Anonymization of statistical data

S. De Capitani di Vimercati, S. Foresti, G. Livraga, P. Samarati

DTI, University of Milan, Italy

Abstract. In the modern digital society, personal information about individuals can be collected, stored, shared, and disseminated much more easily and freely. Such data can be released in *macrodata* form, reporting aggregated information, or in *microdata* form, reporting specific information on individual respondent. Protecting data against improper disclosure is then becoming critical to ensure proper privacy of individuals as well as of public and private organizations, and several data protection techniques have been developed. In this paper, we characterize macrodata and microdata releases and then focus on microdata protection. We provide a characterization of the main microdata protection techniques and describe recent solutions for protecting microdata against identity and attribute disclosure, discussing some open issues that need to be investigated.

Keywords: H.2.7 [**Database Management**]: Database Administration - *Security, integrity and protection*; H.2.8 [**Database Management**]: Database Applications - *Statistical databases*; K.6.5 [**Management of Computing and Information Systems**]: Security and Protection

1 Introduction

We live in a digital society where we continually release personal information as we complete e-commerce transactions, create accounts, query search engines and participate to census surveys. At the same time, the rapid advances of the Information Technology have increased the ability of anyone, anywhere in the world, to easily collect, access, and analyze huge amounts of personal information that are available in the cyberspace. Privacy of the data is then becoming an issue that most people are concerned about and that has captured the attention of many researchers in several areas (e.g., [2, 5, 8, 9]). For the European Census 2011, all European citizens will be asked to reveal their personal data (e.g., date of birth, occupation, family composition, etc.), which will be made publicly available for future statistical analysis. To prevent privacy breaches, it will be necessary to adopt protection techniques that guarantee adequate protection.

Several techniques have been developed to protect from improper disclosure data that have been publicly or semi-publicly released. Typically, these techniques depend on how the data are released. In the past, data were released as *macrodata* or through *statistical databases*. Macrodata are aggregated pieces of information (statistics) on individuals or companies, usually presented as two-dimensional tables. In this case, the protection techniques are based on the selective obfuscation of sensitive cells, where the sensitivity can be defined according to different rules. A statistical database is a database whose users may retrieve only aggregate statistics. The protection techniques for statistical databases mainly consist in restricting the statistical queries that can be submitted, in modifying the query result, or in controlling responses to queries [6].

Today, data are often released in the form of *microdata*, that is, data related to individual respondents. The main advantage of releasing microdata instead of macrodata is an increased flexibility and availability of information for the data recipients since they can apply on microdata the specific analysis that is needed. To protect the anonymity of the *respondents* to whom microdata refer, data holders often remove or encrypt explicit identifiers such as names, addresses, and phone numbers. De-identifying data, however, provide no guarantee of anonymity since they may contain other identifying information (e.g., date of birth, sex, and geographical location) that uniquely or almost uniquely distinguishes the individual. By linking such information to publicly available databases reporting the individual's identity, data recipients can determine to which individual each piece of released data belongs, or restrict their uncertainty to a specific subset of individuals. This disclosure of an individual's identity (*identity disclosure*) often implies leakage of sensitive information (*attribute disclosure*) associated with the individual. For instance, in 2006 Netflix, an online DVD delivery service, started a competition whose goal was the improvement of its movie recommendation system based on users' previous ratings. To this purpose, Netflix released 100 million records about movie ratings of 500,000 of its subscribers. The released records were anonymized removing the personal identifying information of the subscribers. However, by linking the movie recommendations available on the Internet Movie Database (IMD) with the anonymized Netflix dataset, it was possible to re-identify individuals, thus revealing potentially sensitive information (e.g., apparent political preferences). Although people are today most conscious of privacy risks that characterize our digital society, data continues to be released and privacy continues to be an open issue. For instance, 2011 European Census will collect personal information (e.g., date of birth, occupation, family composition) of all European citizens, with the goal of releasing collected data for statistical analysis over the European population. Clearly, to prevent possible re-identifications, and therefore the disclosure of sensitive information, collected data must be adequately protected before their publication. Microdata protection techniques aim at avoiding identity and attribute disclosure by modifying data before their release or reducing the amount of information released.

