Protecting Privacy in Data Release

Sabrina De Capitani di Vimercati, Sara Foresti,
Giovanni Livraga, and Pierangela Samarati

Dipartimento di Tecnologie dell'Informazione
Università degli Studi di Milano
Via Bramante 65 - 26013 Crema, Italy
firstname.lastname@unimi.it

Abstract. The evolution of the Information and Communication Technology has radically changed our electronic lives, making information the key driver for today's society. Every action we perform requires the collection, elaboration, and dissemination of personal information. This situation has clearly brought a tremendous exposure of private and sensitive information to privacy breaches.
In this chapter, we describe how the techniques developed for protecting data have evolved in the years. We start by providing an overview of the first privacy definitions (k-anonymity, ℓ-diversity, t-closeness, and their extensions) aimed at ensuring proper data protection against identity and attribute disclosures. We then illustrate how changes in the underlying assumptions lead to scenarios characterized by different and more complex privacy requirements. In particular, we show the impact on privacy when considering multiple releases of the same data or dynamic data collections, fine-grained privacy definitions, generic privacy constraints, and the external knowledge that a potential adversary may exploit for inferring sensitive information. We also briefly present the concept of differential privacy that has recently emerged as an alternative privacy definition.

Keywords: Privacy, microdata protection, data release

1 Introduction

The advancements in the Information and Communication Technology (ICT) have revolutionized our lives in a way that was unthinkable until few years ago. We live in the Globalization era, where everything we need to do is available within "one mouse click". Global infrastructure, digital infrastructure, digital society are only few examples of terms used at different times for concisely referring to our "computer-based" society. The term that better represents the today's society is however *information society* (or *information age*) since the information has a key role in the daily life activities of everyone. Every time we browse Internet, perform an online transaction, fill in forms to, for example, enter contests or participate in online games, and spend our time online in social

networks, information about us is collected, stored, analyzed, and sometimes shared with third parties. Furthermore, public and private companies have often the need of publishing aggregate statistical data (*macrodata*) as well as detailed data (*microdata*) for research or statistical purposes.

The complexity and variety of the today's information society introduce therefore new risks and pose new research challenges. In fact, the vast amount of personal (user-generated) data collected, stored, and processed, the unclear data ownership, and the lack of control of the users on their own data are creating unprecedented risks of privacy breaches. The problem of properly protecting the privacy of the users is clearly not new and has received (and receives) considerable attention from the research and development communities. In the past, the restricted access to information and its expensive processing represented a form of protection that does not hold anymore. In fact, with the rate at which technology is developing, it is now becoming easier and easier to access huge amount of data by using, for example, portable devices (e.g., PDAs, mobile phones) and ubiquitous network resources. Also, the availability of powerful techniques for analyzing and correlating data coming from different information sources makes it simple to infer information that was not intended for disclosure.

It is interesting to observe how the problem of guaranteeing privacy protection is changing over the years, in line with the evolution of the ICT. Data were principally released in the form of macrodata, that is, tables (often of two dimensions), where each cell contains aggregate information about users or companies, called *respondents*. The macrodata protection techniques were principally based on the identification and obfuscation of sensitive cells [11]. With the growing importance and use of microdata, the research community dedicated many efforts in designing microdata protection techniques able to preserve the privacy of the respondents while limiting the *disclosure risks*. Traditionally, the disclosure risks are related to the possibility, for an adversary, to use the microdata for determining confidential information on a specific individual (*attribute disclosure*) or for identifying the presence of an individual in the microdata table itself (*identity disclosure*). To limit the disclosure risks, names, addresses, phone numbers, and other identifying information are removed (or encrypted) from the microdata. For instance, in the microdata table in Figure 1, which contains medical data, the names of the patients as well as their Social Security Numbers are removed, thus obtaining the de-identified medical data in Figure 2(a).

Although a de-identified microdata table apparently protects the identities of the respondents represented in the table, there is no guarantee of anonymity. The de-identified microdata may contain other information, called *quasi-identifier*, such as birth date and ZIP code that in combination can be linked to publicly available information to re-identify individuals. As an example, consider the de-identified medical data in Figure 2(a) and the voter list for the San Francisco area, publicly released by the local municipality, in Figure 2(b). It is easy to see that the values of attributes `DoB`, `Sex`, and `ZIP` can be exploited for linking the tuples in the microdata with the voter list, thus possibly re-identifying individuals and revealing their illnesses. For instance, in the microdata in Fig-

SSN	Name	DoB	Sex	ZIP	Disease
123-45-6789	Diana Smith	1950/06/02	F	94141	H1N1
234-56-7890	Nathan Johnson	1950/06/20	M	94132	Gastritis
345-67-8901	Eric Williams	1950/06/12	M	94137	Dyspepsia
456-78-9012	Liz Jones	1950/06/05	F	94144	Pneumonia
567-89-0123	John Brown	1940/04/01	M	94143	Peptic Ulcer
678-90-1234	Luke Davis	1940/04/02	M	94142	Peptic Ulcer
789-01-2345	Barbara Miller	1940/04/10	F	94139	Peptic Ulcer
890-12-3456	Fay Wilson	1940/04/20	F	94130	Peptic Ulcer
901-23-4567	Anthony Moore	1940/06/07	M	94130	Broken Leg
012-34-5678	Matt Taylor	1940/06/05	M	94131	Short Breath
134-56-7890	Jane Doe	1958/12/11	F	94142	Pneumonia
245-67-8901	Anna Anderson	1940/06/25	F	94142	Broken Leg
356-78-9012	Carol Thomas	1940/06/30	F	94145	Stomach Cancer
467-89-0123	Gabrielle White	1950/05/02	F	94136	H1N1
578-90-1234	Lorna Harris	1950/05/05	F	94134	Flu
689-01-2345	Rob Martin	1950/05/10	M	94147	Stomach Cancer
790-12-3456	Bob Thompson	1950/05/30	M	94148	Gastritis

Fig. 1: An example of microdata table with identifying information

ure 2(a) there is only one female born on *1958/12/11* living in the *94142* area. This combination, if unique in the external world as well, uniquely identifies the corresponding tuple in the table as pertaining to *Jane Doe, 300 Main St., San Francisco*, revealing that she suffers from *Pneumonia*. From a study performed on the data collected for the 2000 US Census, Golle showed that 63% of the US population can be uniquely identified combining their gender, ZIP code, and full date of birth [21]. This percentage decreases if the gender is combined with the County of residence instead of the ZIP code, and with the month/year of birth (see Figure 3).

In the 1990s, several microdata protection techniques were developed [11]. Such techniques can be classified in two main categories: *masking techniques* that transform the original data in a way that some statistical analysis on the original and transformed data produce the same or similar results; *synthetic data generation techniques* that replace the original data with synthetic data that preserve some statistical properties of the original data. Among the microdata protection techniques, k-anonymity [37] is probably one of the most popular, which has inspired the development of algorithms and techniques for both enforcing k-anonymity and for complementing it with other forms of protection (e.g., ℓ-diversity [29], and t-closeness [26]). These techniques are based on the assumptions that: quasi-identifiers are the only attributes that can be used for inferring the respondents to whom information refers; the same microdata table is published only once; and potential adversaries do not have any external knowledge. Clearly, such assumptions do not hold anymore in the today's society, where any information can be used to re-identify anonymous data [33]. Two well-known examples of privacy violations, which testified how ensuring proper privacy protection is becoming a difficult task, are the America OnLine (AOL) and Netflix incidents [3, 32]. AOL is an Internet services and media company that in 2006 released around 20 millions of search records of 650,000 of its customers. To protect the privacy of its customers, AOL de-identified such records

SSN	Name	DoB	Sex	ZIP	Disease
		1950/06/02	F	94141	H1N1
		1950/06/20	M	94132	Gastritis
		1950/06/12	M	94137	Dyspepsia
		1950/06/05	F	94144	Pneumonia
		1940/04/01	M	94143	Peptic Ulcer
		1940/04/02	M	94142	Peptic Ulcer
		1940/04/10	F	94139	Peptic Ulcer
		1940/04/20	F	94130	Peptic Ulcer
		1940/06/07	M	94130	Broken Leg
		1940/06/05	M	94131	Short Breath
		1958/12/11	*F*	*94142*	Pneumonia
		1940/06/25	F	94142	Broken Leg
		1940/06/30	F	94145	Stomach Cancer
		1950/05/02	F	94136	H1N1
		1950/05/05	F	94134	Flu
		1950/05/10	M	94147	Stomach Cancer
		1950/05/30	M	94148	Gastritis

(a) De-identified medical data

Name	Address	City	ZIP	DoB	Sex	Education
...	...	...	...	...	...	...
Jane Doe	300 Main St.	San Francisco	*94142*	*58/12/11*	*female*	secondary
...	...	...	...	...	...	...

