Privacy in Microdata Release: Challenges, Techniques, and Approaches

Giovanni Livraga

1 Introduction

We live in a society that relies more and more on the availability of data to make knowledge-based decisions (Livraga, 2015). The benefits that can be driven by data sharing and dissemination have been widely recognized for a long time now (Foresti, 2011; Livraga, 2015), and are visible to everybody: for instance, medical research is a simple example of a field that, leveraging analysis of real clinical trials made available by hospitals, can improve the life quality of individuals. At the same time, many laws and regulations have recognized that privacy is a primary right of citizens, acknowledging the principle that sensitive information (e.g., personal information that refers to an individual) must be protected from improper disclosure. To resolve the tension between the (equally strong) needs for data privacy and availability, the scientific community has been devoting major efforts for decades to investigating models and approaches that can allow a data owner to release a data collection guaranteeing that sensitive information be properly protected, while still allowing useful analysis to be performed (Bezzi et al., 2012; De Capitani di Vimercati et al., 2011b).

In the past, data were typically released in the form of aggregate statistics (*macrodata*): while providing a first layer of protection to the individuals to whom the statistics pertain, as no specific data of single respondents (i.e., the individuals to whom data items refer) are (apparently) disclosed (De Capitani di Vimercati et al., 2011a), releasing precomputed statistics inevitably limits the analysis that a recipient can do. To provide recipients with greater flexibility in performing analysis, many situations require the release of detailed data, called *microdata*.

G. Livraga (✉)
Dipartimento di Informatica, Università degli Studi di Milano, Crema, Italy
e-mail: giovanni.livraga@unimi.it

N. Crato, P. Paruolo (eds.), *Data-Driven Policy Impact Evaluation*,
https://doi.org/10.1007/978-3-319-78461-8_5

Indeed, since analyses are not precomputed, more freedom is left to the final recipients. The downside, however, comes in terms of major privacy concerns, as microdata can include sensitive information precisely related to individuals.

As will be illustrated in this chapter, the first attempts towards the development of microdata protection approaches pursued what today are typically called *syntactic* privacy guarantees (Ciriani et al., 2007a; Clifton and Tassa, 2013; De Capitani di Vimercati et al., 2012). Traditional protection approaches (e.g., k-anonymity (Samarati, 2001) and its variations) operate by removing and/or generalizing (i.e., making less precise/more general) all information that can identify a respondent, so that each respondent is hidden in a group of individuals sharing the same identifying information. In this way, it is not possible to precisely link an individual to her (sensitive) information. Existing solutions following this approach can be used to protect respondents' identities as well as their sensitive information (Livraga, 2015), also in emerging scenarios (De Capitani di Vimercati et al., 2015b). Alternative approaches based on the notion of differential privacy (Dwork, 2006) have then been proposed. Trying to pursue a relaxed and microdata-adapted version of a well-known definition of privacy by Dalenius (1977), that anything that can be learned about a respondent from a statistical database should be learnable without access to the database, differential privacy aims at ensuring that the inclusion in a dataset of the information of an individual does not significantly alter the outcome of analysis of the dataset. To achieve its privacy goal, differential privacy typically relies on controlled noise addition, thus perturbing the data to be released (in contrast to k-anonymity-like solutions that, operating through generalization, guarantee data truthfulness). There has been a major debate in the scientific community regarding which approach (syntactic techniques versus differential privacy) is the "correct" one (Clifton and Tassa, 2013; Kifer and Machanavajjhala, 2011), and recent studies have pointed out that, while they pursue different privacy goals through different protection techniques, both approaches are successfully applicable to different scenarios, and there is room for both of them (Clifton and Tassa, 2013; Li et al., 2012a), possibly jointly adopted (Soria-Comas et al., 2014). Both the approaches have in fact been used in different application scenarios, ranging from the protection of location data (e.g., Peng et al. 2016; Xiao and Xiong 2015), to privacy-preserving data mining (e.g., Ciriani et al. 2008; Li et al. 2012c), and to the private analysis of social network data (e.g., Tai et al. 2014; Wang et al. 2016), just to name a few.

The goal of this chapter is to illustrate some of the best-known protection techniques and approaches that can be used to ensure microdata privacy. The remainder of this chapter is organized as follows. Section 2 presents the basic concepts behind the problem of microdata protection, illustrating possible privacy risks and available protection techniques. Section 3 discusses some well-known protection approaches. Section 4 illustrates some extensions of the traditional approaches, proposed to relax or remove some assumptions for use in advanced scenarios, with a specific focus on the problem of protecting microdata coming from multiple sources. Finally, Sect. 5 concludes the chapter.

2 Microdata Protection: Basic Concepts

This section illustrates the key concepts behind the problem of protecting microdata privacy. It discusses firstly some privacy issues that can arise in microdata release (Sect. 2.1), and secondly the protection techniques that have been proposed by the research community to protect microdata (Sect. 2.2).

