket chain.

Although it is advantageous to achieve sophisticated analysis there exist several serious privacy issues of the data-mining-as-a-service paradigm. One of the main issues is that

the server has access to valuable data of the owner and may learn sensitive information from it.

A key distinction between this problem and the privacy-preserving data mining and data publishing problems is that, in this setting, not only the underlying data but also the mined results (the strategic information) are not intended for sharing and must remain private. In particular, when the third party possesses background knowledge and conducts attacks on that basis, it should not be able to learn new knowledge with a probability above a given threshold.

The frameworks devised to protect privacy in this setting have also to preserve the data utility. *What does data utility mean in this specific context?* In general, a framework for protecting the corporate privacy in data mining outsourcing must guarantee: (1) to the data owner the possibility to query its data in outsourcing (2) to the server provider to answer the queries of the data owner with an encrypted result that does not allow to infer any knowledge (3) to the data owner to recover the query results within a quantifiable approximation. The approximation of the point (3) specifies the data utility guaranteed by the privacy-preserving framework.

4.1 State-of-the-art on privacy-preserving data mining outsourcing

There have been some works on privacy-preserving data mining outsourcing. Note that, the application of a simple substitution ciphers to the items of the original database is not enough to protect privacy in this setting. Indeed, an intruder could use information about the item frequency for inferring the real identity of the items and as a consequence for breaking the whole database and the possible knowledge represented into it.

The approach in [33] is based on outsourcing a randomized dataset that is transformed by means of Bloom filters: compared with our proposal, the main weakness of this approach is that it only supports an approximate reconstruction of the mined frequent itemsets by the data owner, while our encryption/decryption method supports reconstruction of the *exact* supports.

The works that are most related to ours are [18] and [34]. They assume that the adversary possesses prior knowledge of the frequency of items or item sets, which can be used to try to re-identify the encrypted items. Wong et al. [18] consider an attack model where the attacker knows the frequency of $\alpha\%$ of frequent itemsets to within $\pm\beta\%$, while our attack model focuses on single items with the assumption that the attacker knows the exact frequency of every single item, i.e., ours is a $(100\%, 0\%)$ attack model, but confined to items. Authors in [34] assume the attacker knows exact frequency of single items, similarly to us. Both [18] and [34] use similar privacy model as ours, which requires that each real item must have the same frequency count as $k-1$ other items in the outsourced dataset. The major issue left open by [18] is a formal protection result: their privacy analysis is entirely conducted empirically on various synthetic datasets. Tai et al. [34] show that their outsourced data set satisfies k-support anonymity, but only explores set based attack empirically. Unfortunately, both works have potential privacy flaws: Molloy et al. [35] show how privacy can be breached in the framework of [18]. We have discuss the details of the flaws in the framework of [34] in [21].

4.2 Attack and privacy model

In the proposed framework, in order to achieve a strong data protection, the assumption is that an attacker wants to acquire information on the sale data and the mined patterns by

using rich background information. In particular, the attacker knows with precision the set of items in the original transaction database and their popularity, i.e., how many times each individual item is sold. This information can be obtained from a competing company or from published reports.

More formally, let D be the original transaction database that the owner has and D^* its private version. The server or an intruder, who gains access to it, may possess some background knowledge using which they can conduct attacks on the encrypted database D^*. The attacker knows exactly the set of (plain) items $\mathcal{I}$ in the original transaction database D and their true supports in D. The service provider (who can be an attacker) can use his background knowledge to make inferences on the encrypted transactions D^*. The data owner (i.e., the corporate) considers the true identity of every cipher item, every cipher transaction, and every cipher frequent pattern as the intellectual property which should be protected. Therefore, the following attacks are considered:

- *Item-based attack*: $\forall$ cipher item $e \in \mathcal{E}$, the attacker constructs a set of candidate plain items $\text{Cand}(e) \subset \mathcal{I}$. The probability that the cipher item e can be broken $\text{prob}(e) = 1/|\text{Cand}(e)|$.
- *Set-based attack*: Given a cipher itemset E, the attacker constructs a set of candidate plain itemsets $\text{Cand}(E)$, where $\forall X \in \text{Cand}(E)$, $X \subset \mathcal{I}$, and $|X| = |E|$. The probability that the cipher itemset E can be broken $\text{prob}(E) = 1/|\text{Cand}(E)|$.

We refer to prob(e) and prob(E) as *crack probabilities*. From the point of view of the owner, minimizing the probabilities of crack is desirable. Clearly, Cand(e) and Cand(E) should be as large as possible; in particular, Cand(e) should be the whole set of plaintext items. This can be achieved by bringing each cipher item to the same level of support, e.g., to the support of the most frequent item in D. This option would lead to a dramatic explosion of the frequent patterns making pattern mining at the server side computationally prohibitive. [21] proposes of relaxing the equal-support constraint introducing item k-anonymity as a compromise.

4.3 Privacy-preserving technique

How to counter the above attacks while assuring for the client the ability of obtaining the correct collection of frequent patterns? A possible solution is applying an encryption scheme that transforms the original database by the following steps: (I) replacing each item by a 1-1 substitution function; and (II) adding fake transactions to the database in such a way that each item (itemset) becomes indistinguishable with at least $k-1$ other items (itemsets). On the basis of this simple idea, this framework guarantees that not only individual items, but also any group of items has the property of being indistinguishable from at least k other groups in the worst case, and actually many more in the average case. This protection implies that the attacker has a very limited probability of guessing the actual items contained either in the sale data or in the mining results. On the contrary, the data owner can efficiently decrypt correct mining results, returned by the third party, with limited computational resources. Indeed, the framework provides a very efficient decryption schema that uses very negligibly small information representing in a compact way the information about the fake transactions added during the encryption phase. This research shows interesting results obtained applying this model over large-scale, real-life transaction databases donated by a large supermarket chain in Europe. The architecture behind the proposed model is illustrated in Figure 4. The client/owner encrypts its transaction database (TDB)

