PRUDEnce: a System for Assessing Privacy Risk vs Utility in Data Sharing Ecosystems

Francesca Pratesi[*,**], **Anna Monreale**[**], **Roberto Trasarti**[*], **Fosca Giannotti**[*], **Dino Pedreschi**[**], **Tadashi Yanagihara**[***]

[*]ISTI CNR, Pisa, Italy.

[**]University of Pisa, Pisa, Italy.

[***]KDDI Corporation of Tokyo, Japan.

E-mail: `francesca.pratesi@isti.cnr.it`

Received 3 July 2017; received in revised form 10 April 2018; accepted 23 April 2018

Abstract. Data describing human activities are an important source of knowledge useful for understanding individual and collective behavior and for developing a wide range of user services. Unfortunately, this kind of data is sensitive, because people's whereabouts may allow re-identification of individuals in a de-identified database. Therefore, Data Providers, before sharing those data, must apply any sort of anonymization to lower the privacy risks, but they must be aware and capable of controlling also the data quality, since these two factors are often a trade-off. In this paper we propose PRUDEnce (Privacy Risk versus Utility in Data sharing Ecosystems), a system enabling a privacy-aware ecosystem for sharing personal data. It is based on a methodology for assessing both the empirical (not theoretical) *privacy risk* associated to users represented in the data, and the *data quality* guaranteed only with users not at risk. Our proposal is able to support the Data Provider in the exploration of a repertoire of possible data transformations with the aim of selecting one specific transformation that yields an adequate trade-off between data quality and privacy risk. We study the practical effectiveness of our proposal over three data formats underlying many services, defined on real mobility data, i.e., presence data, trajectory data and road segment data.

1 Introduction

Large dataset recording human activities, such as records of personal mobility, extracted by vehicular GPS-enabled devices, records of personal purchases and records describing connections among friends in a social network, are key enablers of a new wave of knowledge-based services, as well as of new scientific discoveries. The utilization of human data in commercial services is getting common, but at the same time, raises the concern on leakage of personal information or re-identification. In fact, numerous services have been temporarily put to halt or even out of service because of such issues [1,2].

The paradigm shift towards human knowledge discovery comes with unprecedented opportunities and risks. However, the paradoxical situation we are facing today is that we are fully running the risks, without fully catching the opportunities of big data: on the one

[1] Yomiuri - https://goo.gl/Pxiuny

[2] Tom Tom - https://goo.gl/J8tcuc

hand, we feel that our private space is vanishing in the digital world, and that our personal data can be used without feedback and control; on the other hand, the same data are seized in the databases of companies (telecom companies, insurance companies, ...), which use legal constraints on privacy as a reason for not sharing it with science and society at large, keeping this precious source of knowledge locked to data analysts or service developers.

In Europe, policy-makers have responded to this shift with an update to the data protection legislation, replacing the 1995 Data Protection Directive with a General Data Protection Regulation (GDPR). The GDPR responds to privacy and data protection threats associated with new data practices by strengthening protections for individuals, but also by harmonizing the legal framework to better enable data to flow within Europe. While both these changes are welcome, corporate actors, and especially small and medium enterprises, have struggled to develop the expertise that would enable them to use data to develop innovative products and services, or to generate the expected efficiencies associated with big data. Therefore, it is necessary to enable knowledge discovery from raw data by setting the *data free*, i.e, organizations need to exploit the advantage analyzing available big data while preventing privacy violations, which may result in negative economic and social impacts.

Although in the last years several techniques for data protection and anonymization have been proposed (e.g., k-anonymity based techniques, differential privacy based techniques, etc.), their practical application is inhibited by the results that have a negative impact on the data utility, due to the loss of information caused by the data transformation.

In our vision, in order to increase the practical impact of the privacy-preserving techniques, and making them effectively applicable on large scale, we need a framework enabling a systematic reasoning on the trade-off between privacy protection and data quality. This kind of reasoning represents the key step for the selection of a specific privacy-preserving technique suitable for a given dataset. To this end, we propose PRUDEnce, a system for assessing individual privacy risks and data utility of a specific dataset, which enables the reasoning on the balancing between data protection and data quality. Privacy protection is measured in terms of probability that each specific user is re-identified in the released dataset and data quality is measured in terms of amount of information preserved considering only users with risk below any specified thresholds. PRUDEnce is designed for assisting responsible organizations in the sharing of personal data preventing privacy violations and helping them to consciously choose the proper anonymization method and, possibly, apply it only on the portion of risky data. The final goal is to enable organizations in the sharing of only non-risky data. To this end, PRUDEnce provides an approach that, before applying any privacy-preserving transformation, allows looking at the effective risk there is in the data, as well as the service or purpose for which the data are queried, instead of relying only on theoretical results in terms of privacy. Often it is not necessary to apply the privacy transformation on the whole set of raw data because it is rarely the case that raw data are needed to develop a service. In practice, every service is fueled by specific pre-processing of users data, by means of aggregation, selection and filtering, which may alter significantly the real risk of privacy, compared with the original raw data. For instance, raw GPS tracks of users' cars are typically not needed to develop most personal or social mobility services. Real services require aggregations of data by different spatial or temporal resolutions, may be pertinent to either users' presence at certain locations or users' movements across locations, may span over time intervals of different length, may be filtered according to constraints on the frequency of users' activities, and so on. Different pre-processing, that we refer as *dataviews*, may exhibit very different properties w.r.t. risk. This requires that the actual privacy risk should be taken into account either before applying any further privacy-preserving transformation or by the privacy mechanisms themselves.

If this is not the case, we run the risk of destroying the quality of data without a real need, as often it happens applying strong privacy transformations as differential privacy [5, 20, 16]. Indeed, as highlighted in [32] differential privacy may seriously and unjustifiably damage data because of the computation of the global sensitivity that does not consider the actual data to be protected.

In this complex context, PRUDEnce enables a privacy-aware data sharing ecosystem which permits a *Data Provider* (DP) to share data about its users with a *Service Developer* (SD) after exploring the risks and quality of all possible *dataviews* and selecting the one compatible with its privacy expectation. Our proposed system is general and simply requires the implementation of privacy measures for the risk assessment and of privacy-preserving transformations for the risk mitigation. Given any pair of these two components, our analytical framework enables the DP in the exploration of privacy risks related to the data, in combination of the corresponding data quality. We highlight that in this paper we focus on the privacy risk assessment methodology that is the key step for the selection of the most suitable privacy transformation to be applied on a specific dataset. In this sense, we believe that PRUDEnce can provide practical support to be compliant with some legal advice and obligations. Indeed, in the GDPR (and in particular in Article 35) one can find references to the Data Protection Impact Assessment (DPIA), which is a process that, taking into account the nature, scope, context and purposes of the processing, enables the assessment of the impact of the envisaged processing operations on the protection of personal data. Our privacy risk assessment methodology can help covering a big part of the DPIA, i.e., the one related to the privacy risk, even it overlooks other aspects, for example, all the steps related to the analysis of the cost. However, to the best of our knowledge, PIA does not provides data-driven tools that can effectively measure the privacy risk, but it mainly aims to create self-awareness. Moreover, in Article 25 of the GDPR there is also an explicit reference to the data protection by design and by default, and in Recital 26 of the GDPR, it was stated to take into account any reasonable mean that can be used in re-identifying a natural person. This is totally compliant with the definition of the reasonable attacks used by the PRUDEnce framework. Finally, we believe that the output of our elaboration can also be in line with the principle of transparency advocated by the Article 29 Data Protection Working Party[3] and reported in Recital 39 of the GDPR.

We validate PRUDEnce in the context of location and movement data using a large-scale mobility dataset pertaining to the vehicular GPS traces of tens of thousands private cars observed over a month. In particular, we show how to apply in practice the framework analyzing the user privacy risk in typical mobility data formats useful for developing services such as Point-of-Interest (POI) recommendations, geo-marketing, or positioning of charging stations for electric vehicles. Note that PRUDEnce is a general framework that allows the analysis of the privacy risks in any possible context, thus the services and data formats used in our case studies are only examples used to show the effectiveness of our proposal.

The reminder of the paper is organized as follows. Section 2 introduces the privacy-aware ecosystem for data sharing. Section 3 presents the details about our methodology for the privacy risk assessment. In Section 4 we show how to apply our methodology for the analysis of three different kinds of mobility data. In Section 5 we discuss a possible distributed version of our methodology. Section 6 discuss the impact on privacy when multiple dataviews are released. Lastly, Section 7 discusses the related work and Section 8 concludes the paper.

[3] Article 29 Data Protection Working Party, 17/EN, WP260, Guidelines on transparency under Regulation 2016/679

2 Privacy-aware ecosystem

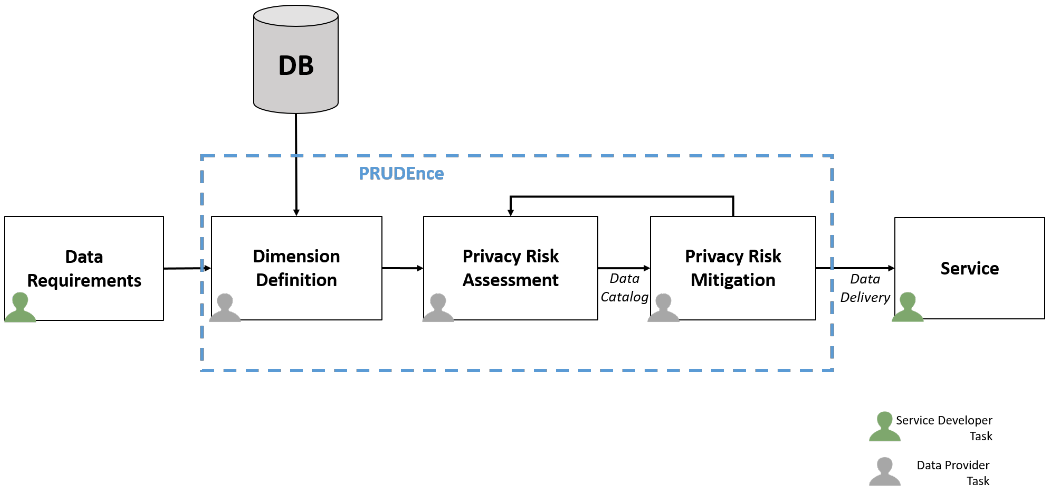

Figure 1: Privacy-aware data sharing ecosystem.

We envision an ecosystem for data sharing, enabled by PRUDEnce, our system of privacy risk assessment, i.e., the measurement of the empirical privacy risk inherent to the data to be transferred from the Data Provider (DP) to the Service Developer (SD). The architecture of the ecosystem is illustrated in Figure 1. Here, SD's aim is to develop services based on information extracted from the raw personal data stored by the DP, which cannot be directly shared without a high risk of compromising users' privacy.[4] Note that the SD may be either another entity with respect to the DP or an untrusted department within the DP organization.