(b) Voter list

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

	Date of Birth		
	year	*month, year*	*full date*
ZIP	0.2%	4.2%	63.3%
County	0%	0.2%	14.8%

Fig. 3: Identifiability of the US population in the 2000 US Census data [21]

by substituting personal identifiers with numerical identifiers. A sample of such data is the following:

```
116874 thompson water seal 2006-05-24 11:31:36 1 http://www.thompsonwaterseal.com
116874 knbt 2006-05-31 07:57:28
116874 knbt.com 2006-05-31 08:09:30 1 http://www.knbt.com
117020 texas penal code 2006-03-03 17:57:38 1 http://www.capitol.state.tx.us
117020 homicide in hook texas 2006-03-08 09:47:35
117020 homicide in bowle county 2006-03-08 09:48:25 6 http://www.tdcj.state.tx.us
```

that shows records related to two different users (116874 and 117020) containing the ID, the term(s) used for the search, the timestamp, whether the user clicked on a result, and the corresponding visited website. With these data, two reporters of the New York Times newspaper were able to identify AOL customer no. 4417749 as Thelma Arnold, a 62 yeas old widow living in Lilburn [3]. In the same year, the on-line movies renting service Netflix publicly released 100 millions

records, showing the ratings given by 500,000 users to the movies they rent. The records were released within the "Netflix Prize" competition that offered $1 million to anyone who could improve the algorithm used by Netflix to suggest movies to its customers. Also in this case, records were de-identified by replacing personal identifiers with numerical identifiers. However, some researchers were able to de-anonymize the data by comparing the Netflix data against publicly available ratings on the Internet Movie Database (IMDB). For instance, the release of her movie preferences damaged a lesbian mother since she was re-identified, thus causing the disclosure of her sexual orientation [32].

From the discussion, it is clear that the protection of microdata against improper disclosure is a key issue in today's information society. The main objective of this chapter is then to provide an overview of how the techniques for protecting microdata releases have evolved in the years, according to the evolution and complexity of the scenarios considered. We will start with a description of the first solutions aimed at protecting microdata against identity and attribute disclosures (Section 2). We will then describe recent approaches that removed some of the assumptions characterizing the first proposals. In particular, we will describe novel solutions for supporting: *i)* multiple data releases and the insertion/removal of tuples into/from the released microdata table (Section 3); *ii)* the definition of fine-grained privacy preferences by each data respondent (Section 4); *iii)* generic sensitive associations among released data that need to be kept confidential (Section 5); and *iv)* external knowledge by malicious data recipients that may cause information leakage (Section 6). We will finally illustrate a new privacy notion, called *differential privacy*, that defines when a computation is privacy-preserving (Section 7).

2 Privacy in microdata publishing

Goal of this section is to present the privacy models specifically targeted to the protection of microdata from identity and attribute disclosures (i.e., k-anonymity [37], ℓ-diversity [29], and t-closeness [26]) that have influenced the work performed by the research community in the data protection area during the last two decades. All these privacy models are based on the following assumptions. First, the attributes composing a microdata table are classified in four categories, as follows.

- *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 the respondents and whose release is not harmful (e.g., `FavoriteColor`).

A microdata table is then protected by applying microdata protection techniques [11] that transform the values of the quasi-identifier attributes. Second, there is a one-to-one correspondence between tuples in the table and data respondents. Third, a microdata table is released only once. Fourth, each table is characterized by a unique quasi-identifier.

In the remainder of this section, after a description of k-anonymity, ℓ-diversity, and t-closeness, we also briefly discuss their recent enhancements that remove some of the assumptions on which they are based.

2.1 k-Anonymity

k-Anonymity has been proposed for protecting microdata from identity disclosure [37]. It captures the well-known requirement, traditionally applied by statistical agencies, stating that any released data should be *indistinguishably* related to no less than a certain number of respondents. Basically, it characterizes the protection degree against re-identification caused by linking the released dataset with external data sources.

Due to the assumption that linking attacks exploit quasi-identifiers only, in [37] the general requirement described above has been translated into the following k-anonymity requirement: *Each release of data must be such that every combination of values of quasi-identifiers can be indistinctly matched to at least k respondents.* A microdata table then satisfies the k-anonymity requirement if each tuple in the table cannot be related to less than k respondents in the population, and each respondent in the population cannot be related to less than k tuples in the released table.

The k-anonymity requirement assumes that the data holder knows any possible external source of information that may be exploited by a malicious recipient for respondents re-identification. This assumption is however limiting and highly impractical in most scenarios. k-anonymity then takes a safe approach by requiring that each combination of values of the quasi-identifier attributes appears with at least k occurrences in the released table. The definition of k-anonymity represents a sufficient (not necessary) condition for the k-anonymity requirement. In fact, since each combination of values of the quasi-identifier attributes appears with at least k occurrences, each respondent cannot be associated with less than k tuples in the microdata table. Analogously, each tuple in the table cannot be related to less than k respondents in the population. As a consequence, the risk of re-identification is reduced by at least a factor k, independently from the external datasets available to a malicious recipient.

Traditional approaches for guaranteeing k-anonymity transform the values of the attributes composing the quasi-identifier and leave the sensitive and non-sensitive attributes unchanged. In particular, k-anonymity relies on *generalization* and *suppression*, which have the advantage of preserving the truthfulness of the released information, while producing less precise and less complete tables. Generalization consists in substituting the original values with more general values. For instance, the date of birth can be generalized by removing the day, or the day and the month, of birth. Suppression consists in removing data from

the microdata table. The intuition behind the combined use of generalization and suppression is that suppression can reduce the amount of generalization necessary to guarantee k-anonymity. In fact, if a limited number of *outliers* (i.e., quasi-identifier values with less than k occurrences in the table) would force a large amount of generalization to satisfy k-anonymity, these values can be more conveniently removed from the table. For instance, consider the microdata table in Figure 2(a) and assume that the quasi-identifier is attribute `DoB` only. Since there is only one person born in 1958, attribute `DoB` should be generalized to the decade of birth to guarantee 4-anonymity. Alternatively, removing the date of birth associated with the eleventh tuple in the table, 4-anonymity can be achieved by generalizing the date of birth to the month of birth, thus reducing the information loss. Both generalization and suppression can be applied at different granularity levels. More precisely, generalization can be applied at the cell or attribute level, while suppression can be applied at the cell, attribute, or tuple level. The combined use of generalization and suppression at different granularity levels leads to different classes of approaches for guaranteing k-anonymity [10]. However, most of the solutions proposed in the literature adopt attribute generalization and tuple suppression (e.g., Samarati's algorithm [37], Incognito [23], and k-Optimize [4]). The reason is that cell generalization produces a table where the values in the cells of the same column may be non homogeneous, since they belong to different domains (e.g., some tuples report the complete date of birth, while other tuples only report the year of birth). On the other hand, cell generalization has the advantage of causing a reduced information loss if compared to attribute generalization. Therefore, recently also solutions adopting cell generalization have been analyzed (e.g., Mondrian [24]).

Consider the microdata table in Figure 2(a) and assume that the quasi-identifier is composed of attributes `DoB`, `Sex`, and `ZIP`. Figure 4 illustrates the 4-anonymous table obtained combining tuple suppression (the eleventh tuple of the table in Figure 2(a) is suppressed) and attribute generalization (`DoB` has been generalized removing the day of birth, `Sex` has been generalized to a unique value, denoted *, and `ZIP` has been generalized removing the last two digits).

Since generalization and suppression cause information loss, it is necessary to compute an optimal k-anonymous microdata table that maximizes the utility while preserving the privacy of the respondents. The computation of an optimal k-anonymous table is an NP-hard problem, independently from the granularity level at which generalization and suppression are applied [1, 12, 31]. As a consequence, both exact and heuristic algorithms have been proposed [12]. The exact algorithms have computational time complexity exponential in the number of quasi-identifying attributes, which is however a small value compared with the number of tuples in the microdata table.