2.1 Microdata Privacy

Microdata can be represented as relational tables including a set of tuples, related to a set of individuals (called *respondents*), and defined over a set of attributes. Traditional data protection approaches classify attributes in a microdata table depending on their identifying ability and sensitivity, as follows (Ciriani et al., 2007a).[1]

- *Identifiers*: attributes that uniquely identify a respondent (e.g., `Name` and `SSN`).
- *Quasi-identifiers* (QI): attributes that, in combination, can be linked to external information to reidentify (all or some of) the respondents to whom information refers, or to reduce the uncertainty over their identities (e.g., `DoB`, `Sex`, and `ZIP`).
- *Sensitive attributes*: attributes that represent information that should be kept confidential (e.g., `Disease`).

The first step in protecting a microdata table to be released is to remove (e.g., by deleting or encrypting) all identifiers from the table. This process, usually referred to as *de-identification*, is unfortunately not sufficient to effectively ensure the *anonymity* of the data, due to the presence of QI attributes (e.g., 63% of the entire US population in the US 2000 Census was *uniquely identifiable* by the combination of their gender, ZIP code, and full date of birth (Golle et al., 2006)). To illustrate, consider the de-identified version of a microdata table including information on a set of hospitalized patients in Fig. 1a. Figure 1b illustrates a sample excerpt of a (fictitious) publicly available voter list for the municipality of New York City. Attributes `DoB`, `Sex`, and `ZIP` can be used to link the two tables, allowing the re-identification (with either full confidence or a certain probability) of some of the de-identified respondents in Fig. 1a. For instance, the de-identified microdata include only one female respondent, born in 1958/12/11 and living in the 10180 area (tuple 11). If this combination of QI values is unique in the external world as well, the voter list can be exploited to uniquely reidentify the eleventh tuple with respondent *Kathy Doe*, also disclosing the fact that she has been hospitalized for

[1] In this chapter, `SSN`, `DoB`, and `ZIP` are attributes representing Social Security Numbers (the de facto US identification number for taxation and other purposes), dates of birth, and ZIP codes (US postal codes).

SSN	Name	DoB	Sex	ZIP	Disease
		1960/05/02	F	10041	stroke
		1960/05/20	M	10032	dyspepsia
		1960/05/12	M	10037	achlorhydria
		1960/05/05	F	10044	epilepsy
		1955/09/01	M	10043	helicobacter
		1955/09/02	M	10042	helicobacter
		1955/09/10	F	10039	helicobacter
		1955/09/20	F	10030	helicobacter
		1955/12/07	M	10030	dermatitis
		1955/12/05	M	10031	retinitis
		1958/12/11	*F*	*10180*	epilepsy
		1955/12/25	F	10042	dermatitis
		1955/12/30	F	10045	gastritis
		1960/04/02	F	10036	stroke
		1960/04/05	F	10034	labyrinthitis
		1960/04/10	M	10047	gastritis
		1960/04/30	M	10048	dyspepsia

(a)

Name	Address	City	ZIP	DoB	Sex	Education
...	...	...	...	...	...	...
Kathy Doe	300 Main St.	New York City	*10180*	*58/12/11*	*female*	secondary
...	...	...	...	...	...	...

(b)

Fig. 1 An example of a de-identified microdata table (**a**) and of a publicly available non-de-identified dataset (**b**)

epilepsy. Given that tremendous amounts of data are generated and shared every day, the availability of non-de-identified datasets that can be used for linking is a realistic threat. Unfortunately, unlike direct identifiers, QI cannot be easily removed to protect privacy, since QI attributes can represent a large portion of the attributes in the table, and their complete removal would reduce the utility of the anonymized data too much (e.g., removing also the QI from the de-identified microdata in Fig. 1a would leave only a list of diseases, most probably of limited interest to the final recipients).

Given a de-identified microdata table, two different kinds of improper disclosure can occur, as follows (Federal Committee on Statistical Methodology, 2005).

– *Identity disclosure*, occurring whenever the identity of a respondent can be somehow determined and associated with a (de-identified) tuple in the released microdata table.

– *Attribute disclosure*, occurring when a (sensitive) attribute value can be associated with an individual (without necessarily being able to link the value to a specific tuple).

2.2 Protection Techniques

Various *microdata protection techniques* have recently been proposed by the scientific community (Ciriani et al., 2007b; Federal Committee on Statistical Methodology, 2005). An initial distinction can be made between *masking techniques* and *synthetic data generation techniques*: while these latter aim to release a new, synthetic dataset that preserves some statistical properties of the original data, masking techniques operate directly on the original microdata, to sanitize them before release, and can be classified as follows.

– *Non-perturbative techniques* do not directly modify the original data, but remove details from the microdata table: they sacrifice data completeness by releasing possibly imprecise and/or incomplete data to preserve data truthfulness. Examples of non-perturbative techniques include *suppression*, *generalization*, and *bucketization*. Suppression selectively removes information from the microdata table. Generalization, possibly based on ad hoc generalization hierarchies, selectively replaces the content of some cells in the microdata table (e.g., a complete date of birth) with more general values (e.g., year of birth). Bucketization operates on sets of attributes whose joint visibility should be prevented (e.g., the name and the disease of a patient), and operates by first partitioning tuples in buckets and attributes in groups, and then shuffling the semi-tuples within buckets so as to break their correspondence (De Capitani di Vimercati et al., 2015a, 2010; Li et al., 2012b; Xiao and Tao, 2006).
– *Perturbative techniques* distort the microdata table to be released by modifying its informative content, hence sacrificing data truthfulness. Examples of perturbative techniques include *noise addition* and *microaggregation*. Noise addition intuitively adds controlled noise to the original data collection. Protection is provided by the fact that some values (or combinations among them) included in the released table might not correspond to real ones, and vice versa. Microaggregation (originally proposed for continuous numerical data and then extended also to categorical data (Torra, 2004)) selectively replaces original tuples with new ones. It operates by first clustering the tuples in the original microdata table in groups of a certain cardinality in such a way that tuples in the same cluster are similar to each other, and then by replacing the tuples in a cluster with a representative one computed through an aggregation operator (e.g., mean or median).