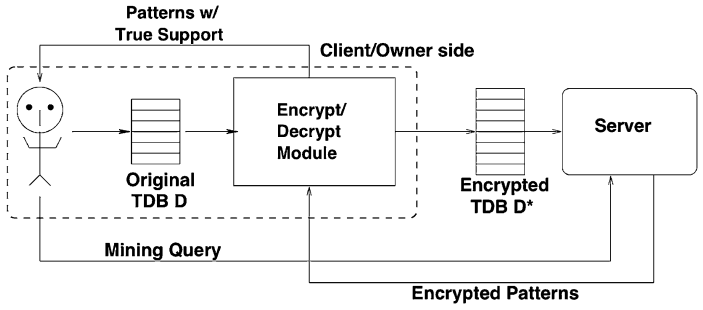

Figure 4 Architecture of Mining-as-Service Paradigm showing that the client before sending data to the server encrypts it and the server answers the mining queries with encrypted patterns that the client can decrypt before their use.

using an encrypt/decrypt module, which can be essentially treated as a 'black box' from its perspective. This module is responsible for transforming the TDB D into an encrypted database D^*. The server conducts data mining and sends the (encrypted) patterns to the owner. The encryption scheme has the property that the returned number of occurrences of the patterns are not true. The encrypt/decrypt module recovers the true identity of the returned patterns as well their true number of occurrences.

The strong theoretical results in [21] show a remarkable guarantee of protection against the attacks presented in Section 4.2, and the practicability and the effectiveness of the proposed schema. The application of this framework on real-world databases showed that the privacy protection is much better than the theoretical worst case. *Why?* The explanation is that the probability of crack generally decreases with the size of the itemset: $\frac{1}{k}$ is an upper bound that essentially applies only to individual items, not itemsets (under the hypothesis that the adopted grouping is robust).

4.4 Frequent pattern mining and privacy protection

The framework is applied to real-world data donated us by Coop, a cooperative of consumers that is today the largest supermarket chain in Italy. The data contain transactions occurring during four periods of time in a subset of Coop stores, creating in this way four different databases with varying number of transactions: from 100k to 400k transactions. In all the datasets the transactions involve 15,713 different products grouped into 366 marketing categories. Two distinct kind of databases are considered: (i) product-level (*CoopProd*) where items correspond to products, and (ii) category-level databases (*CoopCat*), where items are category of the products.

Crack probability. The analysis of the crack probability for transactions and patterns in both databases *CoopProd* and *CoopCat* highlighted that after the data transformation around 90% of the transactions can be broken with probability strictly less than $\frac{1}{k}$. For example, considering the encrypted version of *CoopProd* with 300K transactions, the experiments showed the following facts, even for small k. For instance, for $k = 10$, every transaction E has at least 10 plain itemset candidates, i.e., $\text{prob}(E) \leq \frac{1}{10}$. Around 5% of transactions have exactly a crack probability $\frac{1}{10}$, while 95% have a probability strictly smaller than $\frac{1}{10}$. Around 90% have a probability strictly smaller than $\frac{1}{100}$.

Frequent pattern mining. The schema proposed, i.e., the *encryption* of the transactions and the *decryption* of the patterns enable the client to recover the true identity of the

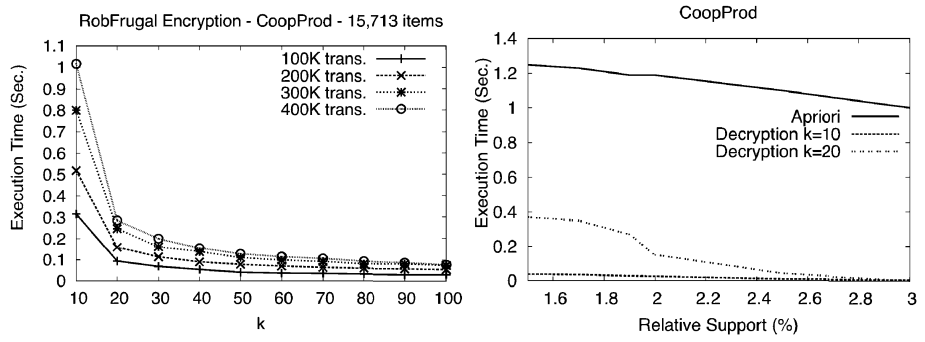

Figure 5 Encryption and Decryption Overhead. (Left) Encryption Overhead by varying the *k* value and the number of transactions; (Right) Comparison of Decryption Overhead and mining cost on the original data using Apriori by varying the minimum support threshold for the pattern extraction.

returned patterns as well their true number of occurrences. Therefore, for the client there is no quality loss for the set of the mined patterns.

An important aspect in the data mining outsourcing is the efficiency of the encryption/decryption schema because one of the motivation of the data mining outsourcing is the lack of computational resources for mining of some companies. In Figure 5(left) we can see that, when there are multiple mining queries, which is always the case for the outsourcing system, the encryption overhead of the proposed scheme is negligible compared with the cost of mining. Figure 5(right) shows that the decryption time is about one order of magnitude smaller than the mining time; for higher support threshold, the gap increases to about two orders of magnitude. The situation is similar in CoopCat. The results also show that the encryption time is always small; it is under 1 second for the biggest *CoopProd*, and below 0.8 second for the biggest *CoopCat*. Moreover, it is always less than the time of a single mining query, which is at least 1 second by Apriori (Figure 5(right)).

5 Privacy-by-design in distributed analytical systems

The previous Section 3 and Section 4 how we can apply the privacy-by-design methodology for guaranteeing individual privacy in a setting where we have a central trusted aggregation center that collects data and before releasing it can apply a privacy transformation strategy to enable collective analyses in a privacy-aware fashion.