2.2 ℓ-Diversity

Although k-anonymity represents an effective solution for protecting respondents' identities, it has not been designed to protect the released microdata table against attribute disclosure. Given a k-anonymous table it may then be

SSN	Name	DoB	Sex	ZIP	Disease
		1950/06	*	941**	H1N1
		1950/06	*	941**	Gastritis
		1950/06	*	941**	Dyspepsia
		1950/06	*	941**	Pneumonia
		1940/04	*	941**	Peptic Ulcer
		1940/04	*	941**	Peptic Ulcer
		1940/04	*	941**	Peptic Ulcer
		1940/04	*	941**	Peptic Ulcer
		1940/06	*	941**	Broken Leg
		1940/06	*	941**	Short Breath
		1940/06	*	941**	Broken Leg
		1940/06	*	941**	Stomach Cancer
		1950/05	*	941**	H1N1
		1950/05	*	941**	Flu
		1950/05	*	941**	Stomach Cancer
		1950/05	*	941**	Gastritis

Fig. 4: An example of 4-anonymous table

possible to infer (or reduce the uncertainty about) the value of the sensitive attribute associated with a specific respondent.

Two well-known attacks that may lead to attribute disclosure in a k-anonymous table are the *homogeneity attack* [29, 37] and the *external knowledge attack* [29]. To illustrate the homogeneity attack, consider a k-anonymous table where all the tuples composing an *equivalence class* (i.e., all the tuples having the same value for the quasi-identifying attributes) have the same value also for the sensitive attribute. If a data recipient knows that an individual is represented in the microdata table and the value of her quasi-identifier, the recipient can easily identify the equivalence class representing the target respondent. Under the above homogeneity assumption, the recipient can also infer the value of the sensitive attribute of the target respondent. For instance, consider the 4-anonymous table in Figure 4 and suppose that *Alice* knows that her friend *Barbara* is a female living in 94139 area and born on 1940/04/10. Since all the tuples in the equivalence class with quasi-identifier value equal to ⟨1940/04,*,941**⟩ have *Peptic Ulcer* as value for attribute `Disease`, *Alice* can infer that her friend suffers from *Peptic Ulcer*. The external knowledge attack occurs when the data recipient may reduce her uncertainty about the sensitive attribute value of a target respondent, exploiting some additional (external) knowledge about the respondent. For instance, consider the 4-anonymous table in Figure 4 and suppose that *Alice* knows that her neighbor *Carol* is a female living in 94145 area and born on 1940/06/30. Observing the 4-anonymous table, *Alice* can only infer that her neighbor suffers either from *Short Breath* or *Stomach Cancer*, or has a *Broken Leg*. Suppose that *Alice* sees *Carol* running every day: as a consequence, *Alice* can infer that *Carol* can neither have broken her leg nor suffer from short breath. Hence, *Alice* discovers that *Carol* suffers from *Stomach Cancer*.

The definition of ℓ-diversity has been proposed to counteract homogeneity and external knowledge attacks [29]. ℓ-diversity extends k-anonymity requiring

SSN	Name	DoB	Sex	ZIP	Disease
		1940	M	941**	Peptic Ulcer
		1940	M	941**	Peptic Ulcer
		1940	M	941**	Broken Leg
		1940	M	941**	Short Breath
		1950	F	941**	H1N1
		1950	F	941**	Pneumonia
		1950	F	941**	H1N1
		1950	F	941**	Flu
		1940	F	941**	Peptic Ulcer
		1940	F	941**	Peptic Ulcer
		1940	F	941**	Broken Leg
		1940	F	941**	Stomach Cancer
		1950	M	941**	Gastritis
		1950	M	941**	Dyspepsia
		1950	M	941**	Stomach Cancer
		1950	M	941**	Gastritis

Fig. 5: An example of 4-anonymous and 3-diverse table

that each equivalence class must be associated with at least ℓ *well-represented* values for the sensitive attribute. In [29], the authors propose different definitions for "well-represented" values. The simplest formalization of ℓ-diversity requires that each equivalence class includes at least ℓ different values for the sensitive attribute. Even with this straightforward definition, ℓ-diversity counteracts the homogeneity attack, and reduces the effectiveness of external knowledge attacks, since the data recipient needs more external knowledge to be able to associate a single sensitive value with a target respondent. For instance, consider the microdata table in Figure 5, which is 4-anonymous and 3-diverse and suppose that *Alice* knows that her neighbor *Carol* is a female living in 94145 area and born on 1940/06/30. Observing the 3-diverse table, *Alice* can infer that her neighbor either suffers from *Peptic Ulcer*, *Stomach Cancer*, or has a *Broken Leg*. Since *Alice* only knows that *Carol* does not have a *Broken Leg*, she cannot be certain about her neighbor's disease.

Although any traditional k-anonymity algorithm can be easily adapted to guarantee ℓ-diversity [29], in [44] the authors propose a specific approximation algorithm. This algorithm is based on cell suppression, and aims at minimizing the number of suppressed cells in computing a microdata table that satisfies ℓ-diversity.

2.3 *t*-Closeness

Even if ℓ-diversity represents a first step for counteracting attribute disclosure, this solution may still produce a table that is vulnerable to privacy breaches. In fact, the definition of ℓ-diversity does not take into consideration different factors that might be exploited by a malicious data recipient for attribute disclosure, such as: *i)* the frequency distribution of the values in the sensitive attribute domain; *ii)* the possible semantic relationships among sensitive attribute values;

and *iii)* the different sensitivity degree associated with different values of the sensitive attribute domain (e.g., HIV is usually considered more sensitive than Flu).

Two attacks that may cause attribute disclosure in an ℓ-diverse table are the *skewness attack* and the *similarity attack* [26]. Skewness attack exploits the possible difference in the frequency distribution of sensitive attribute values within an equivalence class, with respect to the frequency distribution of sensitive attribute values in the whole population (or in the released microdata table). Indeed, a difference in the distribution may change the probability with which a respondent in the equivalence class is associated with a specific sensitive value. For instance, consider the 3-diverse table in Figure 5 and suppose that *Alice* knows that her friend *Diana* is a female living in 94141 area and born on 1950/06/02. Since two out of the four tuples in the equivalence class with quasi-identifier value equal to ⟨1950,F,941**⟩ have value *H1N1* for attribute `Disease`, *Alice* can infer that her friend has 50% probability of suffering from this disease, compared to the 12.5% of the whole released table. Also, although ℓ-diversity guarantees that each equivalence class has at least ℓ different values for the sensitive attribute, it does not impose constraints on their semantics. As a consequence, the sensitive attribute values within an equivalence class may be different, but semantically similar, thus causing attribute disclosure. For instance, consider the 3-diverse table in Figure 5 and suppose that *Alice* knows that her friend *Eric* is a male living in 94137 area and born on 1950/06/12. Observing the 3-diverse table, *Alice* can infer that her friend suffers from either *Gastritis*, *Dyspepsia*, or *Stomach Cancer* and, therefore, that *Eric* suffers from a stomach-related disease.

The definition of t-closeness has been proposed to counteract skewness and similarity attacks [26]. t-Closeness extends the definition of k-anonymity, requiring that the frequency distribution of the sensitive attribute values in each equivalence class has to be *close* (i.e., with distance less than a fixed threshold t) to the frequency distribution of the sensitive attribute values in the released microdata table. In this way, the frequency distribution of sensitive attribute values in each equivalence class is similar to the frequency distribution characterizing the released microdata table. As a consequence, the knowledge of the quasi-identifier value for a target respondent does not change the probability for a malicious recipient of correctly guessing the sensitive value associated with the respondent. The effect of similarity attacks is also mitigated since the presence of semantically similar values in an equivalence class can only be due to the presence, with the same relative frequencies, of these values in the microdata table.

To determine if a microdata table satisfies the t-closeness property, it is necessary to measure the distance between the frequency distributions of sensitive attribute values characterizing each equivalence class and the whole microdata table. In [26], the authors present different techniques to measure the distance between two frequency value distributions, and propose to adopt the *Earth Mover Distance* (EMD) technique.