The protection techniques illustrated above can be adopted to effectively protect the confidentiality of a microdata collection to be released. Given a data collection to be protected and released, some key questions then need to be answered: what

technique should be used? Should a combination of techniques be preferred to a single one? To which portion of the data (e.g., the entire table, a subset of tuples, and a subset of attributes) should the technique be applied? Whatever the answers to these questions, an important observation is that all microdata protection techniques cause an inevitable information loss: non-perturbative techniques produce datasets that are not as complete or as precise as the originals, and perturbative techniques produce datasets that are distorted. For these reasons, the scientific community has recently developed protection approaches that, given a privacy requirement to be satisfied (e.g., the protection of the identities of the microdata respondents), rely on a controlled adoption of some of these microdata protection techniques to protect privacy while limiting information loss, as illustrated in the remainder of this chapter.

3 Microdata Protection Approaches

This section illustrates the most important protection approaches that have driven research in microdata protection in the past couple of decades, together with the privacy requirements they pursue and the microdata protection techniques (see Sect. 2) that are typically adopted for their enforcement.

3.1 k-Anonymity

The first and pioneering approach for protecting microdata against identity disclosure is represented by k-anonymity (Samarati, 2001), enforcing a protection requirement typically applied by statistical agencies that demands that any released information be *indistinguishably related* to no less than a certain number k of respondents. Following the assumption that re-identification of de-identified microdata takes advantage of QI attributes, such general requirement is translated into the k-anonymity requirement: each release of data must be such that every *combination of values of the QI* can be indistinctly matched to *at least k respondents* (Samarati, 2001). A microdata table satisfies the k-anonymity requirement iff each tuple cannot be related to less than k individuals in the population, and vice versa (i.e., each individual in the population cannot be related to less than k tuples in the table). These two conditions hold since the original definition of k-anonymity assumes that each respondent is represented by at most one tuple in the released table and vice versa (i.e., each tuple includes information related to one respondent only).

Verifying the satisfaction of the k-anonymity requirement would require knowledge of *all* existing external sources of information that an adversary might use for the linking attack. This assumption is indeed unrealistic in practice, and therefore k-anonymity takes the safe approach of requiring that each respondent be indistinguishable from at least $k - 1$ other respondents in the released microdata.

A table is therefore said to be k-anonymous if each combination of values of the QI appears in it with either zero or at least k occurrences. For instance, the table in Fig. 1a is 1-anonymous if we assume the QI to be composed of `DoB`, `Sex`, and `ZIP`, since at least one combination of their values (i.e., $\langle 1958/12/11, \text{F}, 10180\rangle$) appears only once in the table (i.e., in the eleventh tuple). Since each combination of QI values is shared by at least k different tuples in the microdata table, each respondent cannot be associated with fewer than k tuples in the released table and vice versa, also satisfying the original k-anonymity requirement (being the definition of a k-anonymous table a sufficient, though not necessary, condition for the satisfaction of the k-anonymity requirement).

Traditional approaches to enforcing k-anonymity operate on QI attributes by modifying their values in the microdata to be released, while leaving sensitive and nonsensitive attributes as they are (recall that direct identifiers are removed from the microdata as the first step). Among the possible data protection techniques that might be enforced on the QI, k-anonymity typically relies on the combined adoption of *generalization* and *suppression*, which have the advantage of preserving data truthfulness when compared to perturbative techniques (e.g., noise addition; see Sect. 2.2). Suppression is used to couple generalization, as it can help in reducing the amount of generalization that has to be enforced to achieve k-anonymity; in this way, it is possible to produce more precise (though incomplete) tables. The intuitive rationale is that, if a microdata table includes a limited number of outliers (i.e., QI values with less than k occurrences) that would force a large amount of generalization to satisfy k-anonymity, these outliers could be more conveniently removed from the table, improving the quality of the released data.

Generalization and suppression can be applied at various granularity levels (i.e., generalization at the cell and attribute levels, and suppression at the cell, attribute, and tuple levels), and the combined use of generalization and suppression at different granularity levels produces different classes of approaches to enforcing k-anonymity (Ciriani et al., 2007a). The majority of the approaches available in the literature adopt attribute-level generalization and tuple-level suppression (Bayardo and Agrawal, 2005; LeFevre et al., 2005; Samarati, 2001). Figure 2 illustrates a 4-anonymous table obtained from the microdata in Fig. 1a through attribute-level generalization (`DoB`, `Sex`, and `ZIP` have been generalized by removing the day of birth, sex, and the last two digits of the ZIP code, respectively) and tuple-level suppression (the 11th tuple related to *Kathy* has been suppressed). Cell-level generalization has also been investigated as an approach to producing k-anonymous tables (LeFevre et al., 2006). To reduce the inevitable information loss (the original microdata informative content is either reduced in detail or removed), it is necessary to compute an optimal k-anonymization minimizing generalization and suppression, which has been shown to be an NP-hard problem (Ciriani et al., 2007a), and both exact and heuristic algorithms have been proposed.