Therefore, DP needs to examine a repertoire of possible transformations of the raw data to the purpose of selecting one specific transformation that yields an adequate trade-off between data quality and privacy risk. The systematic exploration of this search space of possible transformations is precisely the scope of our proposed privacy risk system. It is based on a methodology for the measurement of the empirical privacy risk associated to users in the different possible pre-processing transformations of raw data (e.g., aggregations, selections and filtering), or *dataviews*. In the following by *privacy risk* we intend the probability that a specific user is re-identified in a specific released dataset, i.e., the probability that a specific user's identity is correctly associated to own data, while by *empirical privacy risk* we intend the distribution of the risk, i.e., re-identification probability, over the entire population of users represented in the dataset. Within these assumptions, the ecosystem operates according to the following workflow:

1. SD specifies the *data requirements* for a specific service or a set of services that require the same type and set of data.

[4] It should be remarked that the proposed privacy-aware ecosystem is a systematic implementation of the *Privacy-by-Design* principle [6, 24], and is also compliant with the *data minimization principle*, i.e., the use of the minimum information needed for a purpose, as stated in many international regulations, such as EU Directive 95/46/EC and the Regulation EC (No) 45/2001.

2. DP identifies the *dimensions* along which raw data can be aggregated, selected and filtered.

3. DP generates the collection of *dataviews* with reference to all selected dimensions.

4. DP identifies the possible attacks that a malicious adversary might conduct on a dataview to re-identify users.

5. For each dataview, DP performs the *privacy risk assessment* by empirically measuring, on the user population in the data, the distribution of the probability that attacks succeed. The privacy risk assessment is carried out in tandem with data quality, measured in terms of coverage, i.e., amount of information preserved considering only users with risk below any specified thresholds. The complete repertoire of dataviews with associated risk and quality measurements is referred to as *data catalog*.

6. DP explores the data catalog to select the dataview and the risk threshold representing an adequate trade-off between risk and quality, given the data requirements for the target service and the expected level of tolerated risk.

7. Given the candidate dataview and tolerated risk threshold, DP performs *risk mitigation*, i.e., applies a privacy-preserving transformation to eliminate the users with a risk higher than a certain threshold ensuring that all users in the sanitized dataview fall below that risk. The advantage of applying such sanitization after the privacy risk assessment gives the possibility of focusing on the problematic cases, i.e., users above the tolerated risk, when applying privacy-preserving transformations such as removal, generalization, randomization, etc.

8. DP reiterates the privacy risk assessment on the sanitized dataview, and delivers it to the SD with the measurement of the empirical risk and the final coverage of users' data.

Note that, in the first step SD specifies the properties of the data useful for the service(s). In case SD has to be develop more services requiring different dataviews characterized for example by different time constraints or granularity, it has to activate a separated request.

The key analytical tool to reason about privacy risk and data quality, that we introduce in Section 3.3, is the *Risk and Coverage curve*, or *RAC* curve, for privacy risk assessment. The *RAC* represents how the coverage of users' data varies as a function of the tolerated risk, and it is a concept very similar to the R-U maps (risk-utility maps) [11]. The methodology is obviously independent on the chosen threshold for tolerated risk, which may vary in different circumstances, e.g., whether the SD is the general public, an external third party or an internal department of the DP.

Even if in this paper we consider only the re-identification risk, since it is most known problem in the literature, PRUDEnce is able to incorporate any other risk measure such as the risk of inference about sensitive information provided by the attribute value [19] or other disclosure measures like those defined in [37, 38]. This flexibility is a strong point of our proposal, in fact it is possible to implement and integrate specific function simply overriding the *risk assessment module*.

3 Privacy Risk Assessment

In general, given a database, containing the original whole raw data, i.e., all the detailed information about the users, it is possible to select the portion of data $\mathcal{D}$, useful for developing a specific service. Starting from $\mathcal{D}$, the privacy risk assessment methodology takes into account two aspects: *data dimensions*, that determine the level of detail of data (e.g., the temporal and spatial granularity); and *background knowledge dimensions*, that determine the privacy attacks that can be conducted on the data for the user re-identification. For each combination of *data dimensions* and *background knowledge dimensions* values, the goal is to simulate an attack and empirically quantify the privacy risk, i.e., the risk of re-identification for each user, and the data quality w.r.t. a privacy risk threshold.

3.1 Data Dimensions

Given the dataset $\mathcal{D}$ it is possible to generate different *dataviews*, by considering different values for each data dimension. A data dimension is an element of the data on which we may compute aggregation or filtering of information. Typical dimensions of mobility data are the spatial and temporal granularity. For each dimension we can have a set of possible values, and varying them we can generate different dataviews $DV = \{D_1, D_2, \dots, D_m\}$, which are characterized by their own privacy risk. As an example, consider to develop a parking assistance service that takes as input the set of frequent locations visited by each user. An important dimension of these data is the spatial granularity to define the parking areas. In particular, instead of releasing exact locations we can apply a spatial discretization by a grid and the side size of each cell represents the spatial granularity. For example, we can use a grid with side size of the cell equal to $500m$ or $1000m$ and both are the admissible values of this dimension. By setting the two values for the spatial dimension we generate two different dataviews. Probably, in the second one we have that different locations of the first dataview will be grouped.

3.2 Background Knowledge Dimensions

The background knowledge is the external information that a third party (attacker, service developer, adversary, potential hacker, etc.) knows about a specific user. When the background knowledge is combined with the released data, it makes it possible to associate a record to that user and to enable new inferences. The adversary's background knowledge depends on the context and the data and determines the set of possible attacks. An example of adversary's background knowledge could be a set of locations visited by a user. An adversary might use this information to re-identify that user in the released dataview. In general it is clear that the probability of success of an attack increases with the number of observed locations. In other words, a more detailed background knowledge makes the attack stronger. It is important to clarify that the background knowledge is modeled as a predicate on data, in order to provide a gain for the attacker. In general, the predicate must reflect the best effort of the attacker in matching the BK to the entries in the shared dataview D. The privacy risk assessment methodology for the privacy risk evaluation has to take into account all possible kinds of background knowledge and for each one it has to analyze its dimensions. Here, the dimensions represent the elements of the background knowledge of which it is possible to change the value to model different levels of external knowledge. More formally, we denote by $BK = \{BK_1, BK_2, \dots BK_m\}$ a specific kind of background knowledge, where each BK_i represents the set of background knowledge obtained considering a specific configuration i of values of the dimensions. Continuing the above example, if the user u has the locations l_1, l_2, l_3 we have that the number of observed locations of u is a background knowledge dimension. Therefore, BK_1 represents all possible cases in which "the third party knows 1 location visited by u", i.e., $BK_1 = \{l_1, l_2, l_3\}$,

BK_2 represents all possible cases in which "the third party knows 2 locations visited by u", i.e., $BK_2 = \{(l_1, l_2), (l_1, l_3), (l_2, l_3)\}$, and so on. In practice the background knowledge is generated starting from the data in the dataview. The possible background knowledge known by an adversary on the user u is a subset of locations associated to u in the dataview to be shared. As stated in [35] the worst-case scenario considers an adversary knowing the same data of the shared table. In order to evaluate the trend of the risk changing the quantity of background knowledge possessed by the adversary, we generate for each user all the possible levels of external knowledge starting from the minimum knowledge (only one location in the example) to the maximum one (the whole set of user's locations).

Clearly, the definition of attack models and the corresponding background knowledge is a key and critical step in our framework because requires some expertise. However, only defining different levels of possible external knowledge, it is possible to provide organizations the possibility to reason in a systematic way on the balancing between privacy risks and data utility, for helping them in making responsible and aware decisions.

3.3 Privacy Risk Measures

Analyzing the privacy risk intuitively means that for each combination of *data dimensions* and *background knowledge dimensions* values it is possible to simulate the attack and empirically quantify the privacy risk, i.e., the risk of re-identification for each user and the data quality w.r.t. a privacy threshold.

In the following, we denote by $\mathcal{D}$ the database containing the original raw data useful for developing a specific service, and $D \in DV$ a dataview extracted from $\mathcal{D}$ by considering specific values for the data dimensions. As an example, $\mathcal{D}$ is the database containing all information about the user movements in Pisa and D contains for each user the list of locations visited in a specific month and is obtained by using a spatial discretization by a grid with side size of the cells equal to $500m$. The set of users represented in the raw data $\mathcal{D}$ is called U.

For each user, the risk of re-identification can be computed, by measuring the probability of his re-identification in a released data.

Definition 1 (Probability of re-identification given D)**.** *The probability of re-identification $PR_D(d = u|t)$ denotes the probability to correctly associate a record $d \in \mathcal{D}$ to a unique identity u, given $t \in BK_i$ and that D has been published.*

Let D_u be the set of records that represent the data of the user u in D. The probability of re-identification depends on: (a) the number of records related to the user u in D, compatible with the background knowledge t, i.e., $supp_{D_u}(t)$ and, (b) the number of records in D compatible with the background knowledge t, i.e., $supp_D(t)$. Here, the *compatibility* is expressed by a function that defines if a record of the database matches the background knowledge t. Therefore, we have: $PR_D(d = u|t) = \frac{supp_{D_u}(t)}{supp_D(t)}$. Note that, if the user u is not represented in D we have $PR_D(d = u|t) = 0$.

Now, we define the risk of re-identification, as the worst harmful situation for the user.

Definition 2 (Risk of re-identification)**.** *Given a user u his risk of re-identification is his maximum probability of re-identification given the set of BK_i, i.e., $Risk(u, D) = \max PR_D(d = u|t)$ for $t \in BK_i$.*

This risk has the lower bound $\frac{|D_u|}{|D|}$ measuring the random choice in D. We highlight that if a user $u \notin D$ we have $Risk(u, D) = 0$.

Clearly, the above definitions are related to an exact (even if partial) background knowledge of an adversary; thus, both probability and risk of re-identification are based on a perfect matching. If a partial matching is considered in the re-identification process, then some alternative approach should be used. In the literature, there are two major approaches to solve the problem of partial matching: probabilistic record linkage [18, 39] and distance based record linkage [1, 35].

For each dataview $D \in DV$ and for each background knowledge $t \in BK_i$ the risk of re-identification can be computed and used to describe how the risk is distributed over the user population; to this aim we introduce the following curve.

Definition 3 (Risk and Coverage curve w.r.t. users)**.** *The RAC_U curve is the function that for each risk value r, quantifies the percentage of users in the dataset $\mathcal{D}$ having at most that risk. It is defined as:*

$$RAC_U(r, D) = \frac{|\{u \in U | Risk(u, D) \leq r\}|}{|U|}.$$

In order to compare distinct RAC_U curves of different dataviews in DV, we define the following index representing the whole curve with a single number:

$$i - RAC_U(r, D) = \int_0^1 RAC_U(r, D).$$

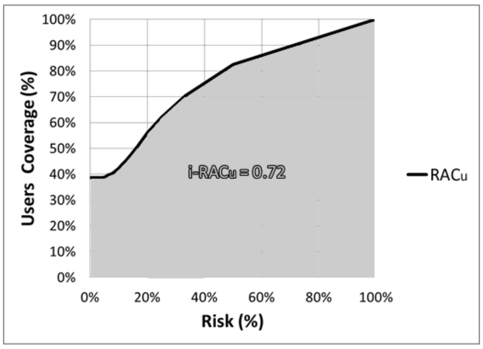

Figure 2: An example of RAC and the relative $i - RAC$ index.

