Data Privacy*

Vicenç Torra[1] and Guillermo Navarro-Arribas[2]

[1] IIIA, Institut d'Investigació en Intel.ligència Artificial - CSIC,
Consejo Superior de Investigaciones Científicas.
[2]Dep. of Information and Communications Engineering, UAB,
Universitat Autònoma de Barcelona

Abstract

Data privacy studies methods, tools, and theory to avoid the disclosure of sensitive information. Its origin is in Statistics with the goal to ensure the confidentiality of data gathered from census and questionnaires. The topic was latter introduced in computer science and more particularly in data mining, where due to the large amount of data currently available, has attracted the interest of researchers, practitioners, and companies. In this paper we will review the main topics related to data privacy, and privacy-enhancing technologies.

1 Introduction

Data privacy and privacy-enhancing technologies study techniques and tools to avoid the unintentional disclosure of sensitive information. They have been studied in the areas of computer science and statistics. Statistical Disclosure Control (SDC) was developed first, to solve the needs of statistical offices to publish data from census and questionnaires avoiding confidentiality problems. Within computer science, tools for data privacy have been developed in relation to communications, security, databases, and data mining. Tools and methods related to communications and security are often classified as privacy-enhancing technologies (PET) while the ones related to data mining are studied in privacy-preserving data mining (PPDM). While there exist these different communities focusing on different types of applications and data uses, background and concepts, as well as some methods are common.

In this overview we will discuss the main concepts and tools in data privacy, giving a general perspective of the field and presenting them independently of

*Postprint of: Vicenç Torra, Guillermo Navarro-Arribas, Data Privacy, *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, Volume 4, Issue 4, 2014, Pages: 269-340, ISSN:1942-4795. https://doi.org/10.1002/widm.1129

the community in which they were originated. This is a broad and relatively non-technical description intended for readers without a strong background in the field. We provide, through the paper, several references which should allow the interested reader to get a deeper understanding in specific topics of the field.

The structure of the paper is as follows. First, we will present a classification of the methods for data privacy. We will review different dimensions. In particular, we will see that one of the dimensions is about the subjects involved in the data privacy process: the respondent, the owner, and the user. We will focus then on user-privacy, methods to be implemented and used by the users of a system to ensure their own privacy. Then, we will focus on respondent and owner privacy. The paper finishes with some conclusions and some references for further study.

2 Classification

The literature presents different classifications of the methods for data privacy [3, 21, 23, 34, 55, 84, 90, 93]. In this section we review three of these classifications. We will use them as different dimensions to classify and review the main methods. The three dimensions are as follows.

- **On whose privacy is being sought.** In the whole process of data collection, data protection, and data analysis several subjects (individuals or entities) are involved. This dimension focuses on the subject whose privacy is considered the main motivation for the application of a method. Three subjects are considered: respondent, owner, and user.
- **On the computations to be done.** Data are protected for a certain use. That is, for the application of a certain algorithm or to do some type of analysis. In this dimension, methods are distinguished according to the type of computation or analysis a data miner (or another user) will perform with the protected data. For example, before data protection we may know that clustering algorithms will be applied to the data, and this can help on the selection of an appropriate method for data protection.
- **On the number of data sources.** Data have to be protected, and then used. It is different when a single data set is considered and when a collection of data sets are considered. This dimension focuses on the number of such data sets.

We discuss these dimensions in more detail in the sections that follow. Figure 1 outlines the classification.

2.1 On whose privacy is being sought

As enumerated above, in the data privacy protection process we typically consider three subjects. They are the respondent, the owner, and the user. We describe their specific meaning below.

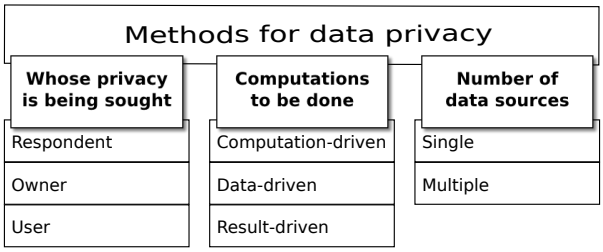

Figure 1: Classification of methods for data privacy based on three dimensions.

- **The respondent**. Following the terminology in statistical disclosure control, the respondent is the person whose data have been collected and are in the database.
- **The owner**. Data from respondents are collected and stored by a company or an administration who is the holder of the data, liable for disclosure of confidential information, and possibly with economic interest on the data. This is the owner of the data.
- **The user**. When a person accesses a system (e.g., a database, a search engine, or an email system) some trails are left which can be recorded in a secondary database or sniffed and used or disclosed later on. From a certain point of view, the user is just a respondent of this secondary database. Nevertheless, we distinguish the user from a respondent when the user can perform some actions against the system or sniffers. That is, the user corresponds to an individual that can act to avoid disclosure of his own information, i.e., an active subject, while a respondent is a passive subject almost a strange to his own data.

We can consider data privacy focusing on the three subjects discussed above. That is, we can consider respondent privacy, owner privacy, and user privacy. This dimension is based on the one in [19].

- **Respondent privacy.** It focuses on technologies that avoid the disclosure of sensitive information about the respondents of a database. Some specific goals of respondent privacy are to avoid that data are linked to a particular individual, to avoid increasing the knowledge about particular individuals, and to avoid that someone can find that a particular individuals' data is in a database.
- **Owner privacy.** It focuses on tools to avoid the disclosure of information that is relevant to the owner of the database. On the one hand this information can be information on particular individuals, as in respondent privacy. On the other hand, this information can be knowledge that can only be inferred from the whole database. An example of the later is when a database owner pretends to publish a database but avoiding that third

parties are able to mine certain rules which are of high relevance in the business.