As a last remark on k-anonymity, it should be noted that some recent approaches have been proposed to obtain k-anonymity through microaggregation (see Sect. 2.2) (Domingo-Ferrer and Torra, 2005; Soria-Comas et al., 2014). To this end, the QI undergoes microaggregation, so that each combination of

Fig. 2 An example of 4-anonymous table

SSN	Name	DoB	Sex	ZIP	Disease
		1960/05	*	100**	stroke
		1960/05	*	100**	dyspepsia
		1960/05	*	100**	achlorhydria
		1960/05	*	100**	epilepsy
		1955/09	*	100**	helicobacter
		1955/09	*	100**	helicobacter
		1955/09	*	100**	helicobacter
		1955/09	*	100**	helicobacter
		1955/12	*	100**	dermatitis
		1955/12	*	100**	retinitis
		1955/12	*	100**	dermatitis
		1955/12	*	100**	gastritis
		1960/04	*	100**	stroke
		1960/04	*	100**	labyrinthitis
		1960/04	*	100**	gastritis
		1960/04	*	100**	dyspepsia

Fig. 3 An example of a microdata table (**a**) and of a 3-anonymous version of it (**b**) obtained by adopting microaggregation

Age	Disease
20	flu
25	gastritis
30	dermatitis
35	stroke
40	dyspepsia
45	asthma

(a)

Age	Disease
25	flu
25	gastritis
25	dermatitis
40	stroke
40	dyspepsia
40	asthma

(b)

QI values in the original microdata table is replaced with a microaggregated version. Figure 3b illustrates a 3-anonymous version of the microdata in Fig. 3a obtained through microaggregation, assuming `Age` to be the QI, and `Disease` the sensitive attribute. Note that, being microaggregation a perturbative protection technique, k-anonymous tables computed adopting this approach do not preserve data truthfulness.

3.2 ℓ-Diversity and t-Closeness

While k-anonymity represents an effective solution to protect respondent identities, it does not protect against attribute disclosure (Samarati, 2001). A k-anonymous table can in fact still be vulnerable to attacks allowing a recipient to determine with non-negligible probability the sensitive information of a respondent, as follows (Machanavajjhala et al., 2007; Samarati, 2001).

– *Homogeneity attack*. A homogeneity attack occurs when all the tuples in an equivalence class (i.e., the set of tuples with the same value for the QI) in a k-anonymous table assume the same value for the sensitive attribute. If a data recipient knows the QI value of a target individual x, she can identify the equivalence class representing x, and then discover the value of x's sensitive attribute. For instance, consider the 4-anonymous table in Fig. 2 and suppose that a recipient knows that *Gloria* is a female living in the 10039 area and born on 1955/09/10. Since all the tuples in the equivalence class with QI value equal to $\langle 1955/09, *, 100 * * \rangle$ assume value *helicobacter* for attribute `Disease`, the recipient can infer that Gloria suffers from a *helicobacter* infection.
– *External knowledge attack*. The external knowledge attack occurs when the data recipient possesses some additional knowledge (not included in the k-anonymous table) about a target respondent x, and can use it to reduce the uncertainty about the value of x's sensitive attribute. For instance, consider the 4-anonymous table in Fig. 2 and suppose that a recipient knows that a neighbor, *Mina*, is a female living in the 10045 area and born on 1955/12/30. Observing the 4-anonymous table, the recipient can infer only that the neighbor suffers from *dermatitis*, *retinitis*, or *gastritis*. Suppose now that the recipient sees *Mina* tanning without screens at the park every day: due to this external information, the recipient can exclude the likelihood that *Mina* suffers from *dermatitis* or *retinitis*, and infer that she suffers from *gastritis*.

The original definition of k-anonymity has been extended to ℓ-diversity to counteract these two forms of attack. The idea behind ℓ-diversity is to take into account the values of the sensitive attributes when clustering the original tuples, so that at least ℓ *well-represented* values for the sensitive attribute are included in each equivalence class (Machanavajjhala et al., 2007). While several definitions for "well-represented" values have been proposed, the simplest formulation of ℓ-diversity requires that each equivalence class be associated with at least ℓ different values for the sensitive attribute. For instance, consider the 4-anonymous and 3-diverse table in Fig. 4 and suppose that a recipient knows that a neighbor, *Mina*, a female living in the 10045 area and born on 1955/12/30, tans every day at the park (see example above). The recipient can now only exclude value *dermatitis*, but she cannot be sure about whether *Mina* suffers from *gastritis* or a *helicobacter* infection.