However, privacy-by-design paradigm can also be applied with success to distributed analytical systems where we have a untrusted central station that collects some aggregate statistics computed by each individual node that observes a stream of data. In this section we discuss an instance of this case [22]; in particular, we show as the privacy-by-design methodology can help in the design of a privacy-aware distributed analytical processing framework for the aggregation of movement data. We consider the data collector nodes as on-board location devices in vehicles that continuously trace the positions of vehicles and periodically send statistical information about their movements to a central station. The central station, which we call *coordinator*, will store the received statistical information and compute a summary of the traffic conditions of the whole territory, based on the information collected from data collectors.

We show how privacy can be obtained before data leaves users, ensuring the utility of some data analysis performed at 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. As discussed in Section 3, the aim of this framework is to provide both *individual* privacy protection by the differential privacy model and acceptable *collective* data utility.

5.1 State-of-the-art on privacy-preserving distributed data analytics

A privacy model particularly suitable for guaranteeing individual privacy while answering to aggregate queries is *differential privacy* [36]. Recently, much attention has been paid to use differential privacy for distributed private data analysis. In this setting n parties, each holding some sensitive data, wish to compute some aggregate statistics over all parties' data with or without a centralized coordinator. [37, 38] prove that when computing the sum of all parties' inputs without a central coordinator, any differentially-private multiparty protocol with a small number of rounds and small number of messages must have large error. Rastogi et al. [39] and Chan et al. [40] consider the problem of privately aggregating sums over multiple time periods. Both of them consider malicious coordinator and use both encryption and differential privacy for the design of privacy-preserving data aggregation methods. Compared with their work, we focus on semi-honest coordinator, with the aim of designing privacy-preserving techniques by adding meaningful noises to improve data utility. Furthermore, both [39, 40] consider aggregate-sum queries as the main utility function, while we consider network flow based analysis for the collected data. Different utility models lead to different design of privacy-preserving techniques. We agree that our method can be further enforced to against the malicious coordinator by applying the encryption methods in [39, 40].

5.2 Attack and privacy model

As in the case analyzed in Section 3, we consider as sensitive information any data from which the typical mobility behavior of a user may be inferred. This information is considered sensitive for two main reasons: (1) typical movements can be used to identify the drivers who drive specific vehicles even when a simple de-identification of the individual in the system is applied; and (2) the places visited by a driver could identify peculiar 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 intruder between the node and the coordinator (i.e., attacks during the communications), and from any intruder 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 participant, like in the previously (and diffusely) discussed linking attack.

The solution proposed in [22] is based on *Differential Privacy*, a recent model of randomization introduced in [41] by Dwork. The general idea of this paradigm 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 ε-differential privacy, due to the

level of privacy guaranteed ε. Note that when ε tends to 1 very little perturbation is introduced and this yields a low privacy protection; on the contrary, better privacy guarantees are obtained when ε tends to zero. Differential privacy assures a record owner that any privacy breach will not be a result of participating in the database since anything, or almost nothing, that is learnable from the database with his record is also learnable from the one without his data. Moreover, in [41] is formally proved that ε-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.

Here, we do not provide the formal definition of this paradigm, but we only point out that the mechanism of differential privacy works by adding appropriately chosen random noise (from a specific distribution) to the true answer, then returning the perturbed answer. A little variant of this model is the (ε, δ)-differential privacy, where the noise is bounded at the cost of introducing a privacy loss. A key notion used by differential privacy mechanisms is the *sensitivity* of a query, that provides a way to set the noise distribution in order to calibrate the noise magnitude on the basis of the type of query. 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 a 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.

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 τ? 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 (ε, δ)-differential privacy, and provides a good balance between privacy and data utility.

5.3 Privacy-preserving technique

First of all, each participant must share a common partition of the examined territory; 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 Section 3.3. Once the partition is shared, each trajectory is generalized as a sequence of crossed areas (i.e., a sequence of movements). For convenience's sake, this information is mapped onto a *frequency vector*, linked to the partition. Unfortunately, releasing frequency of moves instead of raw trajectory data to the coordinator is not privacy-preserving, as the intruder may still infer the sensitive typical movement information of the driver. As an example, the attacker could learn the driver's most frequent move; this information can be very sensitive because such move usually corresponds to a user's transportation between home and workplace. Thus, the proposed solution relies on the differential privacy mechanism, using a Laplace distribution [36]. At the end of the predefined time interval τ, before sending the frequency vector to the coordinator, for each element in the vector the node 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 step the node V_j transformed its frequency vector f_{V_j} into its private version $\tilde{f_{V_j}}$. This ensures the respect of the ε-differential privacy. This simple general strategy has some drawbacks:

first, it could lead to large amount of noise that, although with small probability, can be arbitrarily large; second, adding noise drawn from the Laplace distribution could generate negative frequency counts of moves, which does not make sense in mobility scenarios. To fix these two problems, it is possible to bound the noise drawn from the Laplace distribution, reducing to an (ε, δ) 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 should be in the interval $[0, 2x]$. This approach satisfies (ε, δ)-differential privacy, where δ measures the privacy loss. Note that, since in a distributed environment a crucial problem is the overhead of communications, it is possible to reduce the amount of transmitted information, i.e., the size of frequency vectors. In [22], a possible solution of this problem is reported, but given that this is beyond the purpose of the current paper, we omit this kind of discussion.

5.4 Analytical quality

So far we presented the formal guarantees to individual privacy preservation, but we have to show yet if the individually transformed values are still useful once they are collected and aggregated by the coordinator, i.e., if they are still suitable at collective level for analysis. In the proposed framework, the coordinator collects the perturbed frequency vectors from all the vehicles in the time interval τ and sums them movement by movement. This allows obtaining the resulting global frequency vector, which represents the flow values for each link of the spatial tessellation. Since the privacy transformation operates on the entries of the frequency vectors, and hence on the flows, we present 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 4,200 vehicles, generating around 15,700 trips. The τ interval is one day, so the global frequency vector represents the sum all the trajectories crossing any link, at the end of each day.