- **User privacy.** As stated above, the role of users is similar to the one of respondents in the fact that their data are collected. The difference is that here users can act to protect their privacy. Therefore, user privacy focus on tools that can be implemented and used by users. For example, to avoid leaving trails on their interests when accessing a search engine.

We have two sections below, one titled "User privacy" and the other titled "Respondent and owner privacy".

2.2 On the computations to be done

Data are protected to be used by third parties. If we know how data will be used, and what the third party wants to compute we can take that into account in the data protection process. For example, we may know that a statistician wants to apply a linear regression of income with respect to age or know that a data miner will apply some non supervised machine learning algorithm (e.g., some clustering algorithms).

In this dimension we consider three situations. We describe them below.

- **Computation-driven or specific-purpose data privacy protection methods.** In this case, we know which type of algorithm a researcher will apply to the data, and we can tailor the protected data to this type of method.

- **Data-driven or general-purpose data privacy protection methods.** This corresponds to the case in which there is only rough information or no information at all on the type of analysis to apply to the data. This is the case when data are published in the web.

- **Result-driven data protection privacy methods.** In this case we know the analysis that the researcher will do on the data. Nevertheless, there is a fundamental difference with computation-driven methods because protection is not focused on the original database but on the results obtained from the analysis. That is, we want to avoid that the data miner or statistician gets some particular results. The case described above in which we want to avoid a data miner to obtain a certain association rule from the data belongs to this case.

We can find in the literature another classification (dimension) of methods distinguishing perturbative and cryptographic methods. Perturbative methods are those that add some kind of noise to the data. That is, they mask the original data so that the true values are no longer found. Examples are adding noise or swapping values of the data. In this way, disclosure risk is reduced at the cost of some information (or utility) loss. On the contrary, cryptographic methods describe protocols so that researchers get their desired result without

accessing to the original data. Most perturbative methods can be considered as data-driven while cryptographic methods can be considered as computation-driven. The latter can only be defined when we know which is the intended computation of the user.

We will discuss in more detail these methods later in relation to respondent and owner privacy.

2.3 On the number of data sources

The number of data sources is another way to distinguish data privacy protection methods. There are methods to be applied when we only want to publish a single data set, when we want to publish different data sets (e.g. multiple tables of a single database), and when we want to obtain a computation from multiple data sets.

Methods focused on the publication of a single data set typically correspond to data-driven methods. The same case occurs when a single owner publishes several data sets (see e.g. [73, 57, 58]). A concrete case of this last scenario corresponds to the protection of stream or dynamic data. Protecting data streams can be done by producing several protected data sets in a timely basis or providing incremental versions of the data set [46, 10, 100, 9]. Dynamic data, which also considers deletion of elements, produces an evolving protected data set that reflects the updates (deletions and/or insertions) in the original data [87, 97]. Computation driven methods can be applied if the analysis of the data miner or statistician is known.

The case of computing a function from multiple data sets typically corresponds to computation-driven methods. In fact, most computation-driven methods focus on the following type of problem: n data owners decide to compute a function f of their data so that the only additional knowledge each of the owners get after the computation of f is the outcome of f applied to their data. Cryptographic protocols are defined for this purpose [89].

An important advantage of cryptographic methods is that they compute the function exactly (there is no error in the outcome of the function) and they ensure complete privacy. The main inconvenience is that if the function is changed, the protocol has to be changed. This is so, even in the case of a small variation in the function. So, the main disadvantage is that there is no flexibility on the function to be computed. In some scenarios there might be also computational or communication costs, since cryptographic operations have to be performed through a communication protocols in a reasonable time. On the contrary, data-driven (perturbative) methods are not exact because they decrease risk by adding some noise into the data, and this causes a perturbation in the results of any analysis, and in addition they do not ensure 100% privacy. The risk depends on the amount and type of perturbation added into the data. However, these methods permit the use of the same data to compute different functions. So, the main advantage is flexibility.

3 User privacy

As stated above, user privacy focuses on methods that can be implemented and applied by the user to ensure his own privacy. We can distinguish two main families of methods:

- Methods to protect the identity of the user.
- Methods to protect the data generated by the user.

Tools for anonymous communications belong to user privacy. These tools are expected to be implemented by the user to avoid the disclosure of some information related to him. For example, when a user A sends a message m to user B, A may want to avoid that third parties know that he is the sender of m, or hide the content of m (or try to keep others unaware that he sent a message after all).

Tools for user privacy have also been developed in the context of querying databases or search engines. In this case if A queries a search engine Y with query q, we may have the case that we want to avoid Y to know who is the sender of the query, and the case in which Y knows that A is the sender but is not aware of the query q.

We will give some examples of these tools in the next two sections.

3.1 Methods to protect the identity of the user

In the context of communication, we have anonymous communication mechanisms in which the sender of the message (or, in general, the origin of the communication) is not disclosed. Mix networks [12], onion routing [64] and crowds [65] are examples of such systems.

In the context of querying databases, this problem is studied in anonymous database search. One approach to this problem is allowing users to submit queries in behalf of other users. P2P UPIR (Peer-to-peer User Private Information Retrieval) [70, 71, 11, 91, 66] follows this approach. Cryptography is used to define communities of users, and communication spaces.

3.2 Methods to protect the data generated by the user

In the context of communication, cryptographic mechanisms are used to protect the content of the messages. In addition to that, there are systems developed to ensure unobservability, i.e. that third parties do not even know that a message is sent, for example, we have dining cryptographer networks [13].

Private information retrieval (PIR) studies this type of problems in the context of querying databases. Informally, this problem can be stated as finding a way to retrieve an element of a database without the database being able to deduce which element is of interest to the user.