Computing an ℓ-diverse table minimizing the loss of information caused by generalization and suppression is computationally hard. However, since ℓ-diversity basically requires computing a k-anonymous table (with additional constraints on the sensitive values), any algorithm proposed for computing a k-anonymous table that minimizes loss of information can be adapted to also guarantee ℓ-diversity, simply by controlling whether or not the condition on the diversity of the sensitive attribute values is satisfied by all the equivalence classes (Machanavajjhala et al., 2007). As a last remark on ℓ-diversity, it might be possible to obtain ℓ-diverse tables by departing from generalization and adopting instead a bucketization-based approach (see Sect. 2.2), for instance, by adopting the Anatomy approach (Xiao and

Fig. 4 An example of 4-anonymous and 3-diverse table

SSN	Name	DoB	Sex	ZIP	Disease
		1955	M	100**	helicobacter
		1955	M	100**	helicobacter
		1955	M	100**	dermatitis
		1955	M	100**	retinitis
		1960	F	100**	stroke
		1960	F	100**	epilepsy
		1960	F	100**	stroke
		1960	F	100**	labyrinthitis
		1955	F	100**	helicobacter
		1955	F	100**	helicobacter
		1955	F	100**	dermatitis
		1955	F	100**	gastritis
		1960	M	100**	dyspepsia
		1960	M	100**	achlorhydria
		1960	M	100**	gastritis
		1960	M	100**	dyspepsia

Tao, 2006), or other (possibly more general) techniques (Ciriani et al., 2012; De Capitani di Vimercati et al., 2014, 2015a, 2010).

Although ℓ-diversity represents a first step in counteracting attribute disclosure, an ℓ-diverse table might still be vulnerable to information leakage caused by *skewness attacks* (where significant differences can be seen in the frequency distribution of the sensitive values within an equivalence class with respect to that of the same values in the overall population), and *similarity attacks* (where the ℓ sensitive values of the tuples in an equivalence class are semantically similar, although syntactically different) (Li et al., 2007). To counteract these two disclosure risks, it is possible to rely on the definition of t-closeness (Li et al., 2007), requiring that the frequency distribution of the sensitive values in each equivalence class be close (i.e., with distance smaller than a fixed threshold t) to that in the released microdata table.

3.3 *Differential Privacy*

Differential privacy (DP) is a recent privacy definition that departs from the guarantees and enforcement techniques characterizing k-anonymity and its extensions, and aims to guarantee that the release of a dataset does not disclose sensitive information about *any* individual, who may or may not be represented therein (Dwork, 2006). DP aims at releasing a dataset permitting the disclosure of *properties* about the population as a whole (rather than the microdata themselves), while protecting the privacy of single individuals. The privacy guarantee provided by DP relies on ensuring that the probability of a recipient correctly inferring the sensitive value of

a target respondent x be not affected by the presence or absence of x's tuple in the released dataset.

DP can be adopted either to respond to queries (*interactive scenario*) issued against a microdata table or to produce a sanitized dataset to be released (*noninteractive scenario*). In the interactive scenario, DP is ensured by adding *random noise* to the query results evaluated on the original dataset (Dwork et al., 2006), sacrificing data truthfulness. Unfortunately, the interactive scenario limits the analysis that the recipient can perform, as it allows only a limited number of queries to be answered (Soria-Comas et al., 2014). In the noninteractive scenario, a dataset is produced and released, typically based on the evaluation of histogram queries (i.e., counting the number of records having a given value). To reduce information leakage, these counts are computed through a DP mechanism.

Unlike k-anonymity and its variations, which guarantee a certain degree of privacy to the microdata to be released, DP aims to guarantee that the *release mechanism* $\mathcal{K}$ (e.g., the algorithm adopted to compute the data to be released, whether query answers in the interactive scenario or sanitized counts in the noninteractive scenario) is safe with respect to privacy breaches. A dataset to be released satisfies DP if the removal/insertion of one tuple from/to the dataset does not significantly affect the result of the evaluation of $\mathcal{K}$. In this way, the protection offered by DP lies in the fact that the impact that a respondent has on the outcome of a certain analysis (or on the generation of the sanitized dataset) remains negligible. In fact, DP guarantees that the probability of observing a result for the evaluation of $\mathcal{K}$ over T is close to the probability of observing that result for the evaluation of $\mathcal{K}$ over a dataset T' differing from T for a tuple only.

DP offers strong privacy guarantees at the price of imposing strict conditions on what kind of, and how, data can be released (Clifton and Tassa, 2013). In addition, the amount of noise that needs to be adopted can significantly distort the released data (Clifton and Tassa, 2013; Fredrikson et al., 2014; Soria-Comas et al., 2014), thus limiting in practice their utility for final recipients. Some relaxations of DP have therefore been proposed (e.g., Dwork and Smith 2009; Mironov et al. 2009), possibly applicable to specific real-world scenarios (e.g., Hong et al. 2015), with the aim of finding a reasonable tradeoff between privacy protection and data utility.