Figure 2 shows the geometrical representation of a RAC_U and its index.

In general, a user has associated many data in D, therefore an important aspect to be measured is the percentage of data covered by users in D having at most a specific privacy risk. This gives a hint about the data coverage w.r.t. a selected risk and, as above, we can compute a curve representing how the data coverage changes w.r.t. the distribution of risk.

Note that RAC_U is not able to capture this aspect because it may happen that the same user is associated to more than one record in the same dataview. Therefore, we define the RAC_D curve as the function that for each risk value r, quantifies the percentage of records in D that are covered by users having at most the risk r. In other words, given $U_r = \{u \in U | Risk(u, D) \leq r\}$ and let D_{U_r} be the set of data covered by users in U_r we define

$$RAC_D(r, D) = \frac{|D_{U_r}|}{|D|}.$$

This formulation has a limit: it permits the computation of data coverage only in terms of records associated to users. Sometimes, given a kind of data it is necessary to measure the quantity of data in a different way. For example, given a dataset of trajectories one might measure the quantity of data in terms of number of trajectories, number of locations, or space covered by the trajectories. Or, also, in case of spatial clusters of trajectories we can measure the number of clusters involved, their traffic flows or the space covered by them. To solve this problem we assume to have a function $c(U', D)$ that takes a dataset D and a set of users U' and returns a quantification of the data within D covered by only users in U'. Therefore, we define the Risk and Coverage curve w.r.t. data as follows.

Definition 4 (Risk and Coverage curve w.r.t. data)**.** *The RAC_D curve is the function that for each risk value r, quantifies the percentage of data in D that are covered by users having at most the risk r. Given $U_r = \{u \in U | Risk(u, D) \leq r\}$ is defined as:*

$$RAC_D(r, D) = \frac{c(U_r, D)}{c(U, D)}.$$

Note that RAC_D does not takes into consideration any data representing users in $U \backslash U_r$.

Also in this case it is useful to represent in a concise way the RAC_D with a single number, thus we define

$$i - RAC_D(r, D) = \int_0^1 RAC_D(r, D).$$

The RAC_U and RAC_D curves represent a summarization of the privacy risks and data quality associated to the dataview D and the users represented within it.

Essentially, the Risk and Coverage curves are representations very similar to the risk-information loss maps and to the Risk-Utility maps [11], and they correlate with Statistical Disclosure Control [10]. However, our RAC curves are not limited to information loss, but they can express different measures of data utility. Moreover, they offer a more detailed representation of the behavior (w.r.t. the trade-off between privacy risk and data utility) of each situation (i.e., different dataviews or different attacks). In this way the framework for each situation enables the identification of non-risky users and non-risky data, that the organization might share, and the risky users (data) where it is necessary to apply privacy transformations before the sharing. On the contrary, Risk-Utility maps represent only the global situation varying the situations, offering a vision which can be shown also by our Data Catalog (see Section 3.4).

Note that considering several steps of sanitization over a dataview, the RAC_D must be computed considering what we want to publish without considering the privacy process, i.e., the initial dataview. Formally, $RAC_D(r, D) = \frac{c(U_r, D')}{c(U, D)}$ where D' is the resulting dataset at the n-th cycle of sanitization and D is the data dataview without any transformation.

3.4 Data Catalog

The methodology of privacy risk analysis generates a *Data Catalog*, where one can find information regarding data useful for the development of a service or set of services. In particular, given the:

- Data Format, i.e., the data needed to realize the service(s)

- Privacy Assessment Setting, i.e., the set of dataviews and the background knowledge dimensions.

The Data Catalog provides:

- Quantification of Privacy Risk, i.e., the evaluation of the empirical risk of re-identification ($i - RAC_U$)

- Quantification of Data Quality, i.e., the quality level we can achieve with private data, compared with the data quality of original data ($i - RAC_D$).

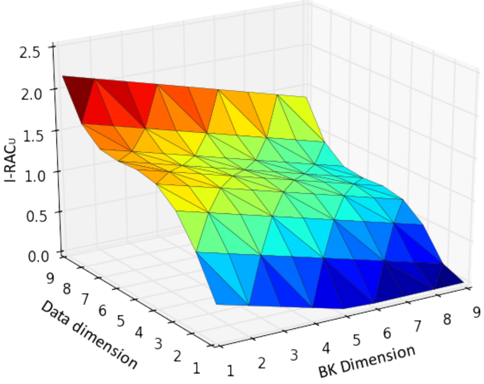

Figure 3: Data Catalog: $i - RAC_U$

The set of $i - RAC$s indexes allows representing the whole catalog as an n-dimensional plane where n is the number of (data and background) dimensions defined. In Figure 3 a representation of a 3-dimensional plane is illustrated where the red area represents a high value of $i - RAC_U$, i.e. the set of datasets where the risk of privacy for the users is lower. The data catalog represents a summary of the results of the measurements applied to the dataset by changing the different dimensions. It becomes an important tool that the DP and the SD may use to identify the most suitable dataview that can be released for developing a specific service. The catalog provides a complete view of the pros and cons related to the release of each dataview thus, since the beginning of the service development, the DP and the SD have the possibility to make a decision considering the privacy issues. Specifically, the DP is perfectly aware about the level of risk of users in the dataview and the SD has a quantative information describing the quality of data which is receiving.

4 Privacy Risk Assessment in Mobility

In this section we show three practical examples of the application of our methodology in mobility data. This is important to show that PRUDEnce is not only a theoretical tool, but it is actually applicable in real contexts. The analysis of movements is made possible by the wide range of wireless technologies able to record the position of the people, such as the satellite-enabled Global Positioning System (GPS) and the mobile phone networks. In the following we present a small overview on the common terminology used in mobility context:

Definition 5. *A point $p_i = (x_i, y_i, t_i)$ is an entity represented in a three dimensional space where x_i, y_i are the spatial coordinates and t_i is the temporal one.*

Definition 6. *A trajectory t is a temporal ordered list of spatio-temporal points $p_1 \dots p_n$ in which the user position has been observed by the device.*

Having additional information like a road network it is possible to map the points to road segments obtaining a richer representation of a trajectory as a sequence of roads traversed by the user. Other common processing of the observed points includes the discretization of the space or time using grids or time-windows. More formally, we define a grid as follows.

Definition 7. *Given a geographical area $A = rect(Ax,\ Ay, Aw, Ah)$, the grid G_m^A is the set $\{c_{(0,0)} \dots c_{(n,k)}\}$ where the squared cell $c_{(i,j)} = rect(Ax + m \times i, Ay + m \times j, m, m)$, $n = \ulcorner \frac{Aw}{m} \urcorner$, $k = \ulcorner \frac{Ah}{m} \urcorner$ and m is the length of the side of the cell.*

Here, $rect(x, y, w, h)$ represents a rectangle with the top left vertex in the coordinates (x, y), width w and height h. The resulting set of cells covers the entire geographical area with same dimension rectangles. A point is generalized with the cell which contains it. Similarly for the time-windows the period under analysis can be divided into time-slots of a specific duration (e.g. a day divided in 24 slots of 1 hour).

In this paper, to show the practical relevance of PRUDEnce, we simulated the ecosystem using a large-scale mobility dataset containing real users GPS traces. This dataset is provided by an Italian company called *OctoTelematics* collecting data for insurance purposes. This dataset is composed of GPS observations of 38,259 private cars active in Tuscany, Italy in a period of 30 days between June and July 2011 for a total of 1,382,892 trajectories. We used the dataset to study the privacy risk assessment for three different kinds of mobility data formats used for developing mobility services such as parking assistance, geo-marketing, and route prediction. Clearly, the data formats, extracted from our mobility dataset and analyzed in the following, are not exhaustive, but they are only three simple examples that allows us to validate our framework and show how different levels of aggregation and granularity in the data may affect the privacy risk. Moreover, they allow us to simulate the typical situation where the DP owns a set of data (in our examples mobility data) and receives many different requests of data from SDs that have to build different services typically unrelated among them.

4.1 Privacy Risk Assessment for Presence Data

The availability of users' presence data in specific locations enables several mobility services such as parking assistance, recommendation, alert/notification and context-aware advertising [31].

In these cases given a specific time window and a geographical area, developing the service requires to know for each user his list of most frequent locations:
$u_i : \langle l_1, f_1 \rangle, \dots, \langle l_n, f_n \rangle$ where in each pair $\langle l_j, f_j \rangle$:

- l_j is the location visited by the user; it is often represented by a cell that is a generalization of the location coordinates. This information may be sensitive because identifies the frequent user visited places, that might permit the user re-identification.

- f_j is the number of times user u_i visited the location l_j in that specific time window.

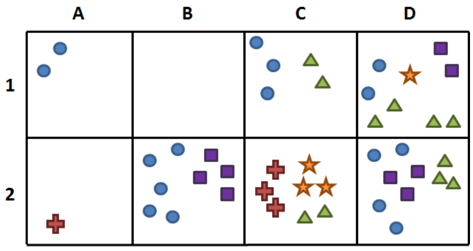

Figure 4: Example of scenario with 5 users and 8 locations

In order to generate the different dataviews starting from the location data, related to a specific time window and a geographical area, we need to define: a strategy for the spatial discretization, to get cells instead of exact locations, and the frequency threshold f for filtering possible infrequent locations. In the following, we discretize the space by a grid where the cells become the locations to be released. Therefore, in this setting we have two data dimensions: the *side size of each cell* that determines the granularity of the spatial information released about each user and the *frequency threshold* that defines a filter on the data to be distributed. Concerning the side size of the cell we considered the following values: $250m$, $500m$ and $750m$; while for the frequency threshold we set $f = 1, 4, 7, 10, 13$. Note that, given a frequency threshold f, the dataview contains only the information on locations which are frequent at least f times. After setting the values for the data dimensions we extracted from the database $3 * 5 = 15$ dataviews.

Example 8. *Consider the case in Figure 4 where there are 8 cells and 5 users (*blue, purple, green, orange *and* pink*). If we set the frequency threshold to 3 the dataview will be:*

$blue(circles) : \langle B_2, 5\rangle, \langle D_2, 4\rangle, \langle C_1, 3\rangle$
$pink(crosses) : \langle C_2, 3\rangle$
$purple(squares) : \langle B_2, 4\rangle, \langle D_2, 3\rangle$
$green(triangles) : \langle D_1, 4\rangle, \langle D_2, 3\rangle$
$orange(stars) : \langle C_2, 3\rangle$

4.1.1 Background Knowledge based attack

We assume the adversary knows a set of places visited by a specific user with a minimum number of visits in each location: $L_h : \langle l_1, mf_1\rangle, \ldots, \langle l_h, mf_h\rangle$. For example, if the attacker knows that during a working week (five days) the user u never missed his work (location l_2), then he also knows that $mf_2 = 5$.