Information Theoretic PIR faces this problem considering the case in which there is no privacy breach even in the case of an unlimited computing power.

However, it has been proven that if we consider a single database, all information theoretic PIR schemes require $\Omega(n)$ bits of information (where n is the number of records in the database). This means (see [16]) that essentially the only thing that the user can do to avoid the database to know his query is to ask for a copy of the whole database.

Because of this theoretical result different alternatives have been considered in the literature. We review some of them below.

First, within the information theoretic PIR the literature considers solutions in which instead of a single database there are replicated copies of this database. Then, the user queries differently each of the copies and from the results of the queries obtains the desired result. In this case, solutions sublinear in n, exist. Some of the solutions are resistant to coalitions of databases. That is, even in the case a certain number of databases collude they will not be able to find out which is the query of the user. See e.g. [16].

Another approach is computational PIR (cPIR). In this case, a server with a limited computational capacity is considered. [14] and [40] are two of the proposed solutions for cPIR.

A third approach is the use of trusted-hardware. See e.g. [99] and [96] on trusted-hardware PIR.

These three approaches are based on cryptographic tools and ensure no privacy leakage. Another approximation consists on methods that mask the real query in a set of other queries. This is the case of GooPIR [20] and TrackMeNot [33] or [77, 92]. They add to the query, either at the query level or at the session level, additional terms with the goal that the server cannot distinguish the real queries of the user among the added ones. This query obfuscation approach can however be attacked by analyzing the user query history from the server side. In [62] authors can re-identify users from their obfuscated queries by using common classifiers and clustering techniques on the user query history.

DisPA (for Dissociating Privacy Agent) [35, 36] follows another approach, also to protect the data generated by the user. In this case, the system (a plug-in for Firefox) generates different identities for a given user, and then distributes the queries among the identities. The basis of this system is to consider that what makes a user unique is the union of all queries. Therefore, the disaggregation of queries permits to keep the profile of the user unknown to the search engine. Disaggregation of queries is done according to topics, so if a user often queries about data privacy, Japanese recipes, and sports/squash it will result that the search engine will just know that there are three individuals one interested on data privacy, another on Japanese recipes and a third one about Sports/squaix.

4 Respondent and owner privacy

According to what has been described in the previous sections, we have that respondent and owner privacy are typically implemented by the owner of a database. According to our discussion on the dimensions about the computa-

tions to be done and the number of sources, we have that there are the following typical scenarios in respondent and owner privacy.

- **Result-driven methods** (mainly used in owner privacy). Given a database D, a data mining algorithm A, and a certain knowledge K that we do not want to disclose, the goal is to modify D into D' so that the algorithm A cannot infer K from D'. [4] is an overview on this topic, and [30, 31] describe algorithms in the case that A are rule mining algorithms.

- **Computation-driven methods** with the typical scenario with several data sources belonging to different data owners. This scenario corresponds to owner privacy. As described above, this type of problem is solved defining cryptographic protocols for the specific function the owners want to compute. [89] describes several computation-driven methods and [37] is a survey on methods for horizontally partitioned data (i.e., different owners have data on different individuals but on the same variables.

- **Computation-driven methods** with a single database release. If the function is completely specified, the most common scenario is when researchers can access to a database and send specific queries (see e.g. [24]). If the function is not completely specified but it is known that the user applies e.g. clustering, then data-driven approaches would be applied with particular emphasis on methods that behave well with respect to this use (clustering). There are studies (see e.g., [41]) comparing different data-driven methods with respect to clustering, supervised learning algorithms, and so on.

- **Data-driven either with one or multiple data releases.** As already cited in a previous section [73, 9, 57, 58] focus on data-driven approaches of multiple data releases or streaming data. Data-driven approaches for a single database are further discussed in the next section.

4.1 Data-driven methods

As described above, data-driven methods, usually referred as masking methods, are appropriate when we do not know before hand what type of analysis will be applied to the data. Given a database D, the usual way to proceed is to modify D into D' so that the risk of disclosure decreases while at the same time we preserve the utility of D. That is, modify D into D' so that the disclosure risk decreases while keeping information loss as low as possible. Note that we use the term information loss as a computation oriented definition of data utility when referring to data utility. More information loss means less utility of the data once it is masked.

Due to the fact that these methods are not disclosure risk free, several disclosure risk measures have been considered in the literature to quantify the risk in D'. At the same time, as the modification of D can decrease the utility of the database, some information loss measures have been defined to measure the

extent of this loss. Naturally, disclosure risk decreases at the expenses of some information loss. Then, a good privacy method is the one that modifies D into D' in such a way that the disclosure risk is very low and the information loss is also very low.

As a summary, we have that research in data-driven methods needs to focus on masking methods, disclosure risk measures and utility measures. We will discuss in the following three sections: disclosure risk measures, information loss measures, and data masking methods.

4.2 Disclosure risk and some definitions of privacy for data-driven methods

In this section we describe several approaches to measure the degree of privacy provided by a given method. This measures are normally referred as disclosure risk measures, or presented as privacy properties to be satisfied by the protected data.

4.2.1 Properties for disclosure risk

Data-driven methods add noise to the data to avoid disclosure. Then, we can either consider risk as a boolean *condition* that is either satisfied or not satisfied, or as a measurable (non-boolean) *condition* and define measures of risk.

Differential privacy [24] and k-anonymity [76, 75, 80, 79] follow this first approach. That is, they define conditions in which we say that the file satisfies our requirements of privacy. At the same time, such definitions permit to define algorithms that given a privacy condition only focus on the minimization of information loss.