It is interesting to note that a recent approach has been proposed using k-anonymity and DP approaches together, with the aim of reducing the amount of noise needed to ensure DP (Soria-Comas et al., 2014). The proposal builds on the observation that, given a microdata table T and a query q for which the outputs are required to be differentially private, if the query is run on a microaggregation-based (see Sect. 3.1) k-anonymous version T_k of T, the amount of noise to be added to the output of q for achieving DP is greatly reduced (compared with the noise that would be needed if q were run on the original T). To this end, microaggregation should be performed carefully so that it can be considered *insensitive* to the input data (i.e., for any pair of datasets T and T' differing by one tuple, given the clusters $\{c_1, \ldots, c_n\}$ produced by the microaggregation over T and the clusters $\{c'_1, \ldots, c'_n\}$ produced by the microaggregation over T', each pair of corresponding clusters differs in at most one tuple). This is a key property required for the microaggregation to succeed

in reducing the noise that will then be employed to ensure DP, as it reduces the sensitivity of the query to be executed (Soria-Comas et al., 2014) and hence the result distortion. This approach can also be used in the noninteractive scenario. To this end, a k-anonymous version T_k of T is first built through an insensitive microaggregation. The differentially private dataset T_{DP} is then built by collating the n differentially private answers to a set of n queries (with n the number of tuples in T_k), where the ith query ($i = 1, \ldots, n$) aims at retrieving the ith tuple in T_k.

4 Extensions for Advanced Scenarios

The traditional microdata protection approaches in the literature (see Sect. 3) are built on specific assumptions that can limit their applicability to certain scenarios. For instance, they assume the data to be released in a single table, completely available for anonymization before release, and never republished. However, it may happen that data are either republished over time or continuously generated, as in the case with data streams: recent proposals (e.g., Fung et al. 2008; Loukides et al. 2013; Shmueli and Tassa 2015; Shmueli et al. 2012; Tai et al. 2014; Xiao and Tao 2007) have extended traditional approaches to deal with these scenarios.

One of the assumptions on which the original formulations of k-anonymity, DP, and their extensions were based is that the microdata to be anonymized are stored in a single table. This assumption represents a limitation in many real-world scenarios, in which the information that needs to be released can be spread across various datasets, and where the privacy goal is that *all* released information be effectively protected. There are two naive approaches that one might think of adopting: join-and-anonymize and anonymize-and-join. The first approach, in which all tables to be released are first joined in a universal relation that is then anonymized by adopting one of the traditional approaches, might not work whenever there is no single subject authorized to see and join all original relations, which might be owned by different authorities. The second approach (i.e., first anonymize each table singularly taken and then release the join among the sanitized versions of all tables) does not guarantee appropriate protection: for instance, if a QI is spread across multiple tables, it could not be effectively anonymized by looking at each relation individually. The scientific community has recently started looking at this problem, and some solutions have been proposed (typically extending k-anonymity and its variations) to address the multiple tables scenario.

A first distinction has to be made depending on whether the multiple tables to be released belong to the same authority (e.g., different relations of a single database) that therefore has a complete view over them, or the tables belong to different authorities, where no subject in the picture has a global view of the entire informative content that needs to be released. In the first scenario, a careful join-and-anonymize approach might do. However, the anonymization has to be performed with extreme care to avoid vulnerability to privacy breaches. For instance, assume n relations, owned by the same authority, to be released together provided that k-

anonymity is satisfied by their join. When computing the join among the n relations, it might be possible that the k-anonymity assumption of one respondent being represented by a single tuple is not satisfied (as different tuples could be related to the same respondent). The risk here is that (some of) the different tuples related to the same individual are "anonymized together": hence, an equivalence class of size k might refer to less than k respondents, violating their privacy despite the relation being apparently k-anonymous. To overcome this issue, MultiR k-anonymity (Nergiz et al., 2007) has been proposed to extend the definition of k-anonymity and ℓ-diversity to multiple relations belonging to a snowflake database schema.

When the relations to be anonymized belong to different authorities, it is clearly not possible to join them beforehand. One might think to first anonymize each relation individually and then join the obtained results on the (anonymized) QI. Unfortunately, this strategy is not trivial: besides possibly exploding in size, the joined tuples could not be used for meaningful analysis, as many tuples in the join would be incorrect (joining over the anonymized QI would join more tuples than using the original values). Some approaches have recently been proposed to address this issue. For instance, distributed k-anonymity (DkA (Jiang and Clifton, 2006)) proposes a distributed framework for achieving k-anonymity. The applicability of this approach is limited to two relations (defined as two views over a global data collection), which can be correctly joined through a 1:1 join on a common key. The framework builds a k-anonymous join of the two datasets, without disclosing any information from one site to the other. In a nutshell, the approach works iteratively in three steps: (1) each data holder produces a k-anonymous version of her own dataset; (2) each data holder checks whether or not joining the obtained k-anonymous datasets would maintain global k-anonymity; and (3) if so, join and release, otherwise go back to step 1 and further generalize the original data. Checking the global anonymity (step 2) is a critical task, as it requires the two parties to exchange their anonymized tables. To avoid information leakage, encryption is adopted and, in this regard, the price to be paid for this approach is in terms of the required encryption and decryption overhead (Jiang and Clifton, 2006; Mohammed et al., 2011). Recent efforts that have recently been devoted to enforce DP in a multi-relational setting (Mohammed et al., 2014) (also focusing on two relations only) should also be highlighted. The solution in Mohammed et al. (2011) instead does not pose assumptions on the number of relations to be joined but requires active cooperation among the parties holding the relations to achieve k-anonymity. In addition, the approach in Mohammed et al. (2011) can be successfully extended to provide privacy beyond k-anonymity (e.g., by ensuring ℓ-diversity). Finally, it should be noted that specific approaches have also been proposed to protect different tables that need to be *sequentially* released (Wang and Fung, 2006).