Figure 6 shows the resulting Complementary Cumulative Distribution Functions (CCDFs) of different privacy transformation varying ε from 0.9 to 0.01. Figure 6(left) shows the reconstructed flows per link: fixed a value of flow (x) we count the number of links (y) that have that flow. Figure 6(right) shows the distribution of sum of flows passing for each zone: given a flow value (x) it shows how many zones (y) present that total flow. From the distributions we can notice how the privacy transformation preserves very well the distribution of the original flows, even for more restrictive values of the parameter ε. Also considering several flows together, like those incident to a given zone (Figure 6(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 high utility.

Qualitatively, Figure 7 shows a visually comparison of results of the privacy transformation with the original ones. This is an example of two kind of visual analyses that can be performed using mobility data. Since the global complementary cumulative distribution functions are comparable, we can choose a very low epsilon (ε = 0.01) with the aim to

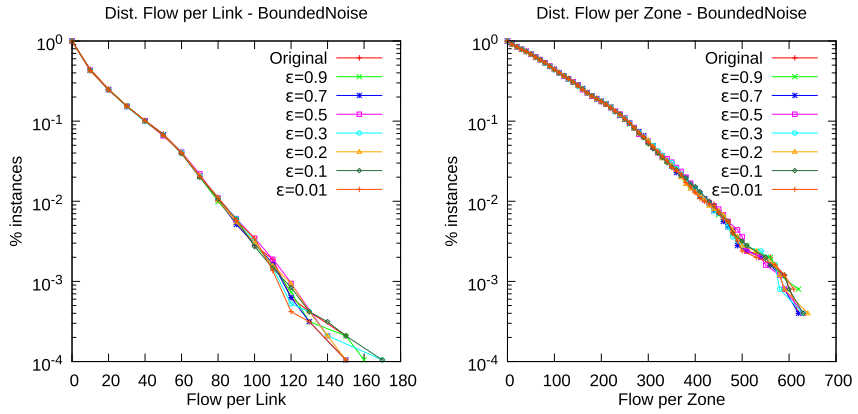

Figure 6 CCDFs of collective flows. (Left) CCDFs of *Flow per Link* for different levels of protection ε; (Right) CCDFs of *Flow per Zone* for different levels of protection ε.

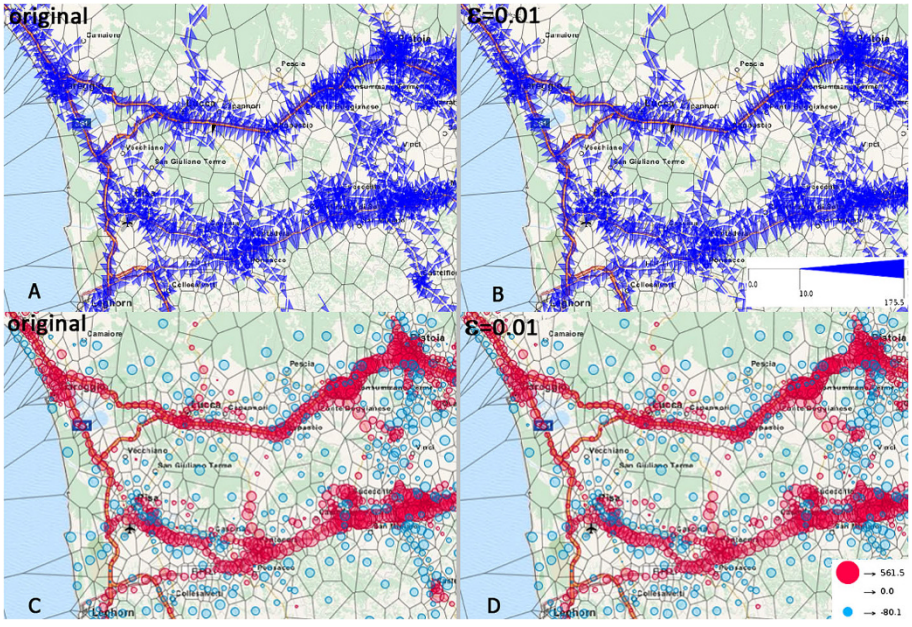

Figure 7 Visual comparison of *Flow per Link* and *Flow per Zone* measures: the resulting overview after the privacy transformation preserves relevant information and properties. Top: flows are rendered as arrows with width proportional to the flow between two regions. Bottom: flow in a region is represented as a circle whose radius is proportional to deviation from median volume in all zones.

emphasize the very good quality of mobility analysis that an analyst can obtain even if the data are transformed by using a very low ε value, i.e. obtaining a better privacy protection. In Figure 7(A) and (B) each flow is drawn with arrows with thickness proportional to the volume of trajectories observed on a link. From the figure it is evident how the relevant flows are preserved in the transformed global frequency vector, revealing the major highways and urban centers. Similarly, the *Flow per Zone* is also preserved, as it is shown in Figure 7(C) and (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

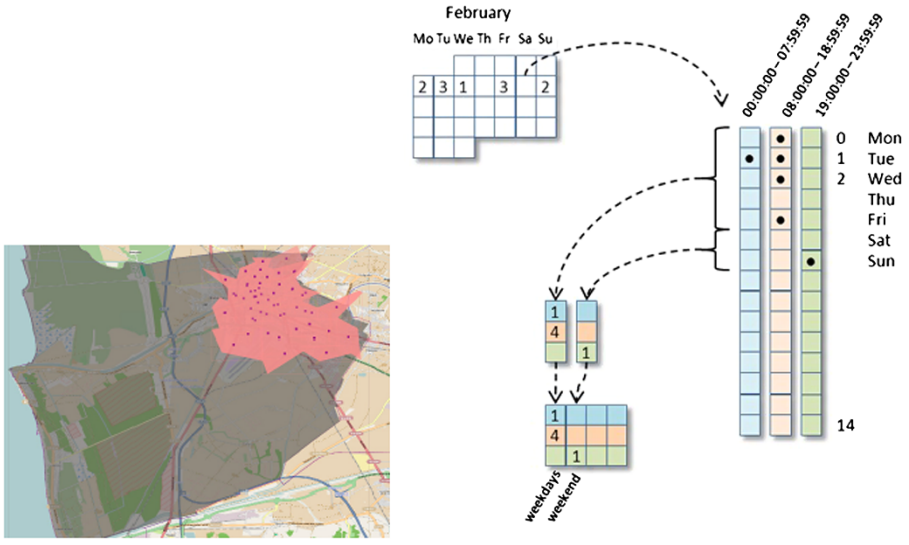

Figure 8 Methodology for GSM User Profile Construction. (Left) GSM Cell coverage in the area of Pisa; (Right) Schematic representation of reconstruction of the Temporal Profile for users in Pisa.

areas (blu circle below the median). The high density traffic zones follow the highways and the major city centers along their routes. The two comparisons proposed above give the intuition that, while the transformations protect individual sensitive information, the utility of data is preserved.