2.4 Extensions

k-anonymity, ℓ-diversity, and t-closeness are based on some assumptions that make them not always applicable in specific scenarios. Such assumptions can be summarized as follows.

- *Each respondent is represented by only one tuple in the microdata table*. In many real-world scenarios different tuples in a microdata table may refer to the same respondent. For instance, consider the relation in Figure 2(a): each respondent may be associated with as many tuples in the table as the number of diseases she suffers from. k-anonymity is not suited for these scenarios, since an equivalence class composed of k tuples may represent less than k respondents. As a consequence, a k-anonymous table may however violate the k-anonymity requirement.
- *A unique microdata table is released*. Data may be stored in different relations, characterized by (functional) dependencies among them. Since different portions of these data may be released at different times (e.g., upon request) or to different recipients, it is necessary to guarantee that the combined view over all the released pieces of information does not cause privacy violations. Traditional solutions may not be applicable in these scenarios since they assume that the data to be released are stored in a unique microdata table.
- *The microdata table is characterized by one quasi-identifier*. Although traditional solutions assume that the released microdata table is characterized by a unique quasi-identifier, different data recipients may posses different external data collections to be linked with the released table. The release of a different microdata table to each data recipient, anonymized considering the specific knowledge of the recipient, is however not effective since collusion among recipients would violate respondents' privacy. On the other hand, a solution that considers the quasi-identifier as the union of the quasi-identifiers of any possible data recipient would cause an excessive information loss.
- *Respondents can be re-identified through a predefined quasi-identifier*. As discussed in Section 1, data respondents can be re-identified through any piece of information that uniquely (or almost uniquely) pertain to them. As an example, in transactional data (e.g., AOL and Netflix datasets) a pattern of items in a transaction (or a set thereof related to the same respondent) may re-identify a respondent. Traditional anonymization solutions are however based on a preliminary identification of a quasi-identifier, and are therefore not applicable in this scenario, where the information exploited for re-identification cannot be determined a-priori and may not be represented by a set of attributes.

In the following, we present a brief overview of some solutions that extend k-anonymity, ℓ-diversity, and t-closeness by removing (some of) the above assumptions (see Figure 6).

	(X,Y)-Privacy	MultiR k-anonymity	Butterfly	k^m-Anonymity
multiple tuples per respondent	×			×
multiple tables	×	×		
multiple quasi-identifiers			×	
non-predefined quasi-identifiers				×

Fig. 6: Extended scenarios (rows) and applicable techniques (columns)

(X,Y)-Privacy [39] addresses the problem of protecting the privacy of the respondents when multiple relations are released and each respondent is possibly associated with a set of tuples in the released tables. In particular, in [39] the authors assume that different tables, obtained as the projection of a subset of attributes of a unique microdata table, are sequentially released. To guarantee that a data release, either singularly taken or combined with previous releases, does not compromise the privacy of the respondents, the definitions of (X,Y)-*anonymity* and (X,Y)-*linkability* have been introduced, which are both referred with the term (X,Y)-*privacy*. Given a microdata table and two disjoint sets of attributes X and Y, the microdata table satisfies (X,Y)-anonymity if each value of X is linked to at least k different values of Y. Note that the definition of (X,Y)-anonymity is more general than the original k-anonymity requirement. Indeed, the k-anonymity requirement can be satisfied by a (X,Y)-anonymous table where X is the set of attributes composing the quasi-identifier and Y is the set of identifying attributes. For instance, consider the microdata table in Figure 2(a) and assume that $X = \{\texttt{DoB},\texttt{Sex},\texttt{ZIP}\}$ and $Y = \{\texttt{SSN}\}$. A (X,Y)-anonymous table guarantees that each combination of values for attributes `DoB`, `Sex`, `ZIP` is associated with at least k different values for attribute `SSN`. (X,Y)-linkability states that no value of Y can be inferred from a value of X with confidence higher than a fixed threshold. To guarantee privacy protection in sequential releases, the join among all the released tables has to satisfy (X,Y)-privacy (i.e., it must satisfy both (X,Y)-anonymity and (X,Y)-linkability).

MultiR *k*-anonymity [35] extends the definition of k-anonymity to multiple relations, while guaranteeing a limited information loss. The considered scenario is characterized by a set $S = \{T_1, \ldots, T_n\}$ of relations that preserves the lossless join property and that includes a *person specific table* PT, where each row corresponds to a different respondent. The release of a set S of relations satisfies *multiR k-anonymity* if each view over the join $JT = PT \bowtie T_1 \bowtie \ldots \bowtie T_n$ is k-anonymous, meaning that it includes either 0 or at least k occurrences for each combination of values for the attributes in the quasi-identifier. In this scenario, the quasi-identifier is defined as a subset of the attributes of the join relation JT, which can be exploited to re-identify the respondents in the person specific table

PT. To protect the privacy of the respondents against attribute (besides identity) disclosure, in [35] the authors propose to extend the definition of ℓ-diversity to the multirelational scenario. A release of a set S of relations satisfies multiR ℓ-diversity if each view over the join $JT = PT \bowtie T_1 \bowtie \ldots \bowtie T_n$ among all the released relations satisfies ℓ-diversity with respect to the considered sensitive attribute.

Butterfly [36] aims at protecting the privacy of the respondents, while minimizing information loss, when the microdata table to be released is characterized by different quasi-identifiers. Traditional approaches consider, as a quasi-identifier, the set of all the attributes that may be externally available (although not in combination) to some data recipient. Even if effective, this assumption causes an excessive information loss that can be reduced if the set of attributes available to each recipient is considered separately. Indeed, as shown in [36], a table that is k-anonymous with respect to quasi-identifier QI_1 and with respect to QI_2 may not be k-anonymous with respect to $QI_1 \cup QI_2$. In [36] the authors then introduce the *k-butterfly* principle that allows the anonymization of a microdata table so that the k-anonymity requirement is satisfied with respect to multiple quasi-identifiers while reducing the information loss. More precisely, given a microdata table and two quasi-identifiers QI_1 and QI_2, a subset of tuples with the same value for the attributes in $QI_1 \cap QI_2$ satisfies k-butterfly with respect to QI_1 and QI_2 if the equivalence classes induced by $QI_1 \setminus QI_2$ ($QI_2 \setminus QI_1$, resp.) include at least k tuples each. A microdata relation satisfies k-butterfly if all the equivalence classes induced by $QI_1 \cap QI_2$ satisfy k-butterfly. It is important to note that a set of tuples that satisfies k-butterfly with respect to QI_1 and QI_2 is k-anonymous with respect to both QI_1 and QI_2.

k^m-Anonymity [38] aims at protecting the privacy of respondents of transactional data, where a subset of the items composing a transaction may represent a quasi-identifier. For instance, consider a transactional dataset including all the items purchased in a supermarket and assume that *Alice* has seen her neighbor *Bob* buying parmesan cheese, strawberries, and soy milk. If the supermarket dataset includes only one transaction with this combination of items, *Alice* can find out the complete list of items purchased by *Bob*, which may be sensitive (e.g., if it contains a medicine). The *k^m-anonymity* concept solves this problem by requiring that each combination of at most m items must appear in at least k transactions, where m is the maximum number of items per transaction that may be known to a malicious data recipient.

3 Multiple releases and data streams

The proposals described in Section 2 assume that only one instance of the microdata is published. There are however many scenarios where data are subject to frequent changes, due to the insertion, deletion, and update of tuples in the microdata, and need to be published at a regular basis. For instance, a hospital may be forced by law to publish an anonymized version of its data every six months.

Data may also be continuously generated and may need to be immediately released. This happens, for example, when a credit card company outsources to a third party the transactions generated by its customers whenever they use their credit cards. Such transactions form a *microdata stream* that has to be continuously monitored by the third party to detect possible misbehaviors. In this section, we first describe the privacy issues that arise in the case of multiple data releases and data streams, highlighting why the traditional approaches for protecting microdata are not suitable. We then illustrate some recent approaches that counteract such privacy issues.