Note that, the adversary also knows the *minimum frequency* f used to generate the dataview. Given the background knowledge L_h, to perform the attack for the user re-identification, he can use only the knowledge about the locations l_j with $mf_j \geq f$; we call it $L_s \subseteq L_h$.

Given the dataview, the set of candidates matching the background knowledge L_s is defined as $R = \{u' \in U | \forall \langle l_i, mf_i\rangle \in L_s \exists \langle l_j, f_j\rangle \in u'. \, l_i = l_j \wedge f_j \geq mf_i\}$. The probability of re-identification of the user u is $PR_D(d = u|L_s) = \frac{supp_{D_u}(L_s)}{supp_D(L_s)} = \frac{1}{|R|}$. Note that $supp_{D_u}(L_s) = 1$ since in the dataview we have a transaction for each user.

In the case of $L_s = \emptyset$, i.e., each location in L_h has frequency less than f, the set of candidates R is equal to the whole set of users U.

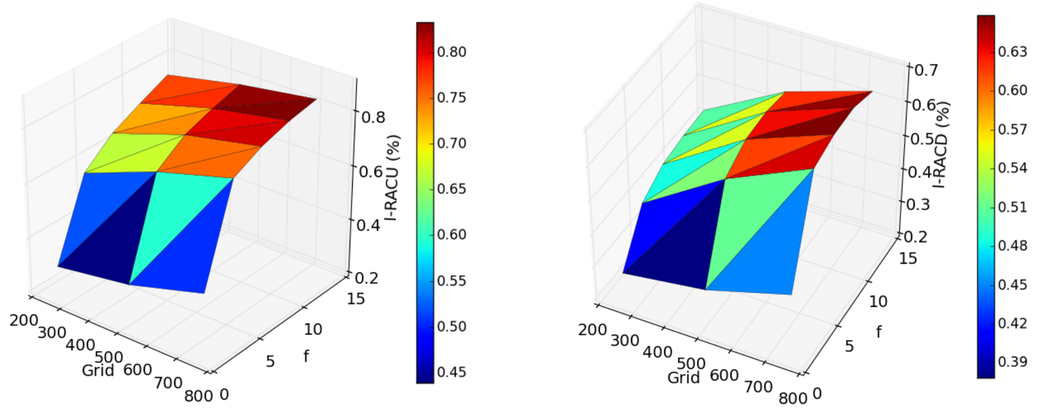

Figure 5: Data Catalog: $i{-}RAC_U$ and $i{-}RAC_D$ for each combination of locations frequency and grid size and setting the level of background knowledge to $h = 2$.

In this context, the dimensions of the background knowledge are: (a) the number of locations known by the adversary h and (b) the minimum number of visits per location. In the following, we simulate the attack with number of known locations $h = 1, 2, 3$ and with a minimum number of known visits per location equal to the exact number of real visits (i.e, 100%), the 50% of the real visits, and considering only the information about the presence of the user in the location, i.e., for each location $count(l_j) = 1$. In the following, we present an example to clarify this attack.

Example 9. *Consider the example in Figure 4, suppose that the attacker wants to discover which are the transactions related to Mr. Smith, who he saw parking 2 times in the bar located in the D_2 area. The attacker search for users who have the location D_2 and frequency greater or equal to 2. The matched users are* purple *(squares) and* green *(triangles) having frequency 3 and* blue *(circles) who has frequency equal to 4. So, in this case, the probability of re-identification is $\frac{1}{3}$.*

The setting for the background knowledge dimensions concludes the definition of the *Privacy Assessment Setting* and generates 9 different attacks to be simulated on the 15 dataviews get by setting of data dimensions. Note that, the background knowledge described above is only one of the possible that we can define for the privacy assessment.

4.1.2 Privacy Risk Analysis

For the simulation of the 9 attacks described above, starting from the GPS data, introduced in Section 4, we selected the users with *at least* a trajectory ending inside Pisa urban area. Moreover, in order to remove the "inactive" users, we filter out the users having less then 7 trajectories in one month. This pre-processing step is needed in order to have a realistic simulation of the users in an urban context (for this application). The result is a set of $247,633$ trajectories related to $3,780$ users. Supposing to create a dataset useful for developing a parking assistance service where locations are parking areas, for each trajectory, we took only its last point that represents a parking position. In this way we constructed the dataset $\mathcal{D}$, that will be used for extracting the dataviews to be analyzed.

In this case, the function $c(U', D)$ measures the number of visited locations related to users U′, so RAC_D shows the ratio between the number of locations related to safe users and the total number of visited locations. For each dataview and for each background knowledge dimension setting, the privacy assessment module computes both indexes $i - RAC_U$ and $i - RAC_D$, for the data catalog.

Both planes depicted in Figure 5 increase moving from the smallest grid to the largest one and from the lowest frequency threshold to the highest. This shape is typical for those cases where the data dimensions have a monotonic effect on the indexes. It is interesting to note how the two dimensions interact with each other in the slope of the plane: using a small grid leads to a slight increment especially for the $i - RAC_D$. This happens due the fact that, with a small grid, the frequency values in the locations are lower, boosting the effectiveness of the frequency threshold. In particular, analyzing the $i - RAC_D$ we can observe how having a frequency threshold equal to 13 or 10 is practically the same because we have the biggest impact on the data with frequency 7. On the other hand, with a larger grid this effect is less evident. Selecting points over the planes it is possible to analyze the distributions of the risk.

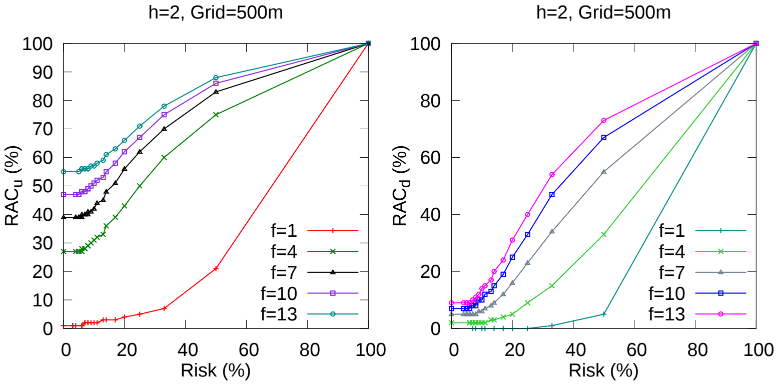

Figure 6: RAC_U and RAC_D varying the frequency and fixing background knowledge to $h = 2$ and grid size to $500m$

Setting the background knowledge to $h = 2$ and the grid size to $500m$, we obtain the distributions shown in Figure 6. They represent the RAC_U and RAC_D varying the frequency threshold value. We observe that the slopes of RAC_U are less steep than the slopes of RAC_D; this means that the risky users have more data than the others, especially when the frequency threshold is lower. Similarly in Figure 7, varying the grid size, we can see how a larger grid reduces the risk of re-identification due the fact that there are less locations in the area, and, therefore, the users are *hidden* in the crowd.

In Figure 8 we show the RAC_U and RAC_D curves that, for $h = 2$ and for $h = 3$, are very similar because the average number of locations per user is low, i.e., most of the users have only two locations which probably are the *home* and the *workplace*. The results for h values higher than 3 do not show relevant changes; for sake of readability we did not report them. Focussing on the differences between $h = 1$ and $h = 2$, it is clear that knowing two locations makes a big difference in both RAC_U and RAC_D; therefore this is a crucial point in the decision of releasing the data.

Figure 9 shows the impact of changing, in the background knowledge, the information about the minimum number of user's visits known per location. While in the previous experiments this parameter was set to 50% of the original one (*average case*), here we show what happens in the following two cases: (*worst case*) the attacker knows the exact number

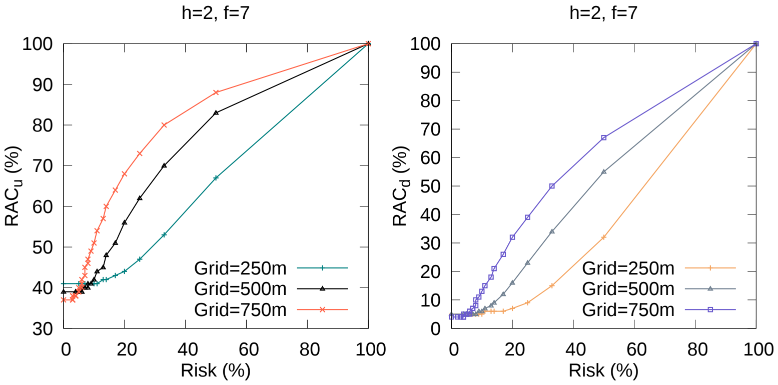

Figure 7: RAC_U and RAC_D varying the grid and fixing background knowledge to $h = 2$ and frequency to $f = 7$

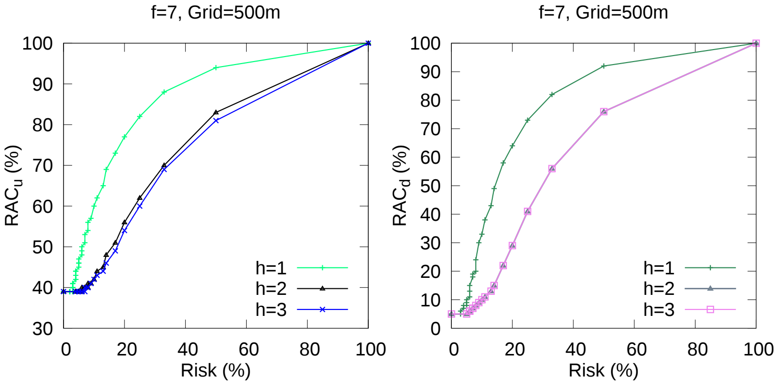

Figure 8: RAC_U and RAC_D varying the background knowledge $h = 1, 2, 3$ and fixing frequency to $f = 7$ and grid size to $500m$.

of visits, and (*best case*) he does not have any information about the user's visits. The plot clearly highlights that the absence of information about the number of visits per location leads to no risk for users because the frequency constraint makes useless the knowledge information owned by the attacker.

In conclusion, we can make a decision about which is the best dataview considering: (i) we understood that taking a value of h higher than 2 does not change greatly the results, so we can select $h = Max_{u \in U}(|u|)$; (ii) The value 7 represents a critical frequency threshold, i.e. with a higher value the risk distribution and quality slightly increase, while for lower values they have an important decrease; (iii) a large grid leads to a better risk distribution and a better data quality. Obviously, the last two considerations affect the kind of service which can be built on top of this data. For example, if we consider the parking assistance service, a large grid leads to a larger spatial granularity of each location, and the value of the frequency threshold set the minimum number of user's parkings for enabling the service in a specific location.