5 Conclusions

This chapter has addressed the problem of protecting privacy in microdata release. After a discussion of the privacy risks that can arise when microdata need to be shared or disseminated, some of the best-known microdata protection techniques and approaches developed by the scientific community have been illustrated. Some recent extensions of traditional approaches, proposed to fit advanced scenarios, have also been highlighted.

Acknowledgements This paper is based on joint work with Sabrina De Capitani di Vimercati, Sara Foresti, and Pierangela Samarati, whom the author would like to thank. This work was supported in part by the European Commission through the Seventh Framework Programme under grant agreement 312797 (ABC4EU) and through the Horizon 2020 programme under grant agreement 644579 (ESCUDO-CLOUD).

References

Bayardo RJ, Agrawal R (2005) Data privacy through optimal k-anonymization. In: Proceedings of ICDE 2005, Tokyo, April 2005

Bezzi M, De Capitani di Vimercati S, Foresti S, Livraga G, Samarati P, Sassi R (2012) Modeling and preventing inferences from sensitive value distributions in data release. J Comput Secur 20(4):393–436

Ciriani V, De Capitani di Vimercati S, Foresti S, Samarati P (2007) k-anonymity. In: Yu T, Jajodia S (eds) Secure data management in decentralized systems. Springer, Berlin

Ciriani V, De Capitani di Vimercati S, Foresti S, Samarati P (2007) Microdata protection. In: Yu T, Jajodia S (eds) Secure data management in decentralized systems. Springer, Berlin

Ciriani V, De Capitani di Vimercati S, Foresti S, Samarati P (2008) k-Anonymous data mining: a survey. In: Aggarwal C, Yu P (eds) Privacy-preserving data mining: models and algorithms. Springer, Berlin

Ciriani V, De Capitani di Vimercati S, Foresti S, Livraga G, Samarati P (2012) An OBDD approach to enforce confidentiality and visibility constraints in data publishing. J Comput Secur 20(5):463–508

Clifton C, Tassa T (2013) On syntactic anonymity and differential privacy. Trans Data Priv 6(2):161–183

Dalenius T (1977) Towards a methodology for statistical disclosure control. Statistik Tidskrift 15:429–444

De Capitani di Vimercati S, Foresti S, Jajodia S, Paraboschi S, Samarati P (2010) Fragments and loose associations: respecting privacy in data publishing. Proc VLDB Endow 3(1):1370–1381

De Capitani di Vimercati S, Foresti S, Livraga G, Samarati P (2011) Anonymization of statistical data. Inform Technol 53(1):18–25

De Capitani di Vimercati S, Foresti S, Livraga G, Samarati P (2011) Protecting privacy in data release. In: Aldini A, Gorrieri R (eds) Foundations of security analysis and design VI. Springer, Berlin

De Capitani di Vimercati S, Foresti S, Livraga G, Samarati P (2012) Data privacy: definitions and techniques. Int J Uncertainty Fuzziness Knowl Based Syst 20(6):793–817

De Capitani di Vimercati S, Foresti S, Jajodia S, Livraga G, Paraboschi S, Samarati P (2014) Fragmentation in presence of data dependencies. IEEE Trans Dependable Secure Comput 11(6):510–523

De Capitani di Vimercati S, Foresti S, Jajodia S, Livraga G, Paraboschi S, Samarati P (2015) Loose associations to increase utility in data publishing. J Comput Secur 23(1):59–88

De Capitani di Vimercati S, Foresti S, Livraga G, Paraboschi S, Samarati P (2015) Privacy in pervasive systems: social and legal aspects and technical solutions. In: Colace F, Santo MD, Moscato V, Picariello A, Schreiber F, Tanca L (eds) Data management in pervasive systems. Springer, Berlin

Domingo-Ferrer J, Torra V (2005) Ordinal, continuous and heterogeneous k-anonymity through microaggregation. Data Min Knowl Disc 11(2):195–212

Dwork C (2006) Differential privacy. In: Proceedings of ICALP 2006, Venice, July 2006

Dwork C, Smith A (2009) Differential privacy for statistics: what we know and what we want to learn. J Priv Confid 1(2):135–154

Dwork C, Mcsherry F, Nissim K, Smith A (2006) Calibrating noise to sensitivity in private data analysis. In: Proceedings of TCC 2006, New York, NY, March 2006

Federal Committee on Statistical Methodology (2005) Statistical policy working paper 22 (Second Version). Report on statistical disclosure limitation methodology, December 2005

Foresti S (2011) Preserving privacy in data outsourcing. Springer, Berlin

Fredrikson M, Lantz E, Jha S, Lin S, Page D, Ristenpart T (2014) Privacy in pharmacogenetics: an end-to-end case study of personalized warfarin dosing. In: Proceedings of the 23rd USENIX security symposium, San Diego, August 2014

Fung BCM, Wang K, Fu AWC, Pei J (2008) Anonymity for continuous data publishing. In: Proceedings of EDBT 2008, Nantes, March 2008

Golle P (2006) Revisiting the uniqueness of simple demographics in the US population. In: Proceedings of WPES 2006, Alexandria, October 2006