3.1 Problem

The multiple releases of a microdata table whose content is updated over time may cause information leakage since a malicious data recipient can correlate the released datasets to gain information about respondents. Also, if a combination of quasi-identifier values appears in one of the released tables only, it may be easier for a malicious recipient to correctly associate it with one of those respondents who have been removed from the subsequent releases (possibly exploiting the information available through external sources). To illustrate, consider the two subsequent releases of the medical data of the patients of a hospital illustrated in Figure 7. Some patients are included in both tables, while others are represented in one of the two tables only (tuples marked with a bullet in Figures 7(a) and 7(b)). Both the first and the second release satisfy 4-anonymity and 3-diversity. However, suppose that *Alice* knows that her friend *Fay*, who is a female living in the 94130 area, born on 1940/04/20, was in the hospital when both tables have been released. *Alice* then knows that *Fay* is represented by a tuple in the tables of Figure 7, and she can identify the two equivalence classes to which *Fay* belongs in the two tables. In particular, in the first release *Fay* is included in the third class of the table in Figure 7(a), and hence her disease could be either *Peptic Ulcer*, *Broken Leg*, or *Stomach Cancer*. Instead, in the second release *Fay* is included in the second class of the table in Figure 7(b), and hence her disease could be either *Gastritis*, *Short Breath*, *Pneumonia*, or *Peptic Ulcer*. By comparing the two sets of possible illnesses, *Alice* can easily infer that *Fay* suffers from *Peptic Ulcer*, since this is the unique disease that appears in both sets.

When we consider a data stream scenario where only new tuples can be inserted into the released microdata table, it is not necessary to compute different releases of the dataset. In fact, it may be more convenient to release only the new tuples. In this case, traditional approaches cannot be adopted, since they need the whole dataset that is instead not available in advance. Furthermore, a data stream is possibly an infinite sequence of tuples, which must be released as soon as they are available since data utility decreases as time passes. As a consequence, it is not possible to assume that the whole dataset can be collected before its publication. For instance, consider a credit card company that releases the stream of data generated by purchases represented as tuples in a microdata table. Such data should be immediately published to check, for example, possible

	DoB	Sex	ZIP	Disease
	1940	M	941**	Peptic Ulcer
	1940	M	941**	Peptic Ulcer
•	1940	M	941**	Broken Leg
	1940	M	941**	Short Breath
	1950	F	941**	H1N1
	1950	F	941**	Pneumonia
	1950	F	941**	H1N1
•	1950	F	941**	Flu
•	1940	F	941**	Peptic Ulcer
	1940	F	941**	Peptic Ulcer
•	1940	F	941**	Broken Leg
	1940	F	941**	Stomach Cancer
•	1950	M	941**	Gastritis
	1950	M	941**	Dyspepsia
•	1950	M	941**	Stomach Cancer
	1950	M	941**	Gastritis

(a)

	DoB	Sex	ZIP	Disease
	1940	*	9414*	Peptic Ulcer
	1940	*	9414*	Peptic Ulcer
•	1940	*	9414*	Measles
	1940	*	9414*	Stomach Cancer
•	1940	*	9413*	Gastritis
	1940	*	9413*	Short Breath
•	1940	*	9413*	Pneumonia
	1940	*	9413*	Peptic Ulcer
	1950	*	9414*	H1N1
	1950	*	9414*	Pneumonia
•	1950	*	9414*	Measles
	1950	*	9414*	Gastritis
•	1950	*	9413*	Infract
	1950	*	9413*	Dyspepsia
	1950	*	9413*	H1N1
•	1950	*	9413*	Thrombosis

(b)

Fig. 7: An example of two subsequent releases of a table including the medical data of a hospital

frauds, while preserving the privacy of the card holders. We note that data streams may include more than one tuple for each respondent (e.g., a card holder may perform multiple purchases).

3.2 Solutions

Recently, the scientific community has proposed different approaches addressing the privacy issues previously discussed. These solutions can be classified in two categories, depending on whether they consider the *re-publication* of data or the publication of *data streams*. In the following, we summarize some of the approaches proposed in both scenarios.

Data re-publication (e.g., [39, 43]). Most of the solutions proposed to protect the privacy of data respondents in data re-publication scenarios only focus on supporting the insertion of new tuples in the dataset and implicity assume that no tuple is removed from the microdata table. *m-Invariance* [43] is the first technique addressing the problem of data re-publication that takes both insertion and deletion of tuples into account. The possible removal of tuples from the dataset can cause information leakage, due to the *critical absence* phenomenon. To illustrate, consider the two subsequent releases in Figure 7 and suppose that *Alice* knows that *Gabrielle* is represented in both microdata tables and that she suffers from either *Flu* or *H1N1*. The value *Flu*, however, does not appear in the second release. As a consequence, *Alice* can conclude with certainty that *Gabrielle* contracted *H1N1*. It is important to note that this inference can be drawn independently from how the two released datasets have been generalized to prevent disclosure. To counteract this privacy breach, in [43] the authors introduce the m-invariance property. A sequence $T_1, \ldots, T_n$ of released microdata

tables satisfies m-invariance if the following properties hold: *i)* each equivalence class in T_i, $i = 1, \ldots, n$, includes at least m tuples; *ii)* no sensitive value appears more than once in each equivalence class in T_i, $i = 1, \ldots, n$; and *iii)* for each tuple t, the equivalence classes to which t belongs in the sequence $[T_i, T_j]$, $1 \leq i \leq j \leq n$, are characterized by exactly the same set of sensitive values. The rationale of m-invariance is that all the equivalence classes to which a published tuple t belongs must be associated with exactly the same set of (at least) m different sensitive values. In this way, the correlation of the tuples in $T_1, \ldots, T_n$ does not permit a malicious recipient to associate less than m different sensitive values with each respondent in the released datasets. The technique proposed in [43] to achieve m-invariance is incremental, meaning that the n-th release T_n can be determined taking into account only the previous release T_{n-1}, and is based on the possible insertion of counterfeits, when it is needed to prevent critical absences.

Data streams (e.g., [25, 40, 45]). One of the most important requirements when releasing data as a stream is timeliness, since these data are usually time critical and need to be published in a timely fashion to be useful for the recipients. The proposed solutions mainly aim at satisfying the k-anonymity requirement. To this purpose, they rely on generalization and on the introduction of a limited delay in data publication. The first solution in this direction has been proposed in [45] and is based on the principle that all the tuples in an equivalence class must be published at the same time. Therefore, the data holder locally maintains a set of equivalence classes, which are all initially empty. As a new tuple is generated by the stream, it is inserted into a suitable equivalence class, if such class exits; a new equivalence class suitable for the tuple is generated, otherwise. We note that each equivalence class can include at most one tuple for each respondent. As soon as one of the equivalence classes includes k tuples that, by construction, are related to k different respondents, these tuples are generalized to the same quasi-identifier value and published. This technique, although simple, guarantees privacy protection and a timely data publication, at the price of a possibly high information loss. To limit this information loss, in [45] the authors introduce an improvement of their technique that makes a probabilistic estimation of the tuples that will be generated by the data stream, to choose the most convenient generalization strategy. An alternative approach for protecting the privacy of respondents in data streams has recently been proposed in [40], where the authors assume that data are generated and published as "snapshots" (i.e., sets of records available at a given moment of time) of d tuples each. This technique combines generalization and tuple suppression with *tuple relocation* to guarantee ℓ-diversity. Relocation consists in moving a tuple from one snapshot to a more recent one, if this delay in data publishing could be useful to satisfy the ℓ-diversity principle. The approach illustrated in [40] guarantees that window queries (i.e., queries that need to be evaluated only on the data released in a specific time window) evaluated on the released dataset produce the same result as if they were evaluated on the original data stream. Another approach for protecting data streams has been illustrated in [25]. This approach is based

on noise addition and therefore is not suitable for all those scenarios that require truthfulness of released data.

4 Fine-grained privacy preferences

The privacy-aware publishing techniques illustrated in previous sections guarantee the same amount of privacy to all the respondents represented in the released microdata table. Privacy requirements may however depend on respondents' preferences, or on the sensitivity of the released values. In the following of this section, we illustrate both the issues that may arise when enforcing the same protection degree to all the respondents, and the advantages of solutions that permit a fine-grained specification of privacy preferences. We also describe some recent approaches supporting fine-grained privacy preference specifications.