In this paper, we survey the main techniques proposed for guaranteeing privacy in data publication. The remainder of this paper is organized as follows. Section 2 introduces the problem of protecting privacy in the publication of macrodata and microdata. Section 3 focuses on microdata and presents some protection techniques. Section 4 discusses research challenges that need to be taken into consideration in the future development of data protection techniques. Finally, Section 5 concludes the paper.

2 Privacy issues in data publication

Data publishing can be roughly classified in two main categories: *macrodata*, resulting from the aggregation of data; and *microdata*, reporting information referred to individual *respondents*, that is, the entities to which collected information refers (e.g., individuals or organizations). In this section, we present the main characteristics of macrodata and microdata, and illustrate the privacy issues arising from their publication.

2.1 Macrodata

Figure 1 An example of count table reporting, for each region, the number of patients with a given disease

The term *macrodata* refers to aggregated data, representing an estimation of a *statistical characteristic* of a given population. For instance, a statistical characteristic can be the number of patients by region and disease. Macrodata are usually represented in *tabular form*, that is, by means of tables. Each dimension of a macrodata table represents an attribute of the collected data (e.g., region, disease), and each cell contains the aggregate value of a statistical characteristic over the dimensions of the macrodata table. Typically, macrodata tables have two dimensions and report the row and column totals (i.e., *marginal totals*). Depending on the statistical characteristic considered, macrodata can be classified in the following three main categories [7].

- *Count tables*: each cell contains the *number* of respondents that have the same value over all the attributes in the table.
- *Frequency tables*: each cell contains the *percentage* of respondents over the total population that have the same value over all the attributes in the table.
- *Magnitude tables*: each cell contains an *aggregate value* of a quantity of interest, computed over the population that have the same value over all the attributes in the table.

Figure 1 illustrates an example of count table, representing the number of patients, aggregated over attribute `Region` (A, B, and C) and attribute `Disease` (*ulcera*, *cancer*, *gastritis*, and *broken bone*).

Although macrodata tables do not report information that refers to a single respondent, still there is the possibility to infer data about (a subset of) the macrodata respondents. For instance, consider the macrodata table in Figure 1. It is easy to see that all the respondents living in region C suffer from a stomach disease, since only the cells associated with *ulcera* and *gastritis* have a value different from zero. To avoid this inference problem, different macrodata protection techniques have been proposed in the literature [4]. A first solution is *sampling*, that is, the released macrodata table is obtained from a sample of the population (in contrast to the whole population). The original data in the sample are then multiplied by the sampling weight (i.e., the ratio between the number of respondents in the sample and the number of respondents in the population), before computing the aggregate function. Sampling helps in decreasing the risk of inferring information related to a specific respondent. As a matter of fact, the aggregate values in the macrodata table are not directly computed on the data referred to single respondents, but on their product with the sampling weight. Also, there is uncertainty about whether or not a specific respondent belongs to the sample. Sampling is however not sufficient for completely eliminating the possibility of inferring information related to individual respondents. Specific macrodata protection techniques have to be applied, which can operate directly on the original data (microdata) used for computing the macrodata, or on the macrodata themselves. An example of technique working on the original microdata is *confidential edit*. This technique was developed by the U.S. Census Bureau for protecting the tables prepared from the 1990 Census. The basic idea consists in: *1)* selecting a sample of the records in microdata; *2)* finding a match for the selected records in other geographical regions and on specific set of attributes; and *3)* swapping all attributes of the matching records.

The techniques directly working on macrodata typically consist in identifying *sensitive cells* (i.e., cells that can be easily associated with a specific respondent, or a limited subset thereof), and in protecting them. The techniques that can be adopted for identifying and protecting sensitive cells vary depending on the kind of macrodata table (i.e., count, frequency, or magnitude table). A well-known technique used to identify sensitive cells in every kind of macrodata tables is the *threshold rule*. According to this rule, a cell is considered sensitive if the number of respondents contributing to the value of that cell is lower than a given threshold [7]. For instance, consider Figure 1 and suppose that the threshold is set to 3. In this case, the second cell in the first row and the last cell in the second row of the table are sensitive, since only 2 respondents contribute to their value. A technique specifically designed to identify sensitive cells in magnitude tables is the *(n,k)-rule*. According to this rule, a cell is considered sensitive if less than n respondents contribute to more than k% of the total cell value [7]. Sensitive cells can be protected by applying *primary suppression*, which consists in not releasing their value. Since genuine (i.e., not modified) marginal totals are often released, a data recipient can however compute an interval that contains the value of the suppressed cells, or even determine their exact value. To prevent this leakage of information, additional cells may need to be suppressed (*secondary suppression*). Linear programming techniques can then be used to minimize the number of cells suppressed. Besides primary and secondary suppression, other solutions can be applied to protect sensitive cells, such as *rounding* (i.e., the value of a sensitive cell is rounded up or down to the nearest multiple of a chosen base value), and *roll up categories* (i.e., a subset of the rows and/or columns in the table are merged to release a less specific macrodata table) [7].