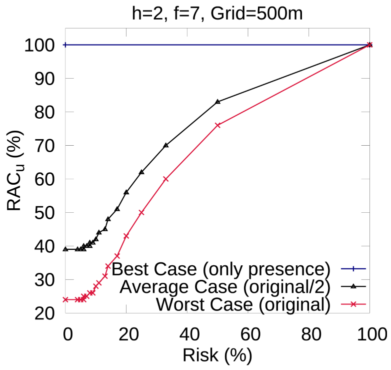

Figure 9: RAC_U for different values of minimum number of visits known per location and fixing $h = 2$, grid size to $500m$ and frequency to $f = 7$.

4.2 Privacy Risk Assessment for Trajectory Data

The distribution of trajectory data describing the movements of users are an important source of information useful for developing services, based on a mechanism of route prediction, such as navigational services, car pooling service and location recommendation services. This kind of data can be used also for supporting the traffic management thanks to different mobility analyses that enable the understanding of urban mobility: identification of access points in a urban area, the mobility atlas of a city, the identification of urban areas with high traffic volume.

For developing this kind of services, given a specific time window and a geographical area, it is necessary to obtain a set of user trajectories i.e., a dataset that for each user contains a sequence of temporally annotated road segments:

$$u_i : \langle s_1, t_{o_1}, t_{d_1}\rangle, \ldots \langle s_n, t_{o_n}, t_{d_n}\rangle.$$

Here, each element of the sequence is composed of:

- s_i that represents a road segment driven by the user starting from the origin location o_i and ending in the destination location d_i;
- t_{o_i} and t_{d_i} represent the time associated to origin and destination locations respectively.

In order to generate the different dataviews starting from the trajectory data, we need to define the *time granularity*, i.e., the precision of the times t_{o_i} and t_{d_i} associated to each link. This represents the only data dimension.

We approximated the time to 30 minutes, 60 minutes and 300 minutes obtaining 3 dataviews.

4.2.1 Background Knowledge based attack

The attack model that we consider in this case is the typical *sequence linking attack* on movement data [23]. Before explaining the details of the attack we introduce the notion of subtrajectory.

Definition 10. *Let tr* $= \langle s_1, t_{o_1}, t_{d_1}\rangle, \ldots \langle s_n, t_{o_n}, t_{d_n}\rangle$ *be a user trajectory. We say that tr′* $= \langle s'_1, t'_{o_1}, t'_{d_1}\rangle, \ldots \langle s'_m, t'_{o_m}, t'_{d_m}\rangle$ *is a sub-trajectory of tr (tr′* $\preceq$ *tr) if there exist integers* $i \leq i_1 < \ldots < i_m \leq n$ *such that* $\forall 1 \leq j \leq m$ $\langle s'_j, t'_{o_j}, t'_{d_j}\rangle = \langle s_{i_j}, t_{o_{i_j}}, t_{d_{i_j}}\rangle$.

We assume an attacker knows a sub-trajectory $tr' = \langle s'_1, t'_{o_1}, t'_{d_1}\rangle, \ldots \langle s'_m, t'_{o_m}, t'_{d_m}\rangle$ of the trajectory of some specific person $u \in U$. He can gain this information for example, by shadowing that person for some time, and could use it to re-identify that person in the released dataview and retrieve the complete trajectory.

The ability to link the published data to external information, which enables various respondents associated with the data to be re-identified is known as *linking attack model*. In relational data, linking is made possible by using a combination of attributes that can uniquely identify individuals, such as birth date and gender; these attributes are called quasi-identifiers. The remaining attributes, called *sensitive*, represent the private information, which may be disclosed by the linking attack, and thus has to be protected. In privacy-preserving data publishing techniques, such as k-anonymity, the goal is to protect personal sensitive data against this kind of attack by suppression or generalization of quasi-identifier attributes. The movement data have a sequential nature; in the case of sequential data the dichotomy of attributes into quasi-identifiers (QI) and private information (PI) does not hold any longer. Thus, in the case of spatio-temporal data a sub-trajectory can play both the role of QI and PI. In a linking attack conducted by a sub-trajectory known by the attacker the entire trajectory is the PI that is disclosed after the re-identification of the respondent, while the sub-trajectory serves as QI. So, in this case study we consider the following attack: the attacker, by using the above background knowledge, constructs the set of candidate trajectories in the released dataset containing the sub-trajectory tr' and tries to identify the whole trajectory relative to u, i.e., $R = \{tr_{u_i} \in D | tr' \preceq tr_{u_i}\}$. In the following we say that $|R|$ is the support of the trajectory tr', i.e., $supp_D(tr')$. The probability of re-identifying the user u by tr' is $PR(d = u|tr') = \frac{supp_{D_u}(tr')}{supp_D(tr')}$. Note that $supp_{D_u}(tr') = 1$, since in the released dataset we have an entry for each user.

In this context the only dimension of the background knowledge is the number of crossed segments known by the adversary h. We simulated the attack by setting the number of known segments equal to $h = 1, 2, 3, 4, 5, n$, where n is the total number of segments in the whole user trajectory.

The setting for the background knowledge dimensions concludes the definition of the *Privacy Assessment Setting* and generates 6 different attacks to be simulated on each one of 3 dataviews determined by setting the data dimension.

4.2.2 Privacy Risk Analysis

For the simulation of the 6 attacks described above, we extracted from the GPS data introduced in Section 4 the trajectories of one day between 8:00AM to 01:00PM in Pisa. Those trajectories are processed by a *map matching* procedure generating, for each user, a single sequence of road segments. This dataset will be used for extracting the dataviews to be analyzed.

In this case $c(U', r)$ quantifies the number of trajectories' segments related to users having risk at most r, so RAC_D indicates the ratio between the number of segments related to safe users and the total number of road segments. For each dataview and for each background knowledge dimension setting, the privacy assessment module computes both indexes $i - RAC_U$ and $i - RAC_D$ generating the planes depicted in Figure 10, which will be part of the data catalog. In this case we can see how the time granularity is not very effective: in all the dataviews the two indexes are very low, exception made for the combination time 300 minutes and $h = 1$. To better understand what happens at the border of the peak, we can study the $RACs$ with $h = 3$. In Figure 11 we show the curves corresponding to those points in the planes, and we observe that a slight change of the curves appears in the case of time approximated to 300 minutes.

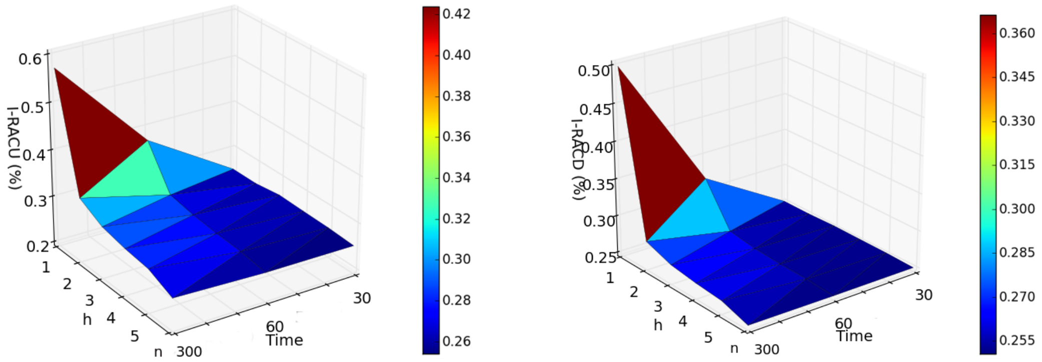

Figure 10: Data Catalog: $i - RAC_U$ and $i - RAC_D$ for each combination of time granularity and level of background knowledge.

The other border of the peak contains all the points of the plane with time 60 minutes; the relative $RACs$ are shown in Figure 12. Here, the height of the curves is lower than the previous case, but we have a change when the background knowledge is $h = 1$.

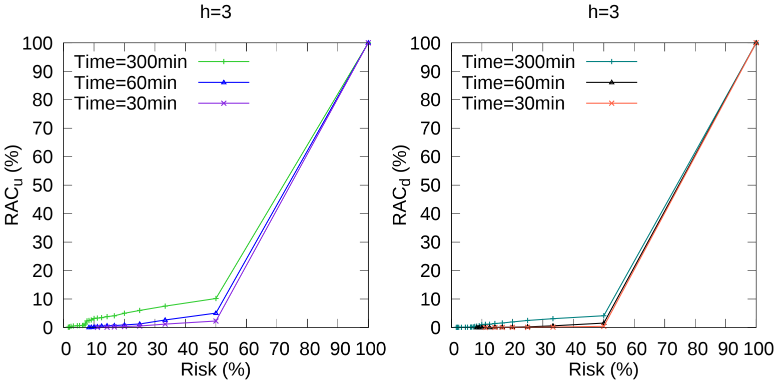

Figure 11: RAC_U and RAC_D for $h = 3$ by varying the time granularity

The analysis of the planes and the curves leads to the conclusion that this kind of data is potentially very dangerous for the users' privacy and some deep mitigation of this risk must be applied before the releasing. In literature several anonimyzation techniques for trajectory data and sequential data exist, such as [23], [34], and [2].

4.3 Privacy Risk Assessment for Road Segment Data

The knowledge about crowded road segments enables several services, ranging from the identification of strategic locations for setting up new facilities (e.g., franchise stores, gasoline/fuel stations) to navigational services optimizing the routing system to avoid congestions.

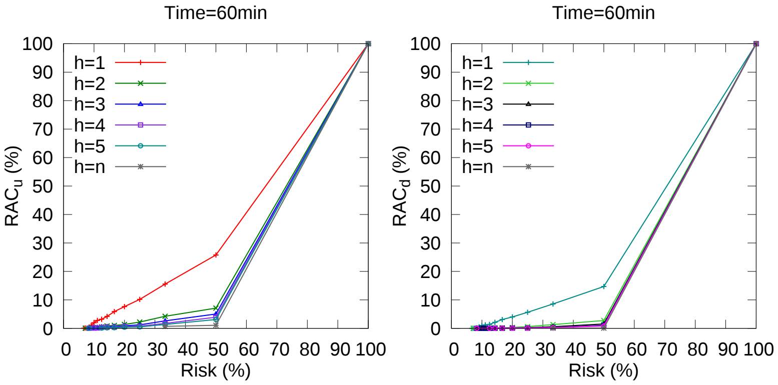

Figure 12: RAC_U and RAC_D varying the background knowledge $h = 1, 2, 3, 4, 5, n(max\ length)$ and setting the time to $60min$

For developing this kind of services it is necessary to get areas with high volume of traffic in a specific slot of the day. This can get by specific analytical processes applied on user movements related to a given period and geographical area.

Each area is represented by:

$$c_i : tw, \{s_1, s_2, \ldots, s_p\}, \mu$$

where $\{s_1, s_2, \ldots, s_p\}$ is a set of road segments characterized by a volume of traffic described by the index μ in the time-window tw.