Hong Y, Vaidya J, Lu H, Karras P, Goel S (2015) Collaborative search log sanitization: toward differential privacy and boosted utility. IEEE Trans Dependable Secure Comput 12(5):504–518

Jiang W, Clifton C (2006) A secure distributed framework for achieving k-anonymity. VLDB J 15(4):316–333

Kifer D, Machanavajjhala A (2011) No free lunch in data privacy. In: Proceedings of SIGMOD 2011, Athens, June 2011

LeFevre K, DeWitt D, Ramakrishnan R (2005) Incognito: efficient full-domain k-anonymity. In: Proceedings of SIGMOD 2005, Baltimore, June 2005

LeFevre K, DeWitt D, Ramakrishnan R (2006) Mondrian multidimensional k-anonymity. In: Proceedings of ICDE 2006, Atlanta, April 2006

Li N, Li T, Venkatasubramanian S (2007) t-closeness: privacy beyond k-anonymity and ℓ-diversity. In: Proceedings of ICDE 2007, Istanbul

Li N, Qardaji W, Su D (2012) On sampling, anonymization, and differential privacy or, k-anonymization meets differential privacy. In: Proceedings of ASIACCS 2012, Seoul, May 2012

Li T, Li N, Zhang J, Molloy I (2012) Slicing: a new approach for privacy preserving data publishing. IEEE Trans Knowl Data Eng 24(3):561–574

Li Y, Chen M, Li Q, Zhang W (2012) Enabling multilevel trust in privacy preserving data mining. IEEE Trans Knowl Data Eng 24(9):1598–1612

Livraga G (2015) Protecting privacy in data release. Springer, Berlin
Loukides G, Gkoulalas-Divanis A, Shao J (2013) Efficient and flexible anonymization of transaction data. Knowl Inform Syst. 36(1):153–210
Machanavajjhala A, Kifer D, Gehrke J, Venkitasubramaniam M (2007) ℓ-Diversity: privacy beyond k-anonymity. ACM Trans Knowl Discov from Data 1(1):3:1–3:52
Mironov I, Pandey O, Reingold O, Vadhan S (2009) Computational differential privacy. In: Proceedings of CRYPTO 2009, Santa Barbara, August 2009
Mohammed N, Fung BC, Debbabi M (2011) Anonymity meets game theory: secure data integration with malicious participants. VLDB J 20(4):567–588
Mohammed N, Alhadidi D, Fung BC, Debbabi M (2014) Secure two-party differentially private data release for vertically partitioned data. IEEE Trans Dependable Secure Comput 11(1): 59–71
Nergiz M, Clifton C, Nergiz A (2007) Multirelational k-anonymity. In: Proceedings of ICDE 2007, Istanbul
Peng T, Liu Q, Meng D, Wang G (2017) Collaborative trajectory privacy preserving scheme in location-based services. Inform Sci 387:165–179. Available online
Samarati P (2001) Protecting respondents' identities in microdata release. IEEE Trans Knowl Data Eng 13(6):1010–1027
Shmueli E, Tassa T (2015) Privacy by diversity in sequential releases of databases. Inform Sci 298:344–372
Shmueli E, Tassa T, Wasserstein R, Shapira B, Rokach L (2012) Limiting disclosure of sensitive data in sequential releases of databases. Inform Sci 191:98–127
Soria-Comas J, Domingo-Ferrer J, Sánchez D, Martínez S (2014) Enhancing data utility in differential privacy via microaggregation-based k-anonymity. VLDB J 23(5):771–794
Tai CH, Tseng PJ, Yu PS, Chen MS (2014) Identity protection in sequential releases of dynamic networks. IEEE Trans Knowl Data Eng 26(3):635–651
Torra V (2004) Microaggregation for categorical variables: a median based approach. In: Proceedings of PSD 2004, Barcelona, June 2004
Wang K, Fung B (2006) Anonymizing sequential releases. In: Proceedings of KDD 2006, Philadelphia, August 2006
Wang Q, Zhang Y, Lu X, Wang Z, Qin Z, Ren K (2016) Real-time and spatio-temporal crowd-sourced social network data publishing with differential privacy. IEEE Trans Dependable Secure Comput (in press)
Xiao X, Tao Y (2006) Anatomy: simple and effective privacy preservation. In: Proceedings of VLDB 2006, Seoul, September 2006
Xiao X, Tao Y (2007) m-Invariance: towards privacy preserving re-publication of dynamic datasets. In: Proceedings of SIGMOD 2007, Beijing, June 2007
Xiao Y, Xiong L (2015) Protecting locations with differential privacy under temporal correlations. In: Proceedings of CCS 2015, Denver, October 2015

Giovanni Livraga is an assistant professor at the Computer Science Department of the Università degli Studi di Milano, Italy, where he acquired a PhD in Computer Science. His PhD thesis "Preserving Privacy in Data Release" received the ERCIM STM WG 2015 award for the best PhD thesis on Security and Trust Management in a European University. He has been a visiting researcher at SAP Labs, France, and George Mason University, VA (USA). He has been serving

as PC chair and PC member for several international conferences, and as reviewer for several international journals. His research interests are in the area of data protection, privacy, and security, in release, outsourcing, and emerging scenarios. He has collaborated in several national and international projects on different aspects of information protection.