2.2 Microdata

Figure 2 An example of a de-identified microdata table (a) and of a publicly available non de-identified dataset (b)

Microdata consist of records, each containing a set of attributes whose values are related to a single respondent. The release of microdata instead of macrodata presents several advantages for data recipients. Indeed, the availability of the microdata provides higher flexibility, since **the** data recipients can perform any analysis and compute any aggregate function considered of interest. However, it is also more difficult to guarantee respondents' privacy.

The first step for protecting the privacy of the respondents consists in removing or encrypting explicit *identifiers*, that is, those attributes that uniquely identify each respondent in the microdata table (e.g., `SSN`). This de-identification process is however not sufficient to guarantee privacy protection, since the microdata table can contain sets of attributes, called *quasi-identifiers*, whose combination of values uniquely, or almost uniquely, pertain to specific respondents. For instance, from a study on the 2000 census data [10], Golle showed that 63% of the US population was *uniquely identifiable* by their gender, ZIP code, and full date of birth. Since databases associating respondents' identities with quasi-identifying attributes are often publicly available, linking attacks can be exploited for re-identifying respondents or for restricting the uncertainty to a specific subset of respondents [13]. This identity disclosure may also imply the disclosure of sensitive information associated with the respondents (attribute disclosure). As an example, consider the de-identified microdata table in Figure 2a, where attributes `SSN` and `Name` have been removed before publication. The municipality register in Figure 2b includes attributes `DoB`, `Sex`, and `ZIP`, in association with the `Name`, `Address`, `City`, and `Education` of data respondents. The common attributes `DoB`, `Sex`, and `ZIP` can then be exploited by a data recipient to re-identify respondents. In the microdata table in Figure 2a, for example, there is only one male born on 61/09/15 living in the 94142 area. This combination, if unique in the external world as well, uniquely identifies the corresponding tuple in the released microdata as pertaining to "*John Doe, 250 Market Street, San Francisco*" (identity disclosure), thus revealing that he suffers from *stomach cancer* (attribute disclosure).

Depending on their identifying ability and sensitivity, the attributes in a microdata table can be classified in the following four categories.

- *Identifiers*: attributes that uniquely identify a respondent (e.g., `Name` and `SSN`).

- *Quasi-identifiers* (*QI*): attributes that, in combination, can be linked with external information to re-identify (all or some of) the respondents to whom information refers, or to reduce the uncertainty over their identities (e.g., `DoB`, `Sex`, and `ZIP`).
- *Confidential attributes*: attributes that represent sensitive information (e.g., `Disease`).
- *Non-confidential attributes*: attributes that are not considered sensitive by respondents, and whose release is not harmful (e.g., `Race`).

Microdata protection techniques typically ignore the non-confidential attributes since they are not critical and assume basic protection of removing identities to be applied. In the remainder of this paper, we will focus on microdata protection techniques proposed in the literature for counteracting the privacy issues discussed above.

3 Microdata protection

The protection of microdata tables is a complex task, since it has to take into consideration different factors that can contribute to identity and attribute disclosure [7]. Such factors include the existence of highly visible tuples (e.g., tuples with unique characteristics such as a rare disease) and, as already discussed, the possibility of linking the microdata table with external information sources. The ability of linking microdata to other data sources may increase when there is a high number of common attributes between the microdata table and the external sources. By contrast, there are factors that may decrease the risk of identity and attribute disclosure, such as: the natural noise characterizing the microdata table and the external sources; the presence in the microdata table of data that are not up-to-date or that may refer to different temporal intervals; the use of different formats for representing the information in the microdata table and in the external sources. All these factors limit the ability to link information and therefore make the disclosure of information more difficult.