The analytical process to extract the above areas from user movements is composed by three steps: i) using a frequency threshold f, we select only segments with a high volume of traffic, i.e., segments crossed by a high number of user in the specific time-window tw; ii) a density based clustering [3] is applied using a spatial tolerance ϵ; iii) for each cluster of segments, representing an area, the traffic index μ is computed as average of the segments frequency.

In order to generate different dataviews we exploit three data dimensions: the *temporal dimension*, that defines the temporal granularity of the time-window, the *frequency threshold*, that defines a filter on the segments to be distributed, and the *spatial tolerance* of the clustering, that affects the clusters composition. Concerning the first dimension, in our experiment we set $tw = 1\ hour$ and $\epsilon = 50m, 100m, 150m, 200m$, while for the frequency threshold we set $f = 200, 250, 300, 350$. The number of the resulting dataviews is 16.

In the following we present a simple example that highlights the main steps to get the clusters of links to be distributed.

Example 11. *Suppose to apply the analytical process described above on a dataset of trajectories. The first step is to split the trajectories in two time slots: 09-10 AM and 10-11 AM getting the following trajectories depicted in Figure 13 (top):*

09:00AM-10:00AM	*10:00AM-11:00AM*
$u_1 : A, B, E$	$u_5 : B, F, P, Q, V$
$u_2 : A, B, C, D$	$u_6 : F, G$
$u_3 : M, N, O, H$	$u_7 : P, S$
$u_4 : H, I, L, M$	$u_8 : P, S, T$
	$u_9 : P, S, R, V$

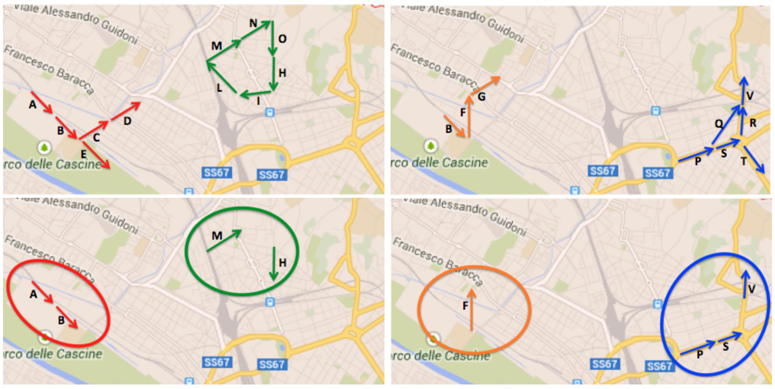

Figure 13: Left: trajectories and clusters in the time slot 09:00AM-10:00AM; Right: trajectories and clusters in the time slot 10:00AM-11:00AM.

Note that above we are indicating each road segment of a trajectory by a letter. If we set the minimum frequency to 2*, we get the clusters in Figure 13 (bottom) for each time window. In particular, in the time-window we obtain the clusters:*

$green,\ 09:00AM-10:00AM,\ \{M,H\},2$
$red,\ 09:00AM-10:00AM,\ \{A,B\},2$
$orange,\ 10:00AM-11:00AM,\ \{F\},2$
$blue,\ 10:00AM-11:00AM,\ \{P,S,V\},3$

4.3.1 Background Knowledge based attack

In this case, the dataview does not contain data directly linkable to a single user but models representing information on groups of road segments visited by several users. Here, each user movement may contribute to several clusters.

The attacker, if he is capable to identify the set of models representing the user movements, may understand his target's habits and infer sensitive information about him.

We highlight that the adversary, knowing only a portion of the user road segments, cannot identify all clusters crossed by that user because he could have crossed an unmatched cluster with the portion of trajectory that the attacker does not know. On the other hand, knowing all user road segments the attacker has already the complete knowledge about the user mobility and cannot gain any new information.

Another external knowledge that an adversary may use in his attack is the period of time in which the user moved. This can help him to reduce the number of candidate clusters. Therefore, we assume that given a user u the attacker knows the time-slots of all his movements $BK = t_1, t_2, \ldots, t_m$.

Let Cl_{BK} be the set of released clusters covering temporally the time-slots in BK and let $Cl_{BK,u}$ be the subset of these clusters crossed by the user u. We define a function $W(.)$ that takes a set of clusters with their average frequency and computes the total frequency $W(Cl_{BK}) = \sum_{c_i \in Cl_{BK}} \mu_{c_i}$.

Therefore, for each user u, we compute the probability of re-identification as the probability to associate a cluster to the user u $PR(d = u|BK) = \frac{W(Cl_{BK,u})}{W(Cl_{BK})}$.

In this setting the *Privacy Assessment* simulates the described attack on the 16 dataviews.

Example 12. *Continue with Example 11. Assuming that the attacker knows that Mr.Smith (u_5) is represented in the released data and that he travelled only in the time slot 10:00AM-11:00AM. The risk of re-identifying the cluster crossed by u_5 is equal to 1 because with his trajectory he crossed both the blue and orange cluster. Instead, if we consider the user u_1, assuming that the attacker knows that he traveled only in the time slot 09:00AM-10:00AM, then the risk is equal to* $\frac{\mu_{red}}{\mu_{red}+\mu_{green}} = \frac{1}{2}$.

4.3.2 Privacy Risk Analysis

In this case study we used the whole dataset described in Section 4 to simulate the privacy risk assessment methodology. The first step is the application of an analytical process based on the computation of a set of clusters of frequent road segments. Figure 14 shows the clusters distribution in different hours of the day by varying the minimum frequency threshold required for the road segments (left) and the spatial tolerance of the clustering ϵ (right).

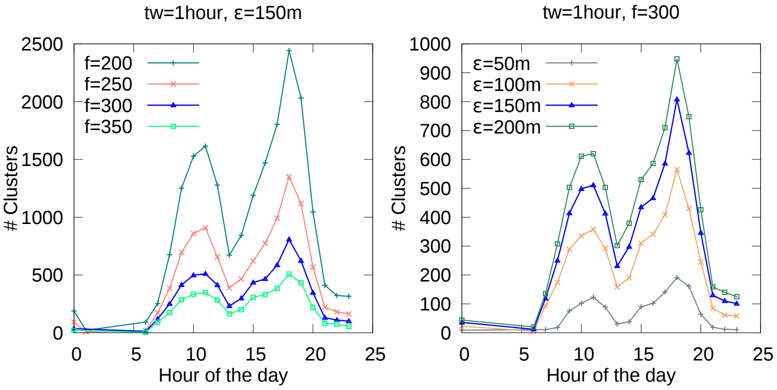

Figure 14: Left: clusters distribution during the day by varying the minimum frequency threshold for the road segments; Right: clusters distribution during the day by varying the spatial tolerance of the clustering ϵ .

We can observe how both distributions have the typical *two peaks* shape well-known in the traffic management field, i.e. the morning rush hour and the evening one.

In this case, the function $c(U', r)$ evaluates the coverage of extracted clusters, i.e., the number of segments that are inside the cluster; here, RAC_D indicates the ratio between the coverage of cluster computed after the remotion of unsafe users and the coverage of cluster computed on the whole dataset. It is worth noting that in this case removing single users might not directly affect the results of this function, since each segment might remain frequent even if some users are no longer present. Starting from the dataset generated by this analytical process, the risk assessment module computes both indexes $i - RAC_U$ and $i - RAC_D$ for each combination of spatial tolerance ϵ and minimum frequency threshold. The two resulting planes are depicted in Figure 15. The $i - RAC_U$ shows almost a constant increase going from the configuration ($\epsilon = 200, f = 350$) to ($\epsilon = 50, f = 200$) suggesting an inverse proportionality to the two data dimensions. More interesting is the $i - RAC_D$ presenting an area with high values surrounded by valleys. This means that a slight change of data dimensions values may lead to a reduction of the data quality.

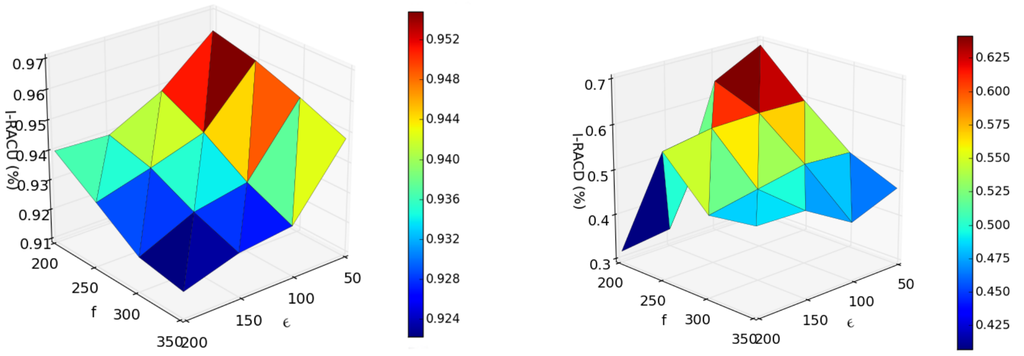

Figure 15: Data Catalog: $i - RAC_U$ and $i - RAC_D$ for a dataset created with time-window 1 hour.

Figure 16 shows the RAC_U and RAC_D curves by setting the minimum frequency threshold for the road segments to $f = 300$ and varying the spatial threshold used by the clustering algorithm. In this case the aggregation of the data leads to very low risk for the users (notice that the $i - RAC_U$ varies between 0.89 to 0.97). The shape of RAC_D is very different from other cases. This is due the fact we are managing patterns and not single user data. Looking at the curves we note how, for all ϵ values, a vertical increase occurs around 45% of risk generating a *jump* in RAC_D value from 15% to 90%; i.e. if we require a privacy guarantee with a risk lower than 45%, the data quality drops.

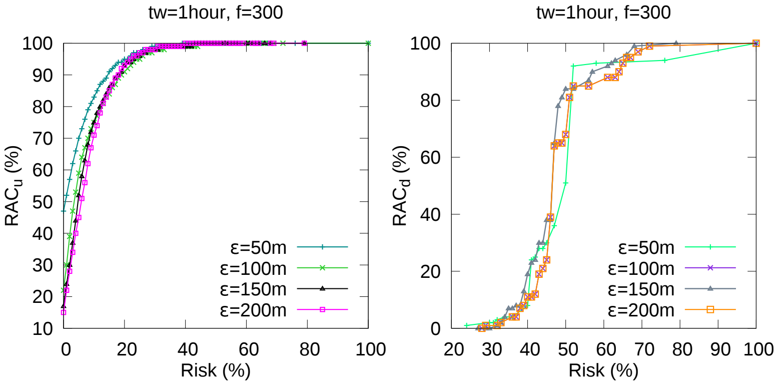

Figure 16: RAC_U and RAC_D by varying the spatial tolerance ϵ and setting the frequency to $f = 300$

Moreover, considering the points on the plane having $\epsilon = 150m$, we have the $RACs$ depicted in Figure 17 showing how the frequency changes the risk distribution. For the RAC_U we observe that different values of frequency produce small changes in the curves. On the contrary, the RAC_D shows that using a lower frequency threshold we can move the vertical increase to lower value of risk. This guarantees more privacy without destroying the data.