The k-anonymity property ensures that in a protected data set there are at least k records indistinguishable from each other. Or, from the point of view of re-identification, that the probability of re-identifying and individual from the data set is $1/k$. This property is very common in statistical data like census, where the perturbation is applied to attributes known as quasi-identifiers (they cannot be used to re-identify an individual by themselves, but their union might be), and sensitive attributes are left without perturbation. Example of quasi-identifiers can be age, sex, or postal code, while typical sensitive attributes are salary or disease. Consider a k-anonymous data set, where the set of k records sharing the same quasi-identifiers (known as anonymity set), also have the same sensitive attribute. Although the table might not be used to directly re-identify a given individual it leaks information about the sensitive attribute. That is an attacker will know the sensitive attribute of the individual knowing to which anonymity set it belongs. This is a well known problem of k-anonymity (see [17] for a detailed description of the problem and discussion of current solutions). To address this issue several properties have emerged. l-diversity [47] requires at least l well represented values of the sensitive attribute in each anonymity set. Moreover t-closeness [45] requires the distribution of sensitive attributes to be close to their distribution in the overall data set. In this same line, p-sensitivity

has also been defined as the concrete case of l-diversity where the number of values for each sensitive attribute is at least p for each anonymity set. See [88] for a review of p-sensitive k-anonymity models.

Some generalizations of k-anonymity have been defined in the literature. E.g. k-confusion [74] and probabilistic k-anonymity [69], in which instead of requiring indistiguishable records the focus is on the probability of re-identification. k-concealment [81] requires computationally indistinguishable records (each record can be matched with $k-1$ generalized records).

Differential privacy states that adding or removing an item from a data set does not significantly affect the outcome of any analysis. That is, the outcomes should be probabilisitically similar. This definition of privacy has boosted a great number of literature on mechanisms to provide differential privacy [15], but it has also raised some concerns. E.g. [68, 5] question the practicality of differential privacy as a general case approach for data privacy.

4.2.2 Disclosure risk measures

As an alternative to boolean conditions, there are measures of disclosure risk defined under the premise that risk is not binary but a measurable condition. Then, it has sense to consider different levels of risk and the trade-off of the risk with respect to the utility of the data. In this setting, the problem is not to define algorithms with the only purpose of optimizing information loss but with the purpose of finding a good trade-off between information loss and disclosure risk. Therefore, from an optimization point of view, we have a multi-objective (two-objective) optimization problem instead of a minimization problem. The perspective of an optimization problem has been exploited in e.g. [48, 67].

Some of the measures of disclosure risk are based on the concept of uniqueness, and on re-identification algorithms. Key references on disclosure risk based on uniqueness are [26, 25] and based on re-identification algorithms are [94, 22, 86]. Disclosure risk measures based on re-identification algorithms model the scenario in which intruders use their knowledge (represented in terms of a database) to attack the published data set. In this case, the intruders will try to link their data with the one in the data set by means of the best available technology for database integration (re-identification algorithms, schema matching and record linkage algorithms). This approach is flexible enough to cope with a large number of scenarios. For example, disclosure risk has been studied for masked data [98], synthetic data [85, 94], and for the case in which the intruder and the protected data are not using the same variables [82, 83] or are using different terms (e.g. ontology-base record linkage in [49]).

As a general purpose estimation of the disclosure risk, re-identification can be attempted on the protected data set assuming the knowledge of all the attributes from the original data set. For example by applying record linkage between the original records and the same protected records. This approach was introduced in [78], and widely used afterwards [94, 42, 95, 85]. The percentage of re-identification is used as a generic index of disclosure risk, that can be used to compare different masking methods [22]. A parametrized record linkage allows

to provide an upper bound index of re-identification by finding the optimal distance between records (one that provides the highest re-identification index) using machine learning techniques [2].

Disclosure risk based on re-identification methods can also be used to model the case in which the intruder uses information about the data masking process to attack the data. That is, in case that an institution publishes a data set giving information on the algorithm applied and the parameters used, we can use this information to attack more effectively the data. This has been proven to be effective in [59, 60, 61] in the case that data was protected using rank swapping and microaggregation. Methods resistant to this type of attacks are needed for the sake of transparency [38].

4.3 Utility and information loss measures

Utility measures are used to measure in what extent the protected database diverges from the original one for some statistics and analysis. We can measure the utility of the data once it is masked as compared to the original one. This measure can be given in terms of the loss of information produced by the masking method. A masking method that yields a higher loss of information will present lower utility. Then, given a database D, a protected database D', and a certain analysis f, an information loss measure is a function

$$IL_f(D, D') = divergence(f(D), f(D')),$$

where *divergence* is a way to compare the result of the analysis f on D and D'.

Naturally, the function *divergence* should be zero when $D = D'$, and increasing the more $f(D)$ and $f(D')$ differ.

We can distinguish between generic utility (or information loss) measures and specific utility (or information loss) measures. We have specific utility measures when they focus on particular uses of the data. This would be the case if we consider clustering as a data use, and then we use clustering algorithms and functions to compare partitions to define an information loss measure. This is the case in [41]. Otherwise, we have generic utility measures when we e.g. aggregate some statistics of the data. This latter approach is used in [51, 21].

4.4 Masking methods

Data masking methods are typically classified in three main classes. See [23, 34] for detailed descriptions of the methods.

- **Perturbative methods.** Given a database D, these methods modify the database adding some noise to D. This can be modeled as follows:

  $$D' = D + \epsilon.$$

  There are several perturbative methods. The simplest one is noise addition where the error to be added to D follows a normal distribution.

Most important methods are noise addition [8], multiplicative noise [39], microaggregation (applicable to all types of data) [18], rank swapping (for data in ordinal or numerical scales) [52], and PRAM (for ordinal or categorical scales) [29].