The microdata protection techniques aim at reducing the amount of released information, masking the data (e.g., not releasing or perturbing data values), or at releasing plausible but synthetic values instead of the real ones. In the remainder of this section, we provide a classification of the microdata protection techniques and then present some recent studies based on the k-anonymity concept [13].

3.1 Classification of protection techniques

The microdata protection techniques can be classified in two main categories [4]: *i) masking techniques*, which transform the original set of data, and *ii) synthetic data generation techniques*, which build a new set of data to be released.

Masking techniques. Masking techniques can be further classified in two categories: *non-perturbative* and *perturbative*. Non-perturbative masking techniques do not modify the original data, but remove details from the microdata table. Perturbative masking techniques modify the data before release, removing unique quasi-identifier values and introducing new ones.

Examples of non-perturbative protection techniques are: *sampling*, *local suppression*, *generalization*, and *global recording*. Sampling consists in releasing data that are related to a subset of the original population. Local suppression blanks those cells that are likely to significantly contribute to the identity or attribute disclosure of a respondent (e.g., high visibility cells). Local suppression can also be applied at the granularity level of tuple or column. Generalization replaces the original values of attributes with more general values (e.g., the year of birth is released instead of the complete date of birth). Global recording partitions the domain of an attribute into disjoint intervals, usually of the same width, and associates a label with each interval. Instead of releasing the original values, the labels of the corresponding intervals are published. Two examples of global recording techniques are *top-coding* and *bottom-coding*. Top-coding substitutes the values above a given threshold with a unique label (e.g., annual incomes over 1 million dollars are substituted with label ">1MM"). *Bottom-coding* substitutes the values under a given threshold with a unique label (e.g., annual incomes less than 50 thousand dollars are represented as <50K).

Examples of perturbative protection techniques are *resampling*, *rounding*, and *swapping*. Resampling replaces the values of a numerical attribute with the average, computed on a given number of samples of the original population. Rounding replaces the original value with the nearest multiple of a chosen base. Swapping exchanges the values of a set of attributes between predefined pairs of tuples.

Synthetic data generation techniques. Synthetic data generation techniques replace (a subset of) the original data with synthetic data before release. The generation of synthetic data follows a statistical model that preserves the *key* statistical properties of interest of the original microdata table. No assurance is instead given on the maintenance of the properties that are not explicitly considered by the data generation process. The released dataset may either contain synthetic data only (*fully synthetic* techniques), or a mix of original and synthetic data (*partially synthetic* techniques).

Examples of fully synthetic techniques are *bootstrap* and *Cholesky decomposition*. Bootstrap modifies the *p-variate cumulative distribution function* characterizing the original microdata table. The published tuples are obtained by sampling the new p-variate cumulative distribution function. Cholesky decomposition is based on the decomposition (through the Cholesky method) of the matrix corresponding to the microdata table. The result obtained is then used to compute the matrix representing the synthetic dataset, which is guaranteed to preserve the mean, variance, and co-variance of the original table.

Examples of partially synthetic techniques are *hybrid masking* and *blank-and-impute*. Hybrid masking first generates a synthetic dataset, starting from the original microdata. Each tuple in the original dataset is then matched with a synthetic tuple and their linear combination (e.g., their sum) is released. Blank-and-impute replaces the value of a subset of the cells in the microdata table with new synthetic values, computed by applying a suitable function (e.g., average) to the original values.

3.2 Protecting identities

The microdata protection techniques described above reduce the risk of disclosure of sensitive information. However, these techniques also limit the utility of the published microdata table. The k-anonymity concept [13] has been introduced for characterizing the degree of data protection with respect to inference by linking. k-Anonymity captures the well known requirement applied by statistical agencies, stating that the released data should be indistinguishably related to no less than a certain number of respondents. Since re-identification occurs exploiting quasi-identifiers, the requirement above has been formulated in [13] as follows: *Each release of data must be such that every combination of values of quasi-identifiers can be indistinctly matched to at least k respondents*. Basically, a microdata table satisfies the k-anonymity requirement if each tuple in the table cannot be related to less than k respondents in the population, and vice versa. Clearly, this limits the possibility for a data recipient to exploit external (possibly non de-identified) datasets for the re-identification of the respondents.