In conclusion, looking at the data catalog we understood that data dimensions do not impact the number of risky users, therefore the attention should be focused on the data quality. The data dimension ϵ has a limited effect on $i - RAC_D$, which is really sensitive to

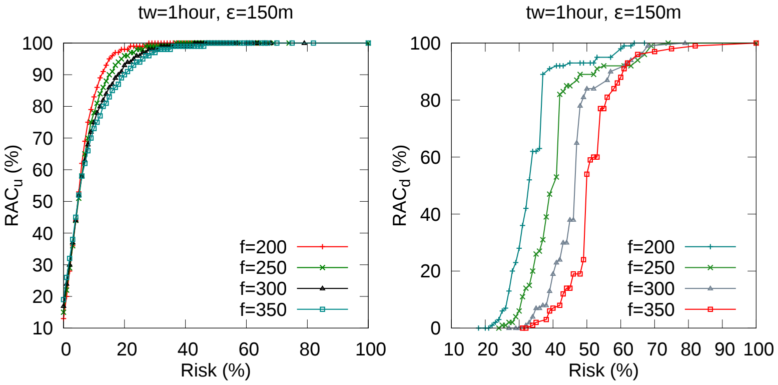

Figure 17: RAC_U and RAC_D by varying frequency of each road segment and setting the spatial tolerance to $\epsilon = 100m$.

the frequency threshold.

5 Towards Distributed Computation

In addition to utilization and privacy risk, computational aspects is also a crucial part of the assessment procedure; it would be difficult to justify its usefulness for actual usage if it took days or weeks to complete the computation. The key issue here is how to distribute the computation of the RAC curves as the number of users in the dataview scales up. This task boils down to the distributed computation of histograms based on the computation of the privacy risk for each user. A possible natural implementation of this task using the Map-Reduce paradigm, where each node compute in parallel the risk of a set of users. Given the dataview D and its partitioning $\{P_1 \dots P_n\}$, each map node is assigned to a partition P_j and the *map* step enumerates all the possible adversary background knowledge $\{bk_1^i \dots bk_m^i\}$ relative to each user u_i, counts the occurrences of each bk while maintaining the set of users sharing the same bk. Then, by using bk as key for the *shuffling* step, the system groups all users having in common the vulnerability against bk. The *reduce* step finally computes the maximum probability of re-identification of each user given the shared background knowledge and uses such information for constructing the RAC curves. Considering the cost of centralized version as $O(n)$, where n is the number of user, we can say that the theoretical lower bound of the distributed version is $O(n/k)$, where k is the number of nodes used to execute the process in parallel. Clearly, other computational costs such as overhead in the distribution of work and network communications may reduce the gain in a real application. In Figure 18 we show the runtime analysis of the distributed version and compare the results with the theoretical lower bound. We tested our implementation for the case study presented in Section 4.1. However, similar optimizations can be achieved also for the data formats used in the other two case studies. Results show that the execution time is very close to the theoretical lower bound highlighting that the overheads have a small effect on the runtime. We ran this test increasing the number of nodes (i.e. from 1, which is the centralized version, to 7) and we can see that the process scale up smoothly.

6 Releasing multiple Dataviews

In this section we discuss the impact on privacy protection when multiple dataviews are released by the DP. The problem of managing the risk in this particular context is already studied in the literature related to the compositional attack [30, 4], to the incremental publication of datasets from an organization [40] and to multiple views analysis [41]. To the best of our knowledge, all the existing works are based on the presence of sensitive attribute, and, at least in the case of compositional attack, on the possibility of discover sensitive information exploiting the uniqueness of quasi-identifiers (QI) and private information (PI) in different datasets. Here, the problem is to check if the combination of datasets, that singularly are safe from a privacy perspective, leads to some possible inferences or privacy threats. In brief, they are based on the unicity of a PI among the possible values given by the QI in two or more datasets. As already said, in case of mobility data we do not have the distinction between QI and PI, therefore given a released dataview, with the attacks defined in Section 4, we already tested the linkability and the protection level for all the possible combinations of locations. In other words, given a dataview, if the risk assessment provides an evaluation that guarantees a level x of privacy, we ensure that this protection is guaranteed also for the potential PI.

However, our framework is able to manage a more general situation without doing assumption on the type of data. The universal solution is based on keeping track of the previously released dataviews, in order to perform the risk assessment (and the mitigation process) incrementally. In practice, if the DP released a certain dataview (D_1) to a certain SD and, then, DP receives a new query from the same SD, instead of assess the privacy risk only on the plain dataview (D_2), DP tests the combination of the two dataviews ($D_1 + D_2$). The key point here is the correct definition of the background knowledge and, consequently, the cautious definition of attacks. This solution has the price of keeping track of data released to each SD, a solution considered reasonable in literature [40, 41], eventually considering also datasets released by other data holders [4].

Considering different kind of data, where there is a distinction between QI and PI, it would be interesting to extend PRUDEnce analyzing the assessment of privacy risk when datasets are independently released by different DPs, i.e., each DP does not have any information about the other anonymized releases, like in the work [14].

7 Related Work

One of the most related work is the LINDDUN methodology [9], where the authors introduce a privacy-aware threat analysis framework based on Microsoft's STRIDE methodology [33] capable to model privacy threats in software-based systems. LINDDUN methodology lacks a quantitative approach for privacy evaluation, as that one presented in this paper. Therefore from this perspective both approaches can be considered complementary.

In the literature, different techniques for risk management has been proposed in the last years such as the OWASP's Risk Rating Methodology[5], Microsoft's DREAD [22], NIST's Special Publication 800-30[6], and SEI's OCTAVE[7]. Many of them does not consider in deep privacy including it only when assessing the impact of a threat.

[5] https://goo.gl/Df98k1

[6] https://goo.gl/3142Mb

[7] http://www.cert.org/octave/

A proposal that quantitatively tries to manage privacy risks was presented by Trabelsi et al. [36], where an entropy-based method is elaborated to evaluate the disclosure risk of personal data. Other works in the literature study the re-identification risk as privacy measure. Hay et al. [17] propose three models of external information used by an adversary to attack naively-anonymized networks. They take into consideration attacks based on the structural knowledge of network data. In the context of online social networks Liu and Terzi [21] propose a framework for computing privacy scores for each user in the network. Such scores indicate the potential risk caused by her participation in the network. Dankar and Emam [8] introduce a re-identification risk metric that measures the proportion of records that are correctly re-identified in a dataset by an adversary, based on the idea that he wishes to re-identify as many records as possible in the disclosed database. Lastly, Ferro et al. [13] propose a methodology for assessing the vulnerability of individuals in a pre-released and pre-anonymized dataset who may be re-identified by using public data.

The main difference with our paper is that all these works do not provide any systematic way to perform a complete privacy assessment for the data. They do not give to the data provider the possibility to understand if there is some possible data aggregation or data generalization that could help the reduction of privacy risks maintaining data useful for the development of some service.

Another work relevant for our purposes is the ARX tool [26, 27], which enables the analysis of the risk of re-identification and the application of various privacy paradigms, very similarly to our PRUDEnce. Even the iterative method that authors applied in [27] is quite similar, with the difference that we want to provide the quantification of the data quality not after but along with the assessment of the privacy risk, in order to be verified before the application of the chosen mitigation strategies. However, the main difference is that ARX tool is based on the existence of typical quasi-identifiers, while it cannot be easily applied to complex high-dimensional identifiers [27], as in the case of mobility data.

In the community of statistical disclosure control some works address the problem of measuring the disclosure risk. Some of them consider the risk of re-identification based on a perfect matching like us. Others instead propose approaches solving the problem of partial matching: probabilistic record linkage [18, 39] and distance based record linkage [1, 35]. The goal of probabilistic record linkage (opposite to the deterministic record linkage) is to establish with a certain probability whether pairs of records from different sources either correspond to the same individual or different individuals. In particular, in [18], Jaro describes the EM (ExpectationMaximisation) algorithm for parameter estimation; this algorithm aims to estimate the maximum likelihood of unknown parameters if some data are missing. Winkler [39] improves the previous algorithm considering additional properties, such as allowing a generic log-linear model and convex constraints. The distance-based record linkage consists of computing distances between records in two different sets of data; the pairs of records at the minimum distance are considered linked pairs. The effectiveness of this record linkage algorithm heavily relies on the effectiveness of the distance function. In case of probabilistic and distance-based record linkage, some works use machine learning algorithms for identifying the worst-case scenario corresponding to the case with the largest number of re-identifications. Typically, machine learning helps to find for example the best set of parameters so that the number of re-identifications is maximum [35]. The inclusion of the partial matching in the PRUDEnce framework might make our method more general. Thus, it would be interesting to extend our framework including also these approaches.

The same community also proposes some methods for reasoning on the balancing between privacy risk and data utility by using Risk-Utility maps [11, 10], that provide the

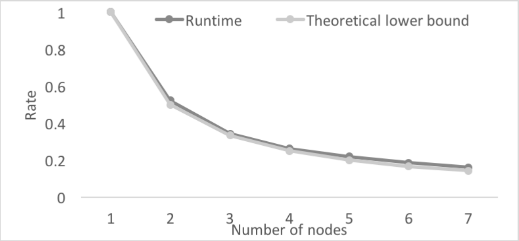

Figure 18: The comparison between theoretical lower bound and the actual runtime using a distributed version w.r.t. the number of nodes used.

trend of the risk-utility curve changing for example the protection strategy. Unfortunately, these approaches do not enable the reasoning by changing different background knowledge, attacks and dataviews. Therefore, these maps enable only a partial vision with respect our Data Catalog (Section 3.4).

Other kind of works related to ours are those introducing techniques for privacy-preserving publishing and analysis especially in the context of spatio-temporal data. In [15] authors discuss the problem of privacy issues in mobility data and provide an overview of techniques for trajectory data publishing. There have been recent work that use generalization/suppression techniques. The mostly widely used privacy model of these work is adapted from what so called k-anonymity[29], which requires that an individual should not be identifiable from a group of size smaller than k based on their quasi-identifiers. [2] proposes the (k, δ)-anonymity model that exploits the inherent uncertainty of the moving object's whereabouts, where δ represents possible location imprecision. [34] assumes that different adversaries own different, disjoint parts of the trajectories. Authors use *suppression* of the dangerous observations from each trajectory. [42] considers timestamps as the quasi-identifiers, and define a method based on *k-anonymity* to defend against an attack called *attack graphs*. [23] proposes a spatial generalization approach to achieve k-anonymity.

Other works instead use the well-know differential privacy model [12]. [24] considered a distributed aggregation framework for movement data and proposed the application of ϵ-differential privacy model to guarantee individual privacy. Chen et al. [7] propose to release a prefix tree of trajectories with injected Laplace noise. However, these approaches use the standard definition of differential privacy that is based on the global sensitivity computation which does not consider the actual data to be protected, leading to the risk of damage data without any need. Recently, some promising variants of the original differential privacy have been proposed like individual differential privacy [32] and the smooth sensitivity based approach [25].