- **Non-perturbative methods.** Given a database, these methods modify the database changing the level of detail of the data but not introducing errors to the data. One masking method is generalization, which replaces a category by a more general one (e.g., town is replaced by county), another one is suppression (suppression can be considered as equivalent to a generalization to the most general category), and finally we have discretization in the case of numerical data (again, a kind of generalization).

- **Synthetic data generators.** Instead of publishing the original data, we generate a model of the data and then replace the original values by the outcomes of the model. This approach can be considered as a kind of perturbative method.

All the methods described here have been used, and compared in terms of their trade-off between information loss and disclosure risk (defined in terms of re-identification algorithms). In the case of using differential privacy as a standard to ensure risk, the most common masking method is to use Laplace noise. See e.g. [24] for details. In the case of using k-anonymity as the standard for risk, the most common masking method is generalization and suppression. See e.g. [43, 44, 79] for details. Note that such methods focus on numerical data for differential privacy and categorical data for k-anonymity.

5 Discussion

For details in the topics presented in this paper, the reader can look to the following books [23, 34, 89, 93] and also to the material in the web page [102]. [23, 34, 93] follow a SDC perspective, while [89] a PPDM perspective. In addition, [55, 84] focus on some specific topics. [55] is a survey on the use of information fusion techniques in data privacy, mainly focusing on the use of aggregation functions and record linkage techniques. [84] focuses on the use of explicit knowledge (either in the data privacy protection process or in re-identification).

We have discussed the main topics related to data privacy. Although the discussion is general and independent on the type of data used, research is not. Initial research in the field focused on standard databases with either numerical and categorical data. Further research has been done in longitudinal/time-series data [1], and there are more recent trends on data privacy for (search) logs [56, 63, 54, 7, 27], locations [32, 50], and graphs [72, 101].

Research in these areas follow the same lines discussed here. There is research on online social networks that focus on respondent and owner privacy, while there is other research focusing on user's perspective (i.e., user-privacy).

There are perturbative approaches (e.g. to avoid re-identification) and non-perturbative approaches (e.g. to achieve k-anonymity) for online social networks, and also results to achieve differential privacy in online social networks. Similarly, there are also such lines of research in location privacy, or in methods for search logs.

In any case, the development of methods has to take into account the specificities of the data. Ignoring them can cause disclosure as an intruder can use such vulnerabilities to attack the data. Some of the scandals [6, 53] in privacy have been due to a lack of understanding of these specificities (e.g., different logs from the same person, which alone are not sensitive, can be combined to re-identify this person).

The data privacy research and application field is gaining popularity and there is a growing community interested in advancing the research field. There are open issues and research fields specially active: data privacy techniques for very large datasets, including stream data is becoming important as the data processing capabilities are rapidly increasing. Moreover, the interest of other research areas in data privacy is also becoming very relevant, examples are machine learning, or game theory.

In this paper we have presented a review of the main techniques related with data privacy. We have presented the main dimensions that permit to classify data privacy protection methods, we have enumerated some of them, and discussed the main concepts in the area.

Acknowledgments

Partial support by the Spanish MEC projects ARES (CONSOLIDER INGENIO 2010 CSD2007-00004), TIN2010-15764, and COPRIVACY (TIN2011-27076-C03-03) is acknowledged. Partial support of the European Project DwB (Grant Agreement Number 262608) is also acknowledged.

References

[1] Abowd, J. M., Woodcock, S. D. (2001) Disclosure limitation in longitudinal linked data, in P. Doyle, J. I. Lane, J. J. M. Theeuwes, L. Zayatz (eds.) Confidentiality, Disclosure, and Data Access: Theory and Practical Applications for Statistical Agencies, Elsevier Science, 215-277

[2] Abril. D, Navarro-Arribas, G., Torra, V., (2012) Improving record linkage with supervised learning for disclosure risk assessment, Information Fusion, 13:4 274-284.

[3] Aggarwal, C.C., Yu, P.S., (2008) A General Survey of Privacy-Preserving Data Mining Models and Algorithms, in: Privacy-Preserving Data Mining, Advances in Database Systems. Springer, 11-52.

[4] Atzori, M., Bonchi, F., Giannotti, F., Pedreschi, D. (2008) Anonymity preserving pattern discovery, The VLDB Journal 17 703-727.

[5] Bambauer, J., Muralidhar, K., Sarathy, R. (2014) Fool's gold: An Illustrated Critique of Differential Privacy Journal of Entertainment & Technology Law, *in press*.

[6] Barbaro, M., Zeller, T. (2006) A Face Is Exposed for AOL Searcher No. 4417749, The New York Times, August 9, 2006. Retrieved April 25, 2010.

[7] Batet, M., Erola, A., Sanchez, D., Castella-Roca, J. (242) Utility preserving query log anonymization via semantic microaggregation. Inf. Sci. 242: 49-63.

[8] Brand, R., (2002) Microdata Protection through Noise Addition, in Inference Control in Statistical Databases, Lecture Notes in Computer Science vol. 2316, Springer 97-116.

[9] Byun, J.-W., Sohn, Y., Bertino, E., Li, N., 2006. Secure Anonymization for Incremental Datasets, Secure Data Management, Lecture Notes in Computer Science, 48-63.

[10] Cao, J., Carminati, B., Ferrari, E., Tan, K.-L., (2011) CASTLE: Continuously Anonymizing Data Streams. IEEE Transactions on Dependable and Secure Computing 8, 337-352.

[11] Castella-Roca, J., Viejo, A., Herrera-Joancomarti, H., (2009) Preserving user's privacy in web search engines. Computer Communications 32(13-14): 1541-1551.