In principle, to check if the k-anonymity requirement is satisfied, it is necessary to know all the external sources of information available for linking to any possible data recipient. Since this assumption is impossible in practice, k-anonymity takes a safe approach by requiring that each respondent is indistinguishable from at least $k - 1$ other respondents in the released microdata table. A microdata table is then k-anonymous if each combination of quasi-identifier values in the microdata table appears with at least k occurrences in the table itself. This definition represents a sufficient condition for the k-anonymity requirement. Indeed, if in the released table each quasi-identifying value appears at least with k occurrences, the combination of the released table with the external sources cannot allow the data recipient to associate each released tuple with less than k respondents. For instance, the table in Figure 3 is k-anonymous for any $k \leq 3$, since each quasi-identifier value (i.e., the combinations of values over attributes `DoB`, `Sex`, and `ZIP`) appears with at least 3 occurrences. In the table, attribute `DoB` has been generalized by releasing only the year of birth, while attribute `ZIP` has been generalized by removing the last two digits (replaced by symbol $*$).

Among the non-perturbative microdata protection techniques previously introduced, k-anonymity relies on *generalization* and *suppression*. The combined use of these techniques guarantees the release of a less precise and less complete, but truthful, data while providing protection of respondents' identities. For instance, the table in Figure 3 has been obtained from the table in Figure 2a by generalizing attributes `DoB` and `ZIP`, and by suppressing the sixth tuple in the original table. Generalization and suppression can be applied at different granularity levels [3]; generalization can operate at attribute and cell levels while suppression can operate at attribute, tuple, or cell levels. The combinations of the different choices for generalization and suppression (including also the choice of not applying one of the two techniques) result in different classes of k-anonymity approaches and algorithms [3]. Many k-anonymity algorithms have been proposed in the literature. All these algorithms compute a k-anonymous microdata table and try to minimize the information loss due to the k-anonymization process, while limiting the computational overhead for the data holder (e.g., [13]).

3.3 Protecting sensitive information

Figure 3 An example of a 3-anonymous and 3-diverse table

k-anonymity has been proposed to guarantee respondents privacy against identity disclosure. It is however still possible, from a k-anonymous table, to identify or reduce the uncertainty about the sensitive attribute associated with a specific respondent. Consider a k-anonymous table where all the tuples in a given *equivalence class* (i.e., characterized by the same quasi-identifier value) have the same value for the sensitive attribute as well. Suppose now that a data recipient knows the quasi-identifier of a specific respondent, who is represented in the released microdata table. Since the data recipient can determine the equivalence class where the respondent is represented, she can also infer the sensitive attribute associated with a specific respondent (*homogeneity attack*). To illustrate, suppose that all tuples in the equivalence class with quasi-identifier $\langle$*1964*,*M*,*941*$**\rangle$ in Figure 3 have *cancer* as `Disease`. If *Alice* knows that her neighbor *Bob* is in the dataset and that he is a male, born in *1964*, living in area *94138*, she can infer that *Bob* suffers from *cancer*. We also note that, even if the tuples in an equivalence class are not homogeneous with respect to the sensitive attribute, some privacy issues can still arise. If the data recipient has some additional (external) knowledge about a given respondent, she can reduce her uncertainty about the respondent's sensitive attribute (*external knowledge attack*). As an example, consider the equivalence class $\langle$*1964*,*M*,*941*$**\rangle$ in Figure 3 and suppose that two tuples in the equivalence class have *cancer* as `Disease`, and the third tuple has *broken leg* as `Disease`. If *Alice* knows that *Bob* does not have a broken leg since she saw him running (external knowledge), she can still infer that he suffers from *cancer*.

To limit attribute disclosure, Machanavajjhala et al. [12] proposed ℓ-diversity, which extends k-anonymity by requiring that every equivalence class has at least ℓ *well represented values* for the sensitive attribute. The authors proposed several definitions of "well represented" values. In the simplest formalization of ℓ-diversity requirement, each equivalence class in the microdata table must have at least ℓ distinct values for its sensitive attribute. ℓ-diversity counteracts homogeneity attack and limits the effectiveness of external knowledge attack. For instance, consider the table in Figure 3, which is 3-anonymous and 3-diverse. *Alice* knows that *Bob* is represented in the table and that he belongs to the first equivalence class. Since *Alice* only knows that *Bob* does not have a broken leg (external knowledge), she can infer that *Bob* has 50% of probability of suffering from *chest pain* or *gastritis*, against the 33% that occurred without the external knowledge.