4.1 Problem

Privacy is an *individual concept* [27], since each individual may have her own privacy needs that may be different from the requirements of another individual. However, traditional privacy protection techniques provide all the data respondents with the same amount of privacy, without considering their preferences. For instance, the k-anonymity requirement demands that each tuple in a released table cannot be associated with less than k respondents in the population, and viceversa. The anonymity threshold k is fixed by the data holder, without considering the specific privacy requirements of the respondents. As a consequence, this value may be adequate for some respondents and inadequate for others. For instance, consider the 4-anonymous microdata table in Figure 4 and assume that *Gabrielle* is a female born on 1950/05/02 and living in the 94136 area, who suffers from *H1N1*. A data recipient knowing her quasi-identifier can infer that *Gabrielle* suffers from either *H1N1*, *Flu*, *Stomach Cancer*, or *Gastritis*, being these values all equally likely (25% of probability each). Although the microdata table satisfies 4-anonymity, *Gabrielle* may not want people to know, with a probability of 25%, that she suffers from *H1N1*. On the other hand, *Lorna*, born on 1950/05/05 and living in the 94134 area and suffering from *Flu*, may agree to release her disease without the need to protect the corresponding tuple in the table. Therefore, the 4-anonymous table in Figure 4 does not protect *Gabrielle*'s privacy, while it over-protects *Lorna*'s sensitive value. ℓ-diversity, t-closeness, m-invariance, and all the other approaches illustrated in the previous sections suffer from this problem, since they are based on a unique protection threshold that is adopted to guarantee the privacy of all the respondents. As a consequence, these techniques cannot be adopted for supporting the following privacy requirements.

- *Value-specific privacy*: different values of the sensitive attribute should enjoy a different protection degree. For instance, consider the medical microdata table in Figure 2(a). *Stomach Cancer* is usually considered more sensitive than *Flu*. As a consequence, tuples representing patients suffering from

Stomach Cancer should be more carefully protected than tuples of patients suffering from *Flu*.

- *Respondent-specific privacy*: different respondents may have a different perception of their privacy, independently from the values of the sensitive attribute. As a consequence, respondents may be willing to specify a different threshold for the protection of their data. As an example, consider two patients suffering from the same disease. A patient may consider the protection offered by a 3-diverse table adequate, while the other patient may require at least a 5-diverse table.
- *Sensitive presence*: the presence of a respondent in the microdata table may violate her privacy. For instance, consider a microdata table representing patients suffering from rare diseases. Even the fact that an individual is represented in the table may violate her privacy.

To satisfy the privacy requirements of all the respondents, the privacy parameter used by the techniques previously discussed should be fixed to the most restrictive threshold (e.g., to the highest value of k among the preferences of the respondents), thus causing an excessive information loss.

4.2 Solutions

Recently, different solutions supporting fine-grained privacy specifications have been proposed. In the following, for each of the privacy requirements described above, we illustrate a specific protection technique addressing it.

Value-specific privacy (e.g., [18]). To permit the data holder to specify different privacy thresholds for different values of the domain of the sensitive attribute, in [18] the authors propose (α_i, β_i)-*closeness*. (α_i, β_i)-closeness is an extension of the t-closeness principle that, instead of adopting a unique threshold value t, associates a range $[\alpha_i, \beta_i]$ with each value s_i in the domain of the sensitive attribute. An equivalence class satisfies (α_i, β_i)-closeness if, for each sensitive value s_i, the percentage of tuples in the class with value s_i for the sensitive attribute is in the range $[\alpha_i, \beta_i]$. A microdata table satisfies (α_i, β_i)-closeness if each equivalence class in the table satisfies (α_i, β_i)-closeness. (α_i, β_i)-closeness presents different advantages over t-closeness. First, (α_i, β_i)-closeness is flexible, since it defines a different threshold for each sensitive value. As a consequence, only the values that are considered highly sensitive are associated with a small range, while the frequency of non-sensitive values is possibly not bounded, thus better preserving data utility. Second, (α_i, β_i)-closeness is easy to check, since it does not require to compute the distance between two frequency distributions, but only to evaluate the number of tuples in the equivalence class associated with each sensitive value. Finally, (α_i, β_i)-closeness is easy to use, since the definition of a range $[\alpha_i, \beta_i]$ for each sensitive value is more intuitive than the choice of the maximum distance t from a known frequency distribution. (α_i, β_i)-closeness is enforced by applying generalization at the level of attributes [18]. In particular, the authors propose to enforce the (α_i, β_i)-closeness requirement by extending

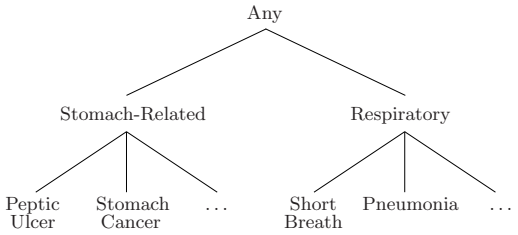

Fig. 8: An example of partial taxonomy tree for attribute `Disease`

traditional algorithms used for solving the k-anonymity problem (see Section 2) and, in particular, the Incognito algorithm [23].

Respondent-specific privacy (e.g., [42]). In [42] the authors observed that different respondents may have different privacy requirements, independently from the value of the sensitive attribute associated with them. As a consequence, the adoption of a unique privacy protection threshold to the whole dataset may result in over-protecting or under-protecting respondents. To overcome this issue, in [42] the authors introduce the concept of *personalized anonymity* that aims at releasing as much information as possible, while satisfying the specific privacy requirement defined by each respondent. Personalized anonymity is based on the definition of a taxonomy tree over the domain of each sensitive attribute in the microdata table. The tree has a leaf for each value in the sensitive attribute domain, and each internal node summarizes the specific values in its subtree. For instance, Figure 8 illustrates an example of a partial taxonomy tree for attribute `Disease` in the microdata table in Figure 2(a). Each respondent specifies her privacy preference by choosing a *guarding node* along the path from the root of the taxonomy tree to the leaf node representing the respondent's sensitive attribute value.

The technique proposed in [42] for enforcing the privacy preferences of the respondents relies on the application of generalization, performed in two steps. In the first step, the attributes composing the quasi-identifier are generalized following traditional generalization schemes [37]. In the second step, each equivalence class generated in the first step is further modified to respect the privacy level required by each respondent. In particular, for each tuple in an equivalence class, the value of the sensitive attribute is generalized (based on the taxonomy tree defined for the attribute) if needed to respect the guarding node chosen by the tuple respondent. A table to be released satisfies personalized anonymity if, for each respondent, a malicious recipient cannot infer with probability higher than p that the respondent is associated with a specific value in the subtree rooted at the guarding node chosen by the respondent herself. The value of p represents the maximum confidence with which a recipient is allowed to infer sensitive attribute values and is set by the data holder. For instance, consider the taxonomy tree in Figure 8 and suppose that *Carol*, who suffers from *Stomach*

Cancer, chooses *Stomach-Related* disease as her guarding node. *Carol*'s privacy is violated if a data recipient can infer with probability higher than p that *Carol* suffers from a disease among *Peptic Ulcer*, *Stomach Cancer*, and the other values children of the *Stomach-Related* node in the taxonomy tree. Another respondent, *Matt*, who suffers from *Short Breath*, does not consider a privacy violation the release of his disease. As a consequence, he sets his guarding node as $\emptyset$ (i.e., a special guarding node denoting that a respondent believes that the release of her sensitive attribute value does not violate her privacy).

Sensitive presence (e.g., [34]). In [34], the authors propose δ-*presence* as a metric to evaluate the risk that a data recipient can identify the presence of an individual in the released table. The released dataset includes a subset of the tuples of a larger data collection D. The released microdata table T^*, obtained by generalizing (at the level of attributes) $T \subseteq D$, satisfies δ-presence if $\delta_{min} \leq P(t \in T|T^*) \leq \delta_{max}$, for all $t \in D$, where $P(t \in T|T^*)$ is the probability that a data recipient correctly guesses that tuple t belongs to the released dataset, observing the released microdata table T^*. If the released dataset satisfies δ-presence, a malicious recipient cannot determine the inclusion of a respondent in T^* with probability lower than δ_{min} or higher than δ_{max}. We note that, by tuning the values of δ_{min} and δ_{max}, it is possible to find a good trade-off between data utility and privacy of the released generalized microdata table T^*. In fact, a small $[\delta_{min},\delta_{max}]$ range favors privacy, while a large $[\delta_{min},\delta_{max}]$ range favors data utility.