6 Privacy-by-design in user profiling in GSM data

In this section we study the privacy guarantees of the knowledge discovery process introduced in [23] and we show that it can be made in a privacy by design manner by applying some small change to the process that will not affect the final result of the analysis.

In [23] the authors present an analytical process for user profiling in GSM data; in other words, the proposed methodology identifies a partition of the users tracked by GSM phone calls into profiles like *resident*, *commuters* and *visitors* and quantifies the percentage of the different profiles.

The profiling methodology is based on an machine learning step using SOM [42] applied to spatio-temporal user profiles extracted from people call habits. In particular, the whole analytical process is composed of the following steps:

1. Select from the whole network the cells overlapping the area to which we are interested for the analysis (see Figure 8(left) as an example);
2. Build a time projection by two temporal operations (Figure 8(right)): (a) the aggregation of the days in weekdays and weekend slots; (b) the splitting of each slot in time intervals representing 3 interesting time windows during the day;
3. Construct for each user the Space Constrained Temporal Profile (SCT profile) [23] by using the CDR logs according to the space constraints (Point 1.) and the time projection (Point 2.). A SCT profile P is an aggregation of call statistics according to a given temporal discretization where only the calls performed in the cells, contained within the a certain area, are considered. In particular, each profile P is a matrix and each position P_{ij} contains the value v that corresponds to the number of days with at least one call from the user in the area of interest during the set of days j and the time

slot i. As an example, in Figure 8(right), $P_{2,1} = 4$ means that the user visited the area of interest 4 days during the weekdays of the first week and always in the time interval [08:00:00-18:59:59]. In the following we denote by $\mathcal{P}$ the set of SCT profiles extracted from CDR logs.

4. The set of SCT profiles, that are a concise representation of the users' behaviors measured by their calls, is then processed by using the SOM algorithm in order to extract the typical global profiles.
5. The SOM output is a set of nodes representing groups of users with similar temporal profiles; therefore, counting the instances in each group, it is possible to estimate the percentage of residents, commuters and visitors.

6.1 State-of-the-art on privacy in GSM data

Relatively, little work has addressed privacy issues in the publication and analysis of GSM data. In the literature, many works that treat GSM data state that in this context there is no privacy issue or at least the privacy problems are mitigated by the granularity of the cell phone. However, recently Golle and Partridge [43] showed that a fraction of the US working population can be uniquely identified by their *home* and *work* locations even when those locations are not known at a fine scale or granularity. Given that the locations most frequently visited by a mobile user often correspond to the home and work places, the risk in releasing locations traces of mobile phone users appears very high.

Privacy risks even in case of releasing of location information with not fine granularity are studied in [44]. In particular, authors look at the same problem of [43] but from a different perspective. They consider the 'top N' locations visited by each user instead of the simple home and work. The basic idea of this work is that more generally the number N of top preferential locations determines the power of an adversary and the safety of a user's privacy. Therefore, we can say that more top locations an adversary knows about a user, the higher is the probability to re-identify that user. The fewer top locations a user has, the safer they are in terms of privacy. [44] presents a study on 30 billion CDRs from a nationwide cellular service provider in the United States with location information for about 25 million mobile phone users on a period of three months. The study highlights important factors that can have a relevant impact on the anonymity. Examples are the value of N in finding the top N locations, the granularity level of the released locations, the fact that the top locations are sorted or not, the availability of additional social information about users, and geographical regions. The outcomes of this study is that the publication of anonymized location data in its original format, i.e. at the sector level or cell level, put at risk the user privacy because a significant fraction of users can be re-identified from the anonymized data. Moreover, it was shown that different geographical areas have different levels of privacy risks, and at a different granularity level this risk may be higher or lower than other areas. When the spatial granularity level of the cell data is combined with time information and a unique handset identifier, all this information can be used to track people movements. This requires that a good privacy-preserving technique has to be applied when analysis such data. Unfortunately, the current proposals, as those presented in [45, 46], do not consider this aspect. However, the work in [46] is very interesting because studies user re-identification risks in GSM networks in the case user historical data is available to characterize the mobile users *a priori*.

6.2 Attack model and privacy by design solution

Given the above overview about the methodology for extracting global profiles and for computing a quantification of the different kinds of global profiles, now we analyze the privacy risks of the users.

We can identify three main phases in this process: (a) the extraction of the SCT profile for each user; (b) the extraction of global profiles; and (c) the quantification of different kinds of global profiles.

It is immediate to understand that the publication of the final result, i.e., the quantification of the global profiles cannot put at risk the individual privacy of any user because this information is a simple aggregation that does not contain any sensitive information about the single users. This means that an attacker by accessing this kind of data cannot infer any information about a user.

The first phase instead is more problematic for the individual privacy of users because requires to access the CDR data that contains all information about the user calls. In particular, for each user call we have the identifiers of the cell where the call starts and ends respectively and the date and time when the call starts, and its duration. The positional accuracy of cells is few hundred meters in a city [47] and when this information is combined with the time information all this information can help to track people movements. In [48] authors studied the user re-identification risks in GSM networks and showed that it is possible to identify a mobile user from CDR records and a pre-existing location profile, based on previous movements. In particular, one of the re-identification methods that they propose allows for the identification of around 80% of users. As a consequence, this kind of data can reveal sensitive user behavior and the telecommunication operator cannot release this data to the analyst without any privacy-preserving data transformation.