Although ℓ-diversity represents a first step to prevent attribute disclosure, it might be not sufficient. As an example, consider again the equivalence class $\langle$*1964*,*M*,*941*$**\rangle$ in Figure 3, and suppose that there are two tuples with *NVCJD* as `Disease`, and the third tuple has *broken leg* as `Disease`. Suppose that *NVCJD* is a rare disease that occurs with a low probability. Since the considered equivalence class has two out of three tuples with this disease, *Alice* can infer that *Bob* has 67% probability of suffering from *NVCJD*, meaning that people in the considered equivalent class have higher probability of suffering from this rare disease. The concept of t-closeness [11] has been introduced to prevent this attack. In particular, t-closeness requires the distribution of values for the sensitive attribute in each equivalence class to be similar to the distribution of values for the same property in the population.

In addition to ℓ-diversity and t-closeness, other proposals have been also developed, addressing different aspects of the privacy problem.

4 Research challenges

The problem of preventing both identity and attribute disclosures in microdata release is today receiving wide attention. However, there are still different research challenges that need to be further investigated. We briefly discuss the most important ones.

Privacy and utility metrics. As already noted, the adoption of microdata protection techniques needs to balance two contrasting needs: the need of the data recipient for complete and detailed data, and the need of respondents for privacy protection. The k-anonymity approach can help in balancing data utility and disclosure risk since it aims at finding a k-minimal table (i.e., a table where data are not generalized more than it is needed to reach the threshold k). There is however still the need of defining precise metrics that can be adopted with any microdata protection technique. It is also important to note that data utility highly depends on the intended use of the released data. Therefore, it should be considered in the definition of precise data utility metrics.

External knowledge. Datasets released by competing companies representing similar information, social networks, and personal knowledge are some examples of possible external information sources that can be exploited by a data recipient to determine (or reduce her uncertainty on) the identity of data respondents and/or the sensitive attributes associated with them. Although recent approaches (e.g., [1]) take external knowledge into consideration in microdata publication, they do not represent a comprehensive solution to the problem. As an example, both ℓ-diversity and t-closeness only consider a specific kind of external knowledge (i.e., the personal knowledge of a respondent and the distribution of sensitive values in the population). Further research is still needed to permit a complete modeling of the external knowledge and for considering it in the process of computing the table to be released.

Data dependencies. The microdata protection process should consider the existence of possible dependencies and correlations among published data, which can be exploited for inferring sensitive information that has not been published. For instance, suppose that attribute `Salary` has been removed from a microdata table before publishing and that attribute `Taxes` has been released. Since there is a dependency between the two attributes, a recipient can immediately infer the salary of a respondent by knowing the taxes she annually pays. Approaches should therefore be devised that take into consideration data dependencies when computing the tables to be released.

Longitudinal data. Most microdata protection techniques typically consider static collections of data. There are however data that are often *longitudinal* by nature. Longitudinal data are repeated observations of the same respondents that are published at different points in time. Longitudinal data are, for example, tables containing information about multiple patients' visits over a period of time. Protecting respondents' privacy in longitudinal data poses new privacy issues. In fact, anonymizing each version of the dataset independently from the others does not provide sufficient guarantees, since information referred to the same respondent appears in different versions of the data and could possibly be exploited to identify her. Although the scientific community has started to study the protection of longitudinal data, a robust solution is still missing.

Multiple views. Microdata protection techniques are based on the assumption that there is only one table that needs to be released, where each tuple corresponds to a respondent. A data holder may however need to publish different views of the same microdata table. These views may then be exploited for inferring information that is not intended for disclosure. The problem of releasing different views providing anonymity, even in presence of joins that can be used to infer new information, needs to be investigated.

5 Conclusions

Nowadays, information is probably the most important and demanded resource. Organizations in public as well as private sectors collect, share, and disseminate huge collections of data that often contain personal information that needs to be properly protected. This situation has led to growing concerns about the privacy of the respondents to whom the information refers and a wide variety of data protection techniques has been proposed in the literature. In this paper, we first illustrated the main privacy issues arising in the publication of macrodata and microdata collections. We then described microdata protection techniques, presenting a classification of known methods, and then focusing on solutions aimed at preventing identity and attribute disclosure. We also illustrated some research challenges that still need to be investigated.

Acknowledgments This work was supported in part by the EU within the 7FP project "PrimeLife" under grant agreement 216483 and by the Italian Ministry of Research within the PRIN 2008 project "PEPPER" (2008SY2PH4).