[12] Chaum, D. L. (1981) Untraceable electronic mail, return addresses, and digital pseudonyms, Communications of the ACM 24:2 84-88.

[13] Chaum, D. (1985) The dining cryptographers problem: unconditional sender and recipient untraceability, J. Cryptology 1 65-75.

[14] Chor, B., Gilboa, N. (1997) Computationally private information retrieval (extended abstract). Proceedings of the twenty-ninth annual ACM symposium on Theory of computing (STOC'97) 304-313.

[15] Dankar, F., El Emam, K. (2013), Practicing Differential Privacy in Health Care: A Review, Transactions on Data Privacy 6:1 35-67

[16] Chor, B., Kushilevitz, E., Goldreich, O., Sudan, M. (1998). Private information retrieval. J. ACM 45:6 965-981.

[17] De Capitani di Vimercati, S., Foresti, S., Livraga, J., Samarati, P. (2012) Data Privacy: Definitions and Techniques, International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems, 20:6 793-818.

[18] Defays, D., Nanopoulos, P., (1993) Panels of enterprises and confidentiality: the small aggregates method, in: Proceedings of the 1992 Symposium on Design and Analysis of Longitudinal Surveys, Statistics Canada, pp. 195-204.

[19] Domingo-Ferrer, J. (2007) A three-dimensional conceptual framework for database privacy, SDM 2007, Lecture Notes in Computer Science 4721, 193-202.

[20] Domingo-Ferrer, J., Solanas, A., Castella-Roca, J. (2009) $h(k)$-private information retrieval from privacy-uncooperative queryable databases, Online Information Review, 33:4 720-244.

[21] Domingo-Ferrer, J., Torra, V. (2001) Disclosure Control Methods and Information Loss for Microdata, in P. Doyle, J. I. Lane, J. J. M. Theeuwes, L. Zayatz (eds.) Confidentiality, Disclosure, and Data Access: Theory and Practical Applications for Statistical Agencies, Elsevier Science, 91-110.

[22] Domingo-Ferrer, J., Torra, V. (2001) A quantitative comparison of disclosure control methods for microdata, in P. Doyle, J. I. Lane, J. J. M. Theeuwes, L. Zayatz (eds.) Confidentiality, Disclosure and Data Access: Theory and Practical Applications for Statistical Agencies, North-Holland, 111-134.

[23] Duncan, G. T., Elliot, M., Salazar, J. J. (2011) Statistical confidentiality, Springer.

[24] Dwork, C. (2006) Differential privacy, Proc. ICALP 2006, Lecture Notes in Computer Science 4052, 1-12.

[25] Elliot, M. (2002) Integrating file and record level disclosure risk assessment, in J. Domingo-Ferrer, Inference Control in Statistical Databases, Lecture Notes in Computer Science 2316 126-134.

[26] Elliot, M. J. Skinner, C. J., Dale, A. (1998) Special Uniqueness, Random Uniques and Sticky Populations: Some Counterintuitive Effects of Geographical Detail on Disclosure Risk, Research in Official Statistics 1:2 53-67.

[27] Erola, A. (2013) Contributions to privacy in web search engines. Phd Thesis, Universitat Rovira i Virgili.

[28] Gehrke, J., Hay, M., Lui, M., Pass, R. (2012), Crowd-Blending Privacy, Advances in Cryptology - Crypto 2012, volume 7417 of Lecture Notes in Computer Science, 479-496.

[29] Gouweleeuw, J., Kooiman, P., Willenborg, L., Wolf, P., (1998) Post randomisation for statistical disclosure control: theory and implementation Journal of Official Statistics, 14 (4), pp. 463-478

[30] Hajian, S. (2013) Simultaneous discrimination prevention and privacy protection in data publishing and mining, PhD Dissertation, Universitat Rovira i Virgili.

[31] HajYasien, A. (2007) Preserving privacy in association rule hiding, PhD Dissertation, Griffith University.

[32] Ho, S.-S., Ruan, S. (2013) Preserving Privacy for Interesting Location Pattern Mining from Trajectory Data, Transactions on Data Privacy 6:1 87 - 106

[33] Howe, D.C., Nissenbaum, H. (2009) TrackMeNot: Resisting Surveillance in Web Search, Lessons from the Identity Trail: Anonymity, Privacy, and Identity in a Networked Society 417-436.

[34] Hundepool, A., Domingo-Ferrer, J., Franconi, L., Giessing, S., Nordholt, E. S., Spicer, K., de Wolf, P.-P. (2012) Statistical Disclosure Control, Wiley.

[35] Juárez, M., Torra, V. (2013) Toward a Privacy Agent for Information Retrieval, Int. J. Intel. Syst. 28:6 606-622.

[36] Juárez, M., Torra, V. (2013) A self-adaptive classification for the dissociating privacy agent, Proc. PST 2013 44-50.

[37] Kantarcioglu, M. (2008) A survey of privacy-preserving methods across horizontally partitioned data, in C. C. Aggarwal, P. S. Yu (eds.) Privacy-Preserving Data Mining: Models and Algorithms, Springer, 313-335.

[38] Karr, A. F. (2009). The role of transparency in statistical disclosure limitation. Joint UNECE/Eurostat Work Session on Statistical Data Confidentiality

[39] Kim, J.J., Windler, W.E., (2003) Multiplicative Noise for Masking Continuous Data. Research Report Series Statistics #2003-01, U.S. Bureau of the Census, Statistical Research Division, April 17, 2003.

[40] Kushilevitz, E., Ostrovsky, R., (1997) Replication is not needed: single database, computationally-private information retrieval, Proceedings 38th Annual Symposium on Foundations of Computer Science, 364-373.