However, we observe that the only information that the analyst needs for computing the global profiles and their quantification is the set of SCT profiles; therefore, we propose an architecture where, the telecommunication operator computes the SCT profiles and then sends them to the analyst for the computation of the step (b) and (c). This solution avoids the access to the CDR logs for the analyst while provides to him the minimum information to performing the target analysis with correctness.

Now the question is: *Can an attacker infer private information about a user by accessing the set of SCT profiles? Is this form of data enough for protecting the individual privacy of each user in the system?* If the answer to this last question is yes, we could have both individual privacy protection and perfect quality of the analytical results.

First of all, we observe that a SCT profile can be seen as a spatio-temporal generalization of the CDR data of a user. Clearly, this form of data is more aggregated w.r.t. the CDR logs because it cannot reveal the history of the user movements, the number of calls and the exact day and time of each call. Moreover, this profile is constructed by considering a specific area such as a city therefore, it is impossible to infer where exactly the user went with a finer granularity. The only information that he can infer is that a specific user visited the city in a specific aggregated period. As an example, an attacker could understand that a given user went to Pisa during a specific week-end if the profiles that he was accessing are related to people in Pisa.

However, in the following we identify two possible attack models, based on the *linking attack*, that use two different background knowledge. Then, we simulate this two attacks on real-world data for showing the privacy protection provided by our schema.

Background knowledge 1. We assume that the attacker knows a set of locations visited by a user U where he called someone and the time of these calls. This means that he can build a SCT profile PB with this background knowledge, where $PB_{ij} = -1$ if the attacker does not have any information about the call activity of the user in the period (i,j) while $PB_{ij} = \nu$, with $\nu > 0$, if from the background knowledge he derives that the user was present in the area ν times in the period (i,j).

Attack model 1. The attacker, who gains access to the set of SCT profiles, uses the background knowledge PB on the user U to match all the profiles that include PB. The set of matched profiles is the set $C = \{P \in \mathcal{P} | \forall PB_{ij} \geq 0, PB_{ij} \leq P_{ij}\}$. The probability of re-identification of the user U is $\frac{1}{|C|}$. Clearly, a greater number of candidates corresponds to a more privacy protection.

Background knowledge 2. In our study we also consider a different background knowledge. We assume that the attacker for some time periods (i,j) knows the exact number of times that a user U visited locations in the area of interest. This means that he can build a profile PB with this background knowledge, where $PB_{ij} = -1$ if the attacker does not have any information about the presence of the user U in the area of interest during the period (i,j) while $PB_{ij} = \nu$ with $\nu \geq 0$, if from the background knowledge he derives that the user was present in the area ν times in the period (i,j). As an example, suppose an adversary knows that during the first week Mr. Smith went to Pisa in the time interval [08:00:00-18:59:59] only 4 times over 5 because Friday he was sick, Then, from this information he can construct a profile PB where $PB_{21} = 4$ while the other entries are equal to -1. Note, that in this case the attacker does not know if the user U did a call during his presence in the area of interest of the analysis and this implies that the malicious part does not know if the user U is represented in the set of profiles.

Attack model 2. The attacker, who gains access to the set of SCT profiles uses the background knowledge PB on the user U to select the set of candidate profiles $C = \{P \in \mathcal{P} | \forall PB_{ij} \geq 0, P_{i,j} \leq PB_{ij}\}$. The re-identification probability of the user U is $\frac{1}{|C|} \times Prob$, where $Prob$ is the probability that one of the profiles in $\mathcal{P}$ belongs to user U.

6.3 Privacy protection analysis

We performed a series of experiments on a real GSM dataset. We obtained a dataset of CDR logs in the Province of Pisa during the period from January 9th to February 8th 2012 reporting the activities of around 232k persons, for a total of 7.8M call records. Focusing on the urban area of the city of Pisa, we extracted the SCT profiles for the 63k users performing at least one call activity in the observation period. We then simulated two attacks according to the two attacking models above and measured the re-identification probability of each SCT profile.

The simulation is performed as follows: we generate a series of profiles PB according to the background knowledge 1 (background knowledge 2). These profiles are derived from the real user SCT profiles in the dataset. Then, we have performed the attack 1 (attack 2) on the set of profiles $\mathcal{P}$.

Concerning the attack 2 we have assumed that the adversary knows the exact number of times that the user visited locations in Pisa for each period (i,j), i.e., for all the 4 weeks in the profiles. Figure 9 shows the cumulative distribution of the re-identification probability. We found that in the worst case the probability of re-identification is 0.027% and only about 5% of users in the set of SCT profiles have this level of risk, while the other users

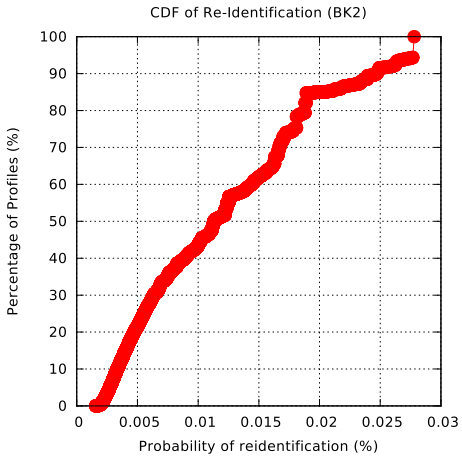

Figure 9 Distribution of the re-identification probability obtained by simulating attack model 2 with a background knowledge of 4 weeks.