5 Group-based approaches for protecting sensitive associations

The approaches described in the previous sections typically apply generalization and suppression for guaranteeing k-anonymity, ℓ-diversity, t-closeness, or other (more enhanced) privacy requirements. Generalization and suppression, however, cause information loss that could be reduced by adopting different protection techniques. In the following of this section, we highlight the disadvantages caused by generalization and illustrate an alternative approach that overcomes these shortcomings, while protecting sensitive data against disclosure.

5.1 Problem

The adoption of well-known generalization and suppression techniques results in tables that are less complete and less detailed than the original microdata tables. In fact, the released table is composed of a set of equivalence classes, including all the tuples that have been generalized to the same value for the quasi-identifier. The values of the quasi-identifier in the released table are then less precise than the values in the original data collection, thus destroying the correlation among the values of the quasi-identifying attributes and the sensitive attribute. The generalization-based approaches are therefore not suitable for those publishing

scenarios where the precision of aggregate queries is of paramount importance and the exact distribution of the values of the quasi-identifier must be released. For instance, if the correlation existing among `Sex`, `ZIP`, and `Disease` is important for the analysis of the impact of an infectious disease, the release of the 4-anonymous table in Figure 4 is of limited utility for the final recipient, since both `Sex` and `ZIP` have been generalized to a unique value.

5.2 Solutions

An alternative technique to generalization that permits the release of the exact distribution of the quasi-identifier values, while guaranteeing to preserve the privacy of the respondents, is *fragmentation*. Basically, fragmentation consists in splitting the original microdata table in vertical fragments, such that the attributes composing the quasi-identifier and the sensitive attribute are not represented in the same fragment. In the following, we illustrate two different solutions adopting a group-based approach to protect the privacy of the respondents.

Anatomy [41] is a group-based proposal addressing the issue of guaranteing ℓ-diversity in microdata release without resorting to generalization. Anatomy first partitions the tuples in the microdata table in groups that satisfy the ℓ-diversity principle (i.e., each group includes at least ℓ well-represented values for the sensitive attribute). Each group is then associated with a unique group identifier and the microdata table is split into two fragments, F_1 and F_2, including the attributes composing the quasi-identifier and the sensitive attribute, respectively. For each tuple, both F_1 and F_2 report the identifier of the group to which it belongs. For simplicity, each group in the fragment storing the sensitive attribute has a tuple for each sensitive value appearing in the group, and reports the frequency with which the value is represented in the group. For instance, consider the microdata table in Figure 2(a) and assume that the data holder is interested in releasing a 3-diverse table. Figure 9 illustrates the two fragments F_1 and F_2 obtained by partitioning the tuples in the table in Figure 2(a) in groups that satisfy 3-diversity. Although a malicious recipient may know the quasi-identifier value of a target respondent, she can only infer that the respondent belongs to one group (say, g_1) in F_1, and that the sensitive value of the target respondent is one of the values in the group in F_2 that is in relation with g_1. To illustrate, assume that *Alice* knows that her friend *Barbara* is a female living in 94139 area and born on 1940/04/10. *Alice* can easily infer that her friend is represented by the ninth tuple of the table in Figure 9(a). However, since the tuples in the third group in Figure 9(a) are in relation with the tuples in the third group in Figure 9(b), *Alice* can only infer that *Barbara* suffers from either *Peptic Ulcer*, *Broken Leg*, or *Stomach Cancer*. Note that the privacy guarantee offered by Anatomy is exactly the same offered by traditional generalization-based approaches. In fact, a malicious data recipient cannot associate less than ℓ different sensitive values with each respondent in the released table. On the other hand, by releasing the exact distribution of the values of the attributes composing the quasi-identifier, the evaluation of aggregate queries can be more precise [41].

DoB	Sex	ZIP	GroupID
1940/04/01	M	94143	1
1940/04/02	M	94142	1
1940/06/07	M	94130	1
1940/06/05	M	94131	1
1950/06/02	F	94141	2
1950/06/05	F	94144	2
1950/05/02	F	94136	2
1950/05/05	F	94134	2
1940/04/10	F	94139	3
1940/04/20	F	94130	3
1940/06/25	F	94142	3
1940/06/30	F	94145	3
1950/06/20	M	94132	4
1950/06/12	M	94137	4
1950/05/10	M	94147	4
1950/05/30	M	94148	4

(a) F_1

GroupID	Disease	Count
1	Peptic Ulcer	2
1	Broken Leg	1
1	Short Breath	1
2	H1N1	2
2	Pneumonia	1
2	Flu	1
3	Peptic Ulcer	2
3	Broken Leg	1
3	Stomach Cancer	1
4	Gastritis	2
4	Dyspepsia	1
4	Stomach Cancer	1

(b) F_2

Fig. 9: An example of two fragments satisfying 3-diversity obtained adopting the Anatomy approach

Loose associations [15] represent a more flexible solution to guarantee privacy in data publication without adopting generalization. Loose associations have been proposed to protect generic sensitive associations among the attributes in a data collection. For instance, consider the microdata table in Figure 1 and suppose that attribute `Treatment` is also represented in the table. A possible set of sensitive associations defined among attributes {`SSN`,`Name`,`DoB`, `Sex`,`ZIP`,`Disease`,`Treatment`} could include: *i)* both the association between the values of attributes `SSN` and `Disease`, and the association between the values of attributes `Name` and `Disease`; *ii)* the association between the values of quasi-identifying attributes `DoB`, `Sex`, `ZIP` and the values of sensitive attribute `Disease`; *iii)* the association between the values of attributes `Disease` and `Treatment`. Given a set of sensitive associations defined among the attributes included in a microdata table, they are broken by publishing a set of different fragments. It is easy to see that the problem of protecting the association of a sensitive attribute with the respondents' quasi-identifier can be modeled through the definition of a sensitive association among the sensitive attribute and quasi-identifying attributes. Like Anatomy, the original microdata table can then be split in different fragments in such a way that the sensitive attribute is not stored together with all the attributes composing the quasi-identifier. It is in fact sufficient to store a subset of the quasi-identifying attributes in a fragment F_1, and all the other quasi-identifying attributes in another fragment F_2, together with the sensitive attribute. For instance, consider the microdata table in Figure 2(a). A fragmentation that would protect against identity and attribute disclosures could be composed of the following two fragments: F_1(`DoB`,`Sex`,`ZIP`) and F_2(`Disease`). Note that a fragmentation is not unique: F_1(`DoB`,`Sex`) and F_2(`ZIP`,`Disease`) is

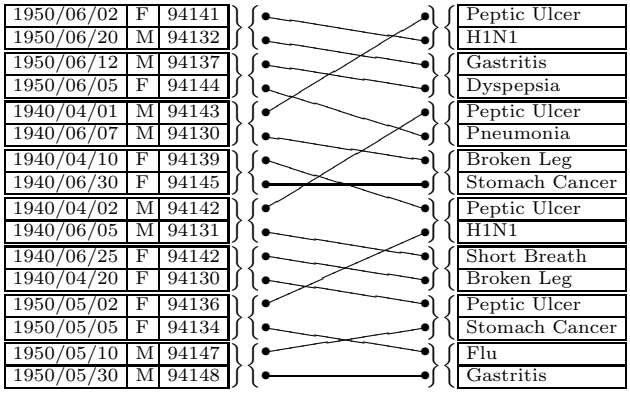

Fig. 10: An example of loose association defined on the table in Figure 2(a)

another solution that still protects the association between the sensitive attribute and the quasi-identifier.