[41] Ladra, S., Torra, V. (2008) On the comparison of generic information loss measures and cluster-specific ones, Intl. J. of Unc., Fuzz. and Knowledge-Based Systems, 16:1 107-120.

[42] Lambert, D. (1993) Measures of disclosure risk and harm Journal of Official Statistics, 9, 313-331

[43] LeFevre, K., DeWitt, D. J., Ramakrishnan, R. (2005) Multidimensional k-anonymity, Technical Report 1521, University of Wisconsin.

[44] LeFevre, K., DeWitt, D. J., Ramakrishnan, R. (2005) Incognito: Efficient Full-Domain K-Anonymity, SIGMOD 2005.

[45] Li, N., Li, T., Venkatasubramanian, S., (2007) t-Closeness: Privacy Beyond k-Anonymity and l-Diversity, IEEE 23rd International Conference on Data Engineering, ICDE 2007, 106-115.

[46] Li, J., Ooi, B.C., Wang, W., (2008) Anonymizing Streaming Data for Privacy Protection, IEEE 24th International Conference on Data Engineering, 2008. ICDE 2008, 1367-1369.

[47] Machanavajjhala, A., Kifer, D., Gehrke, J. (2007) L-diversity: Privacy beyond k-anonymity. ACM Trans. Knowl. Discov. Data 1:1.

[48] Mares, J., Torra, V., (2011) PRAM Optimization Using an Evolutionary Algorithm, Privacy in Statistical Databases, Lecture Notes in Computer Science 6344 970-106.

[49] Martinez, S., Valls, A., Sanchez, D. (2011) An ontology-based record linkage method for textual microdata. CCIA 2011, 130-139

[50] Masoumzadeh, A., Joshi, J. (2013) Top Location Anonymization for Geosocial Network Datasets, Transactions on Data Privacy 6:1 107 - 126

[51] Mateo-Sanz, J. M., Domingo-Ferrer, J. Sebé, F. (2005) Probabilistic information loss measures in confidentiality protection of continuous microdata, Data Mining and Knowledge Discovery, 11:2 181-193.

[52] R. Moore, (1996) Controlled Data Swapping Techniques for Masking Public Use Microdata Sets, US Bureau of the Census (unpublished manuscript).

[53] Narayanan, A., Shmatikov, V., (2008) Robust De-anonymization of Large Sparse Datasets. In Proceedings of the 2008 IEEE Symposium on Security and Privacy (SP '08). IEEE Computer Society, 111-125.

[54] Navarro-Arribas, G., Torra, V. (2010) Privacy-preserving data-mining through microaggregation for web-based e-commerce. Internet Research, 20:3, 366-384.

[55] Navarro-Arribas, G., Torra, V., (2012). Information fusion in data privacy: A survey. Information Fusion 13:4, 235-244.

[56] Navarro-Arribas, G., Torra, V., Erola, A., Castella-Roca, J. (2012) User k-anonymity for privacy preserving data mining of query logs. Information Processing and Management, 48(3):476-487.

[57] Nergiz, M. E., Clifton, C., Nergiz, A. E. (2007) MultiRelational k-Anonymity, Data Engineering, 2007. ICDE 2007. IEEE 23rd International Conference on 2007 1417-1421.

[58] Nergiz, M. E., Clifton, C., Nergiz, A. E.., MultiRelational k-Anonymity (2009) IEEE Trans. on Knowledge and Data Engineering 21:8 1104-1117

[59] Nin, J., Herranz, J., Torra, V. (2007) Rethinking Rank Swapping to Decrease Disclosure Risk, Data and Knowledge Engineering, 64:1 346-364.

[60] Nin, J., Herranz, J., Torra, V. (2008) On the Disclosure Risk of Multivariate Microaggregation, Data and Knowledge Engineering, 67:3 399-412.

[61] Nin, J., Torra, V. (2009) Analysis of the Univariate Microaggregation Disclosure Risk, New Generation Computing, 27 177-194.

[62] Peddinti, S. T., Saxena, N. (2010) On the Privacy of Web Search Based on Query Obfuscation: A Case Study of TrackMeNot, Proc. Privacy Enhancing Technologies, LNCS 6205 19-37.

[63] Poblete, B., Spiliopoulou, M., Baeza-Yates, R. (2010) Privacy-preserving query log mining for business confidentiality protection. ACM Trans. Web 4:3.

[64] Reed, M. G., Syverson, P. F., Goldschlag, D. M. (1998) Anonymous connections and onion routing, IEEE J. of Selected Areas in Communications 16:4 482-494.

[65] Reiter, M., Rubin, A. (1998) Crowds: Anonymity for Web Transactions, ACM Trans. on Information and System Security 1:1 66-92.

[66] Romero-Tris, C., Castella-Roca, J., Viejo, A., (2011) Multi-party Private Web Search with Untrusted Partners. SecureComm 2011: 261-280.

[67] Sebe, F., Domingo-Ferrer, J., Mateo-Sanz, J.M., Torra, V. (2002) Post-Masking Optimization of the Tradeoff between Information Loss and Disclosure Risk in Masked Microdata Sets, Inference Control in Statistical Databases, Lecture Notes in Computer Science 2316 187-196.

[68] Sarathy, R., Muralidhar, K. (2011) Evaluating Laplace Noise Addition to Satisfy Differential Privacy for Numeric Data, Transactions on Data Privacy 4:1 1-17.

[69] Soria-Comas, J., Domingo-Ferrer, J. (2012) Probabilistic k-anonymity through microaggregation and data swapping, FUZZ-IEEE 1-8.

[70] Stokes, K., Bras-Amorós, M. (2010) Optimal configurations for peer-to-peer user-private information retrieval, Computers & Mathematics with Applications 59:4 1568-1577.