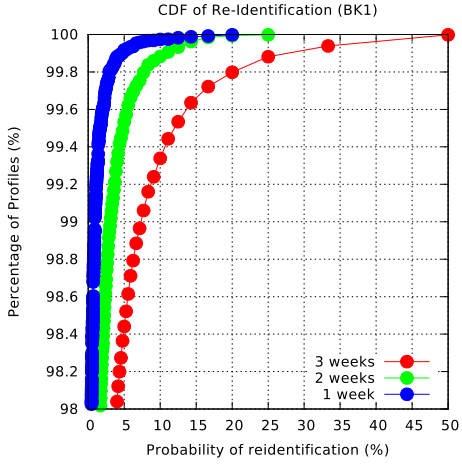

Figure 10 Distribution of the re-identification probability obtained by simulating attack model 1 with different levels of background knowledge.

have a lower risk of privacy violation. This very high protection is due to the fact that with the background knowledge 2 (BK2) the attacker is not sure that the user is in the set of profiles that he is observing. However, even if we assume that he knows that the specific user is represented in the set of SCT profiles, the probability of re-identification is always low. We indeed have observed that the highest probability of re-identification in this case is 0.21%

Concerning the attack 1, that is based on a stronger background knowledge, we have assumed that the attacker knows the user call activities for a specific number of weeks and we have measured the probability to re-identify the user and infer his activities in the remaining weeks. Figure 10 shows the cumulative distribution of the re-identification probability for different levels of background knowledge: 1 week, 2 weeks and 3 weeks. As expected, when we increase the periods of observations of the adversary we have a worst privacy protection. However, when the attacker knows 1 week or 2 weeks of call activities of a specific users the probability of re-identification is always no more than 20% and 25% and this happens for about 0.01% of user in the profile data; 99.99% of users has a lower privacy risk. When we consider a observation period of 3 weeks the privacy protection decreases and for less than 0.1% of users the probability of re-identification is

50%, while for more than 99.9% of people the probability of re-identification is no more than 32%. Moreover, the 99% of users has a risk of re-identification less than about 7%. Clearly, here it is important to note that the background knowledge that we are taking into consideration is very strong. We have also measured the risk of re-identification assuming that the attacker knows the user call activities of different periods of the SCT profile. This kind of attack is similar to that one in [49] where authors discovered that 4 observations are enough to uniquely identify 95% of the individuals. In our experiments by using the SCT profiles instead of CDR logs, we have found that with 10 observations the probability of re-identification is less than 20% for all the users and about 99% of people has a risk of re-identification of about 1%. While if we consider 20 observations the situation is very similar to the case in which the attacker knows 3 weeks of calls of user in Figure 10.

The conclusion is that the illustrated process shows as by knowing the analysis to be performed on the data it is possible to transform the original data in a different form (by aggregations) and find a representation that both contains all the proprieties useful for obtaining a perfect analytical result and preserves the user privacy.

7 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 describes sensitive human activities and the privacy of people is always more at risk. The danger is increasing also thanks to the emerging capability to integrate diversified data. In this paper, 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 the privacy harm, or even prevent the privacy harm. We have discussed how applying the privacy-by-design principle to four 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.

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

Authors' contributions
All authors contributed equally to the writing of this paper. All authors read and approved the final manuscript.

Acknowledgements
This work has been partially supported by EU FET-Open project DATA SIM (FP7-ICT 270833) and EU Project PETRA n. 609042 (FP7-SMARTCITIES-2013).

Received: 4 December 2013 Accepted: 1 August 2014 Published online: 24 September 2014

References

1. Batty M, Axhausen KW, Giannotti F, Pozdnoukhov A, Bazzani A, Wachowicz M, Ouzounis G, Portugali Y (2012) Smart cities of the future. Eur Phys J Spec Top 214(1):481-518
2. Brockmann D, Hufnagel L, Geisel T (2006) The scaling laws of human travel. Nature 439(7075):462-465
3. Giannotti F, Nanni M, Pedreschi D, Pinelli F, Renso C, Rinzivillo S, Trasarti R (2011) Unveiling the complexity of human mobility by querying and mining massive trajectory data. VLDB J 20(5):695-719
4. Gonzalez MC, Hidalgo CA, Barabasi A-L (2008) Understanding individual human mobility patterns. Nature 453(7196):779-782
5. Song C, Koren T, Wang P, Barabasi A-L (2010) Modelling the scaling properties of human mobility. Nat Phys 6(10):818-823
6. Song C, Qu Z, Blumm N, Barabási A-L (2010) Limits of predictability in human mobility. Science 327(5968):1018-1021
7. Wang D, Pedreschi D, Song C, Giannotti F, Barabási A-L (2011) Human mobility, social ties, and link prediction. In: KDD, pp 1100-1108