8 Conclusion

In this paper we proposed PRUDEnce, a system for assessing privacy risk and quality of large scale data sets. PRUDEnce helps realizing a privacy-aware ecosystem to share personal data to and from different organizations of the industry and academia by allowing quantitative measuring of the risk of re-identification and data quality. The system aims at assisting companies in avoiding negative impacts, by enabling the sharing of privacy protected data with a reduced risk. As a consequence, these companies will be able to use personal data to create new products and services and improve the customer experience, share data between different departments or business units to generate additional insights, efficiencies and opportunities.

We validated our proposal in the context of mobility data, specifically on three different examples of data formats, i.e., presence data, trajectory data, and road segments, that are typically used for several specific use cases of mobility services. In particular, we showed how this kind of data can be analyzed by PRUDEnce to give meaningful measures of the users' risk and the data quality varying the privacy guarantee.

As future work we will extend the framework including the measurement of the quality of a service developed on top of the released data for a deeper analysis of the impact of a possible data anonymization over the final service quality. Moreover, it would be interesting to integrate in PRUDEnce re-identification techniques which are not based only on perfect matching like probabilistic and distance based record linkage approaches [35].

Acknowledgment

This work has been partially supported by Project EU H2020-654024 SoBigData Infrastructure.
In a previous version of this work [28], the presented framework was named PRISQUIT (a system for assessing Privacy RISk versus QUalITy in data sharing).

References

[1] Daniel Abril, Guillermo Navarro-Arribas, and Vicenç Torra. Choquet integral for record linkage. *Annals OR*, 195:97–110, 2012.

[2] Osman Abul, Francesco Bonchi, and Mirco Nanni. Never walk alone: Uncertainty for anonymity in moving objects databases. In *International Council for Open and Distance Education (ICDE)*, pages 376–385, 2008.

[3] Mihael Ankerst, Markus M. Breunig, Hans-Peter Kriegel, and Jörg Sander. OPTICS: ordering points to identify the clustering structure. In *SIGMOD 1999*, pages 49–60, 1999.

[4] Muzammil M. Baig, Jiuyong Li, Jixue Liu, Xiaofeng Ding, and Hua Wang. *Data Privacy against Composition Attack*, pages 320–334. Springer Berlin Heidelberg, Berlin, Heidelberg, 2012.

[5] Jane Bambauer, Krishnamurty Muralidhar, and Rathindra Sarathy. Fool's gold: an illustrated critique of differential privacy. *Vanderbilt Journal of Entertainment & Technology Law*, 16:701, 2013.

[6] Ann Cavoukian. Privacy design principles for an integrated justice system - working paper. 2000. https://www.ipc.on.ca/english/resources/discussion-papers/discussion-papers-summary/?id=318.

[7] Rui Chen, Benjamin C. M. Fung, Bipin C. Desai, and Nériah M. Sossou. Differentially private transit data publication: a case study on the Montreal transportation system. In *International Conference on Knowledge Discovery and Data Mining (KDD)*, pages 213–221, 2012.

[8] Fida Kamal Dankar and Khaled El Emam. A method for evaluating marketer re-identification risk. In *Proceedings of the 2010 International Conference on Database Theory (EDBT/ICDT) Workshops*, 2010.

[9] Mina Deng, Kim Wuyts, Riccardo Scandariato, Bart Preneel, and Wouter Joosen. A privacy threat analysis framework: supporting the elicitation and fulfillment of privacy requirements. *Requirements Engineering*, 16(1):3–32, 2011.

[10] Josep Domingo-ferrer, Josep M. Mateo-sanz, and Vicenç Torra. Comparing SDC methods for microdata on the basis of information loss and disclosure. In *Proceedings of Exchange of Tech-*

nology and Know-how - New Techniques and Technologies for Statistics (ETK-NTTS), pages 807–826. Eurostat, 2001.

[11] George T Duncan and S Lynne Stokes. Disclosure risk vs. data utility: The RU confidentiality map as applied to topcoding. *Chance*, 17(3):16–20, 2004.

[12] Cynthia Dwork, Frank McSherry, Kobbi Nissim, and Adam Smith. Calibrating noise to sensitivity in private data analysis. In *Theory of Cryptography Conference (TCC)*, pages 265–284, 2006.

[13] John Ferro, Lisa Singh, and Micah Sherr. Identifying individual vulnerability based on public data. In *International Conference on Privacy, Security and Trust*, pages 119–126, 2013.

[14] Srivatsava Ranjit Ganta, Shiva Prasad Kasiviswanathan, and Adam Smith. Composition attacks and auxiliary information in data privacy. In *Proceedings of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD)*, pages 265–273. ACM, 2008.

[15] Fosca Giannotti, Anna Monreale, and Dino Pedreschi. Mobility data and privacy. In *C. Renso, S. Spaccapietra, E. Zimanyi editors, Mobility Data Modeling, Management, and Understanding*, pages 174–193, 2013.

[16] Andreas Haeberlen, Benjamin C Pierce, and Arjun Narayan. Differential privacy under fire. In *USENIX Security Symposium*, 2011.

[17] Michael Hay, Gerome Miklau, David D. Jensen, Donald F. Towsley, and Philipp Weis. Resisting structural re-identification in anonymized social networks. *Proceedings of the Very Large DataBases PVLDB*, 1(1):102–114, 2008.

[18] Matthew A Jaro. Advances in record-linkage methodology as applied to matching the 1985 census of tampa, florida. *Journal of the American Statistical Association*, 84(406):414–420, 1989.

[19] Diane Lambert. Measures of disclosure risks and harm. *Journal of Official Statistics*, 9(2):313–331, 1993.

[20] Jaewoo Lee and Chris Clifton. How much is enough? choosing ε for differential privacy. In *Information Security*, pages 325–340. Springer, 2011.

[21] Kun Liu and Evimaria Terzi. A framework for computing the privacy scores of users in online social networks. *Transactions on Knowledge Discovery from Data (TKDD)*, 5(1):6:1–6:30, 2010.

[22] J.D. Meier and Microsoft Corporation. *Improving Web Application Security: Threats and Countermeasures*. Patterns & practices. Microsoft, 2003.

[23] Anna Monreale, Gennady L. Andrienko, Natalia V. Andrienko, Fosca Giannotti, Dino Pedreschi, Salvatore Rinzivillo, and Stefan Wrobel. Movement data anonymity through generalization. *Transactions on Data Privacy (TDP)*, 3(2):91–121, 2010.

[24] Anna Monreale, Salvatore Rinzivillo, Francesca Pratesi, Fosca Giannotti, and Dino Pedreschi. Privacy-by-design in big data analytics and social mining. *EPJ Data Science*, 3(1):10, 2014.

[25] Kobbi Nissim, Sofya Raskhodnikova, and Adam D. Smith. Smooth sensitivity and sampling in private data analysis. In *Proceedings of the 39th Annual ACM Symposium on Theory of Computing, San Diego, California, USA, June 11-13, 2007*, pages 75–84, 2007.

[26] Fabian Prasser, Raffael Bild, Johanna Eicher, Helmut Spengler, Florian Kohlmayer, and Klaus A Kuhn. Lightning: Utility-driven anonymization of high-dimensional data. *Transactions on Data Privacy*, 9(2):161–185, 2016.

[27] Fabian Prasser, Florian Kohlmayer, Helmut Spengler, and Klaus A Kuhn. A scalable and pragmatic method for the safe sharing of high-quality health data. *IEEE Journal of Biomedical and Health Informatics*, PP(99), 2017.

[28] Francesca Pratesi, Anna Monreale, Roberto Trasarti, Fosca Giannotti, Dino Pedreschi, and Tadashi Yanagihara. Prisquit: a system for assessing privacy risk versus quality in data sharing. Techical Report 2016-TR-043, ISTI - CNR, Pisa, Italy. FriNov20162291.

[29] Pierangela Samarati and Latanya Sweeney. Generalizing data to provide anonymity when disclosing information (abstract). In *Principles of Database Systems (PODS)*, page 188, 1998.

[30] A. H. M. Sarowar Sattar, Jiuyong Li, Jixue Liu, Raymond Heatherly, and Bradley Malin. A probabilistic approach to mitigate composition attacks on privacy in non-coordinated environments. *Knowledge-Based Systems*, 67:361–372, 2014.

[31] Jochen H. Schiller and Agnès Voisard, editors. *Location-Based Services*. Morgan Kaufmann, 2004.

[32] J. Soria-Comas, J. Domingo-Ferrer, D. Sanchez, and D. Megias. Individual differential privacy: A utility-preserving formulation of differential privacy guarantees. *IEEE Transactions on Information Forensics and Security*, 12(6):1418–1429, June 2017.

[33] Frank Swiderski and Window Snyder. *Threat Modeling*. O'Reilly Media, 2004.

[34] Manolis Terrovitis and Nikos Mamoulis. Privacy preservation in the publication of trajectories. In *Mobility Data Mining (MDM)*, pages 65–72, 2008.

[35] Vicenç Torra. *Data Privacy: Foundations, New Developments and the Big Data Challenge*. 28. Springer International Publishing, 1 edition, 2017.

[36] Slim Trabelsi, Vincent Salzgeber, Michele Bezzi, and Gilles Montagnon. Data disclosure risk evaluation. In *CRiSIS '09*, pages 35–72, 2009.

[37] Traian Marius Truta, Farshad Fotouhi, and Daniel C. Barth-Jones. Disclosure risk measures for microdata. In *Proceedings of the 15th International Conference on Scientific and Statistical Database Management (SSDBM 2003), 9-11 July 2003, Cambridge, MA, USA*, pages 15–22, 2003.

[38] Traian Marius Truta, Farshad Fotouhi, and Daniel C. Barth-Jones. Disclosure risk measures for the sampling disclosure control method. In *Proceedings of the 2004 ACM Symposium on Applied Computing (SAC), Nicosia, Cyprus, March 14-17, 2004*, pages 301–306, 2004.

[39] William E. Winkler. *Probabilistic linkage*. Wiley Series in Probability and Statistics, 2016.

[40] Xiaokui Xiao and Yufei Tao. M-invariance: towards privacy preserving re-publication of dynamic datasets. In *Proceedings of the ACM SIGMOD International Conference on Management of Data 2007*, pages 689–700, 2007.

[41] Chao Yao, Xiaoyang Sean Wang, and Sushil Jajodia. Checking for k-anonymity violation by views. In *31st International Conference on Very Large Data Bases*, pages 910–921, 2005.

[42] Roman Yarovoy, Francesco Bonchi, Laks V. S. Lakshmanan, and Wendy H. Wang. Anonymizing moving objects: how to hide a mob in a crowd? In *International Conference on Extending Database Technology (EDBT)*, pages 72–83, 2009.