References

[1] B.-C. Chen, K. LeFevre, and R.Ramakrishnan. Privacy skyline: Privacy with multidimensional adversarial knowledge. In *Proc. of VLDB 2007*, Vienna, Austria, September 2007.

[2] S. Cimato, M. Gamassi, V. Piuri, R. Sassi, and F. Scotti. Privacy-aware biometrics: Design and implementation of a multimodal verification system. In *Proc. of ACSAC 2008*, Anaheim, USA, Dec 2008.

[3] V. Ciriani, S. De Capitani di Vimercati, S. Foresti, and P. Samarati. k-Anonymity. In T. Yu and S. Jajodia, editors, *Secure Data Management in Decentralized Systems*. Springer-Verlag, 2007.

[4] V. Ciriani, S. De Capitani di Vimercati, S. Foresti, and P. Samarati. Microdata protection. In T. Yu and S. Jajodia, editors, *Secure Data Management in Decentralized Systems*. Springer-Verlag, 2007.

[5] V. Ciriani, S. De Capitani di Vimercati, S. Foresti, and P. Samarati. k-anonymous data mining: A survey. In C.C. Aggarwal and P.S. Yu, editors, *Privacy-Preserving Data Mining: Models and Algorithms*. Springer-Verlag, 2008.

[6] C. Dwork. Differential privacy. In *Proc. of ICALP 2006*, Venice, Italy, 2006.

[7] Federal Committee on Statistical Methodology. *Statistical policy working paper 22*. USA, May 1994. Report on Statistical Disclosure Limitation Methodology.

[8] M. Gamassi, M. Lazzaroni, M. Misino, V. Piuri, D. Sana, and F. Scotti. Accuracy and performance of biometric systems. In *Proc. of IMTC 2004*, Como, Italy, 2004.

[9] M. Gamassi, V. Piuri, D. Sana, and F. Scotti. Robust fingerprint detection for access control. In *Proc. of RoboCare Workshop 2005*, Rome, Italy, May 2005.

[10] P. Golle. Revisiting the uniqueness of simple demographics in the us population. In *Proc. of WPES 2006*, Alexandria, VA, USA, October 2006.

[11] N. Li, T. Li, and S. Venkatasubramanian. t-closeness: Privacy beyond k-anonymity and ℓ-diversity. In *Proc. of ICDE 2007*, Istanbul, Turkey, April 2007.

[12] Ashwin Machanavajjhala, Daniel Kifer, Johannes Gehrke, and Muthuramakrishnan Venkitasubramaniam. ℓ-diversity: Privacy beyond k-anonymity. *ACM Transactions on Knowledge Discovery from Data*, 1(1):3, 2007.

[13] P. Samarati. Protecting respondents' identities in microdata release. *IEEE Transactions on Knowledge and Data Engineering*, 13(6):1010–1027, 2001.

	Ulcera	**Cancer**	**Gastritis**	**Broken bone**	**Total**
A	0	2	5	4	11
B	8	7	10	2	27
C	5	0	7	0	12
Total	13	9	22	6	50

Figure 1: An example of count table reporting, for each region, the number of patients with a given disease

SSN	**Name**	**DoB**	**Sex**	**ZIP**	**Disease**
		64/09/27	M	94139	Chest pain
		63/09/30	F	94139	Broken arm
		64/04/18	M	94139	Gastritis
		63/04/15	F	94139	Ulcera
		63/03/13	F	94138	Short breath
		64/09/15	*M*	*94142*	*Stomach cancer*
		64/09/13	M	94141	Broken leg

(a) De-identified medical data

Name	**Address**	**City**	**ZIP**	**BirthDate**	**Sex**	**Education**
...	...	...	...	...	...	...
John Doe	250 Market St.	San Francisco	94142	64/09/15	male	secondary
...	...	...	...	...	...	...

(b) Municipality register

Figure 2: An example of a de-identified microdata table (a) and of a publicly available non de-identified dataset (b)

DoB	**Sex**	**ZIP**	**Disease**
1964	M	941**	Chest pain
1964	M	941**	Gastritis
1964	M	941**	Broken leg
1963	F	941**	Broken arm
1963	F	941**	Ulcera
1963	F	941**	Short breath

Figure 3: An example of a 3-anonymous and 3-diverse table