8. Balcan D, Gonçalves B, Hu H, Ramasco JJ, Colizza V, Vespignani A (2010) Modeling the spatial spread of infectious diseases: the global epidemic and mobility computational model. J Comput Sci 1(3):132-145
9. Colizza V, Barrat A, Barthelemy M, Valleron AJ, Vespignani A (2007) Modeling the worldwide spread of pandemic influenza: baseline case and containment interventions. PLoS Med 4(1):95-110
10. Colizza V, Barrat A, Barthélemy M, Vespignani A (2006) The role of the airline transportation network in the prediction and predictability of global epidemics. Proc Natl Acad Sci USA 103(7):2015-2020
11. Fumanelli L, Ajelli M, Manfredi P, Vespignani A, Merler S (2012) Inferring the structure of social contacts from demographic data in the analysis of infectious diseases spread. PLoS Comput Biol 8(9)
12. Gallos L, Havlin S, Kitsak M, Liljeros F, Makse H, Muchnik L, Stanley H (2010) Identification of influential spreaders in complex networks. Nat Phys 6(11):888-893
13. El Emam K, Cavoukian A (2014) De-identification protocols: essential for protecting privacy. http://www.privacybydesign.ca/content/uploads/2014/06/pbd-de-identifcation_essential.pdf
14. Abul O, Bonchi F, Nanni M (2008) Never walk alone: uncertainty for anonymity in moving objects databases. In: Proceedings of the 2008 IEEE 24th international conference on data engineering (ICDE), pp 376-385
15. Domingo-Ferrer J, Trujillo-Rasua R (2012) Microaggregation- and permutation-based anonymization of movement data. Inf Sci 208:55-80
16. Monreale A, Pedreschi D, Pensa RG (2010) Anonymity technologies for privacy-preserving data publishing and mining. In: Privacy-aware knowledge discovery: novel applications and new techniques, pp 3-33
17. Pensa RG, Monreale A, Pinelli F, Pedreschi D (2008) Pattern-preserving *k*-anonymization of sequences and its application to mobility data mining. In: PiLBA
18. Wong WK, Cheung DW, Hung E, Kao B, Mamoulis N (2007) Security in outsourcing of association rule mining. In: VLDB, pp 111-122
19. Cavoukian A (2000) Privacy design principles for an integrated justice system. Working paper. www.ipc.on.ca/index.asp?layid=86&fid1=318
20. Monreale A, Andrienko GL, Andrienko NV, Giannotti F, Pedreschi D, Rinzivillo S, Wrobel S (2010) Movement data anonymity through generalization. Trans Data Privacy 3(2):91-121
21. Giannotti F, Lakshmanan LVS, Monreale A, Pedreschi D, Wang WH (2013) Privacy-preserving mining of association rules from outsourced transaction databases. IEEE Syst J 7(3):385-395
22. Monreale A, Wang WH, Pratesi F, Rinzivillo S, Pedreschi D, Andrienko G, Andrienko N (2013) Privacy-preserving distributed movement data aggregation. In: AGILE. Springer, Berlin. doi:10.1007/978-3-319-00615-4_13
23. Furletti B, Gabrielli L, Renso C, Rinzivillo S (2012) Identifying users profiles from mobile calls habits. In: UrbComp'12, pp 17-24
24. (2010) Privacy by design resolution. In: International conference of data protection and privacy commissioners, Jerusalem, Israel, 27-29 october 2010
25. 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)
26. European Data Protection Supervisor (Mar. 18, 2010) Opinion of the European data protection supervisor on promoting trust in the information society by fostering data protection and privacy
27. Federal Trade Commission (Bureau of Consumer Protection) (Dec. 2010) Preliminary staff report, protecting consumer privacy in an era of rapid change: a proposed framework for business and policy makers, at v, 41
28. Samarati P, Sweeney L (1998) Protecting privacy when disclosing information: *k*-anonymity and its enforcement through generalization and suppresion. In: Proc. of the IEEE symp. on research in security and privacy, pp 384-393
29. Samarati P, Sweeney L (1998) Generalizing data to provide anonymity when disclosing information (Abstract). In: PODS, p 188
30. Terrovitis M, Mamoulis N (2008) Privacy preservation in the publication of trajectories. In: Proc. of the 9th int. conf. on mobile data management (MDM)
31. Yarovoy R, Bonchi F, Lakshmanan LVS, Wang WH (2009) Anonymizing moving objects: how to hide a MOB in a crowd? In: EDBT, pp 72-83
32. Nergiz ME, Atzori M, Saygin Y, Güç B (2009) Towards trajectory anonymization: a generalization-based approach. Trans Data Privacy 2(1):47-75
33. Qiu L, Li Y, Wu X (2008) Protecting business intelligence and customer privacy while outsourcing data mining tasks. Knowl Inf Syst 17(1):99-120
34. Tai C, Yu PS, Chen M (2010) *k*-Support anonymity based on pseudo taxonomy for outsourcing of frequent itemset mining. In: KDD, pp 473-482
35. Molloy I, Li N, Li T (2009) On the (in)security and (im)practicality of outsourcing precise association rule mining. In: ICDM, pp 872-877
36. Dwork C, Mcsherry F, Nissim K, Smith A (2006) Calibrating noise to sensitivity in private data analysis. In: Proceedings of the 3rd theory of cryptography conference. Springer, Berlin, pp 265-284
37. Beimel A, Nissim K, Omri E (2008) Distributed private data analysis: simultaneously solving how and what. In: CRYPTO, pp 451-468
38. Chan T-HH, Shi E, Song D (2012) Optimal lower bound for differentially private multi-party aggregation. In: ESA, pp 277-288
39. Rastogi V, Nath S (2010) Differentially private aggregation of distributed time-series with transformation and encryption. In: SIGMOD, pp 735-746
40. Shi E, Chan T-HH, Rieffel EG, Chow R, Song D (2011) Privacy-preserving aggregation of time-series data. In: NDSS
41. Dwork C (2006) Differential privacy. In: Bugliesi M, Preneel B, Sassone V, Wegener I (eds) Automata, languages and programming. Lecture notes in computer science, vol 4052. Springer, Berlin, pp 1-12
42. Kohonen T (2001) Self-organizing maps. Springer series in information sciences, vol 30
43. Golle P, Partridge K (2009) On the anonymity of home/work location pairs. In: Pervasive computing, pp 390-397
44. Zang H, Bolot J (2011) Anonymization of location data does not work: a large-scale measurement study. In: Proceedings of the 17th annual international conference on mobile computing and networking, pp 145-156. ACM

45. Croft NJ, Olivier MS (2006) Sequenced release of privacy accurate call data record information in a GSM forensic investigation. ISSA, Sandton, South Africa
46. De Mulder Y, Danezis G, Batina L, Preneel B (2008) Identification via location-profiling in GSM networks. In: Proceedings of the 7th ACM workshop on privacy in the electronic society, pp 23-32. ACM
47. Trevisani E, Vitaletti A (2004) Cell-ID location technique, limits and benefits: an experimental study. In: WMCSA'04, pp 51-60. IEEE
48. De Mulder Y, Danezis G, Batina L, Preneel B (2008) Identification via location-profiling in GSM networks. In: WPES'08, pp 23-32. ACM
49. de Montjoye Y-A, Hidalgo CA, Verleysen M, Blondel VD (2013) Unique in the crowd: the privacy bounds of human mobility. Scientific reports 3

doi:10.1140/epjds/s13688-014-0010-4
Cite this article as: Monreale et al.: **Privacy-by-design in big data analytics and social mining.** *EPJ Data Science* 2014 **2014**:10.