[71] Stokes, K., Bras-Amorós, M. (2011) On query self-submission in peer-to-peer user-private information retrieval. PAIS 2011: 7

[72] Stokes, K., Torra, T. (2012) Reidentification and k-anonymity: a model for disclosure risk in graphs. Soft Comput. 16(10): 1657-1670

[73] Stokes, K., Torra, V. (2012) Multiple Releases of k-Anonymous Data Sets and k-Anonymous Relational Databases, Int. J. of Unc., Fuzziness and Knowledge-Based Systems 20:6 839-854.

[74] Stokes, K., Torra, V. (2012) n-confusion: a generalization of k-anonymity. EDBT/ICDT Workshops, 211-215

[75] Samarati, P. (2001) Protecting Respondents' Identities in Microdata Release, IEEE Trans. on Knowledge and Data Engineering, 13:6 1010-1027.

[76] Samarati, P., Sweeney, L. (1998) Protecting privacy when disclosing information: k-anonymity and its enforcement through generalization and suppression, SRI Intl. Tech. Rep.

[77] Sanchez, D., Castella-Roca, J., Viejo, A., (2013) Knowledge-based scheme to create privacy-preserving but semantically-related queries for web search engines. Inf. Sci. 218: 17-30.

[78] Spruill, N.L. (1982) Measures of confidentiality, Proc. Survey Research Section American Statistical Association, 260-265.

[79] Sweeney, L. (2002) Achieving k-anonymity privacy protection using generalization and suppression, Int. J. of Unc., Fuzz. and Knowledge Based Systems 10:5 571-588.

[80] Sweeney, L. (2002) k-anonymity: a model for protecting privacy, Int. J. of Unc., Fuzz. and Knowledge Based Systems 10:5 557-570.

[81] Tassa, T., Mazza, A., Gionis, A. (2012) k-Concealment: An Alternative Model of k-Type Anonymity, Transactions on Data Privacy 5:1 189 - 222

[82] Torra, V. (2000) Towards the Re-identification of Individuals in Data Files with Non-common Variables, Proc. ECAI 2000 326-332.

[83] Torra, V. (2004) OWA operators in data modeling and reidentification, IEEE Trans. on Fuzzy Systems 12:5 652-660.

[84] Torra, V., (2011) Towards Knowledge Intensive Data Privacy, Data Privacy Management and Autonomous Spontaneous Security, Lecture Notes in Computer Science 6514 1-7.

[85] Torra, V., Abowd, J. M., Domingo-Ferrer, J. (2006) Using Mahalanobis Distance-Based Record Linkage for Disclosure Risk Assessment, Lecture Notes in Computer Science 4302 233-242.

[86] Torra, V., Stokes, K. (2012) A Formalization of Record Linkage and its Application to Data Protection 20:6 907-920.

[87] Truta, T.M., Campan, A. (2007) K-anonymization incremental maintenance and optimization techniques, in Proceedings of the 2007 ACM symposium on Applied computing, 380-387.

[88] Truta, T.M., Campan, A., Sun, X. (2012) An Overview of P-Sensitive k-Anonymity Models for Microdata Anonymization, International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems, 20:6 819-838.

[89] Vaidya, J., Clifton, C. W., Zhu, Y. M. (2006) Privacy Preserving Data Mining, Springer.

[90] Verykios, V.S., Bertino, E., Fovino, I.N., Provenza, L.P., Saygin, Y., Theodoridis, Y., (2004) State-of-the-art in Privacy Preserving Data Mining. SIGMOD Rec. 33, 50-57.

[91] Viejo, A., Castella-Roca, J. (2010) Using social networks to distort users' profiles generated by web search engines. Computer Networks 54(9): 1343-1357.

[92] Viejo, A., Sanchez, D., Castella-Roca, J. (2012) Preventing automatic user profiling in Web 2.0 applications. Knowl.-Based Syst. 36: 191-205.

[93] Willenborg, L., de Waal, T. (2001) *Elements of Statistical Disclosure Control*, Lecture Notes in Statistics, Springer-Verlag.

[94] Winkler, W.E. (2004) Re-identification methods for masked microdata, PSD 2004, Lecture Notes in Computer Science 3050 216-230.

[95] Winkler, W.E. (2004) Masking and re-identification methods for public use microdata: overview and research problems, Privacy in Statistical Databases, Lecture Notes in Computer Science 3050 231-246.

[96] Wang, S., Ding, X., Deng, R. H., Bao, F. (2006) Private Information Retrieval Using Trusted Hardware, Proc. ESORICS 2006, LNCS 4189 49-64.

[97] Xiao, X., Tao, Y., (2007) M-invariance: towards privacy preserving republication of dynamic datasets, Proceedings of the 2007 ACM SIGMOD International Conference on Management of Data, 689-700.

[98] Yancey, W. E., Winkler, W. E., Creecy, R. H. (2002) Disclosure risk assessment in perturbative microdata protection, in J. Domingo-Ferrer (ed.) Inference Control in Statistical Databases, Lecture Notes in Computer Science 2316 135-152.

[99] Yang, Y., Ding, X., Deng, R. H., Bao, F. (2008) An Efficient PIR Construction Using Trusted Hardware, LNCS 5222 64-79.

[100] Zakerzadeh, H., Osborn, S.L., (2011) FAANST: Fast Anonymizing Algorithm for Numerical Streaming Data, Data Privacy Management and Autonomous Spontaneous Security, Lecture Notes in Computer Science, 6514 36-50.

[101] Zhou, B. and Pei. J. (2008) Preserving privacy in social networks against neighborhood attacks, Proc. ICDE 2008.

[102] `https://www.ppdm.cat/dp/`