To provide the data recipient with some information on the associations between the quasi-identifier and the sensitive attribute values existing in the original relation, provided a given privacy degree of the association is respected, in [15] the authors propose to publish a loose association between the tuples composing F_1 and F_2. The tuples in F_1 and in F_2 are independently partitioned in groups of size at least k_1 and k_2, respectively. Each group in F_1 and in F_2 is then associated with a different group identifier. For each tuple, both F_1 and F_2 report the identifier of the group to which the tuple belongs. The group-level relationships between the tuples in F_1 and in F_2 are represented by an additional table A that includes, for each tuple t in the original microdata table, a tuple modeling the relationship between the group where t appears in F_1 and the group where t appears in F_2. For instance, Figure 10 represents two fragments F_1 and F_2 for the microdata table in Figure 2(a). Both the fragments have been partitioned into groups of 2 tuples each and the lines between the tuples in F_1 and F_2 represent their relationships in the original microdata table. Figure 11 illustrates the three relations, F_1, A, and F_2 that are released instead of the original microdata. It is easy to see that, even if a malicious recipient knows the quasi-identifier of a respondent, she can only identify the tuple related to the target respondent in F_1, but not the corresponding `Disease` in F_2. For instance, assume that *Alice* knows that her friend *Barbara* is a female living in 94139 area and born on 1940/04/10. By looking at the released tables, *Alice* discovers that her friend is represented by the seventh tuple in F_1, which belongs to group *dsz4*. However, since group *dsz4* is associated in A with two different groups in F_2 (i.e., *d4* and *d5*) *Alice* cannot identify the illness *Barbara* suffers from, since it could be either *Peptic Ulcer*, *Broken Leg*, *Stomach Cancer*, or *H1N1*.

The partitioning of the tuples in the two fragments should be carefully designed to guarantee an adequate protection degree. In fact, a loose association enjoys a degree k of protection if every tuple in A indistinguishably corresponds

F_1

DoB	Sex	ZIP	G
1950/06/02	F	94141	dsz1
1950/06/20	M	94132	dsz1
1950/06/12	M	94137	dsz2
1950/06/05	F	94144	dsz2
1940/04/01	M	94143	dsz3
1940/04/02	M	94142	dsz5
1940/04/10	F	94139	dsz4
1940/04/20	F	94130	dsz6
1940/06/07	M	94130	dsz3
1940/06/05	M	94131	dsz5
1940/06/25	F	94142	dsz6
1940/06/30	F	94145	dsz4
1950/05/02	F	94136	dsz7
1950/05/05	F	94134	dsz7
1950/05/10	M	94147	dsz8
1950/05/30	M	94148	dsz8

(a)

A

G_1	G_2
dsz1	d1
dsz1	d2
dsz2	d2
dsz2	d3
dsz3	d1
dsz3	d4
dsz4	d5
dsz4	d4
dsz5	d3
dsz5	d6
dsz6	d7
dsz6	d6
dsz7	d5
dsz7	d8
dsz8	d7
dsz8	d8

(b)

F_2

Disase	G
H1N1	d1
Gastritis	d2
Dyspepsia	d2
Pneumonia	d3
Peptic Ulcer	d1
Peptic Ulcer	d3
Peptic Ulcer	d5
Peptic Ulcer	d7
Broken Leg	d4
Short Breath	d6
Broken Leg	d6
Stomach Cancer	d4
H1N1	d5
Flu	d8
Stomach Cancer	d7
Gastritis	d8

(c)

Fig. 11: An example of 4-loose association

to at least k distinct associations among tuples in the two fragments (i.e., it could have been generated starting from k different tuples in the microdata table). The release of F_1, F_2, and A satisfies *k-looseness*, with $k \leq k_1 \cdot k_2$, if for each group g_1 in F_1 (group g_2 in F_2, respectively), the union of the tuples in all the groups with which g_1 (g_2, respectively) is associated in A is a set of at least k different tuples.

Figure 11 represents an example of a 4-loose association. This implies that it is not possible, for a malicious data recipient, to associate with each quasi-identifier value in F_1 less than 4 different diseases in F_2.

6 Microdata publishing with adversarial external knowledge

Another source of complication for the protection of microdata is the external (or background) knowledge that an adversary may exploit for inferring information about the respondents represented in the microdata. The research community has recently dedicated many efforts for counteracting this problem whose difficulty lies in the fact that the data holder is unaware of the type of knowledge an adversary may have. In this section, we illustrate the privacy problems that may arise due to the adversarial external knowledge, and we present recent protection techniques specifically designed to consider such knowledge.

6.1 Problem

When publishing a microdata table it is necessary to take into consideration the fact that a malicious recipient may exploit, besides the released table, also additional information to infer the sensitive attribute value associated with a

target respondent. This knowledge can be obtained by similar data collections released by other organizations or competitors, by social networking sites, or by personal knowledge. For instance, consider the microdata table in Figure 5 and suppose that *Alice* knows that her neighbor *Gabrielle* is a female born on 1950/05/02 and living in the 94136 area. If *Alice* does not have any additional knowledge, by looking at the released microdata table she can only infer that her friend suffers from either *H1N1*, *Pneumonia*, or *Flu*. However, *Alice* can improve this inference by exploiting external sources of information. For instance, *Alice* is a close friend of *Liz* (who appears in the same equivalence class as *Gabrielle* in the table in Figure 5) and therefore knows that she suffers from *Pneumonia*. As a consequence, *Alice* can infer that *Gabrielle* either suffers from *Flu* or *H1N1*. However, *Gabrielle*'s sister *Hellen* has been recently hospitalized for having contracted *H1N1*, and she posted it on her Facebook profile. Since *H1N1* is a contagious disease, *Alice* can infer with high probability that also *Gabrielle* contracted *H1N1*.

The problem of protecting released data against malicious recipients exploiting external knowledge has been acknowledged since the first introduction of proposals addressing privacy issues in microdata release (see Section 2). However, traditional approaches only consider a few specific types of external knowledge. For instance, k-anonymity assumes that data recipients only know publicly available datasets associating the identity of respondents with their quasi-identifier. ℓ-diversity considers also the fact that a recipient may have additional (personal) knowledge about a subset of the respondents, which permits her to discard a subset of the sensitive values in the equivalence class of a target respondent.

Taking external knowledge into consideration when releasing a microdata collection requires the definition of an adequate modeling of the knowledge that a malicious recipient may possess. This task is complicated by the fact that it is not realistic to assume that data holders have complete knowledge of all the data available to recipients. Furthermore, information is collected and publicly released every day and, consequently, the external information that could be exploited for re-identifying respondents is continuously changing. The external knowledge modeling should therefore be flexible enough to possibly capture any kind of information that might be available to the data recipient.

6.2 Solutions

We summarize some of the most important results modeling the external knowledge of an adversary.

(c,k)-Safety [30] introduces a formal modeling of external knowledge assuming a worst case scenario, where the malicious recipient knows the set of respondents represented by a tuple in the published microdata table, and the values of both quasi-identifying and non-sensitive attributes associated with each respondent. Besides this *identification information*, a malicious recipient may also possess additional information modeled with the concept of *basic unit* of knowledge. A basic unit of knowledge is defined as an implication formula $(\wedge_i A_i) \rightarrow (\vee_j B_j)$,

where A_i and B_j are atoms of the form $t_p[S] = s$ that represent the fact that respondent p is associated with value s for sensitive attribute S. For instance, suppose that *Alice* knows that *Bob* and *Carol*, who are married, are both sick. Since *Flu* is highly contagious, *Alice* knows that if *Bob* has *Flu* also *Carol* suffers from the same disease. This knowledge can be modeled through the following basic unit: $t_{\text{Bob}}[\texttt{Disease}]=Flu \rightarrow t_{\text{Carol}}[\texttt{Disease}]=Flu$. To avoid unrealistic scenarios where the malicious recipient has unbounded external knowledge, in [30] the authors assume that the knowledge available to a data recipient is composed of at most k basic units of knowledge. Therefore, the overall external knowledge of a recipient is represented as the conjunction φ of all her basic units of knowledge. Given the released microdata table T (either obtained using a generalization-based or a group-based approach), the *maximum disclosure risk* to which the released table is exposed can be computed as the maximum probability that a data recipient with knowledge φ can infer that respondent p is associated with value s for sensitive attribute S. More formally, the maximum disclosure of table T is defined as $max\ P(t_p[S] = s|T \wedge \varphi)$, computed on the tuples t_p of table T that are consistent with external knowledge φ, and that take value s in the domain of S. A released table T satisfies (c, k)-safety if the maximum disclosure risk of T is lower than a threshold $0 \leq c \leq 1$ fixed by the data holder, assuming that malicious recipients have at most k basic units of external knowledge. Given a microdata table, the goal is therefore to determine a corresponding (c, k)-safety microdata table that minimizes the loss of information due to the adoption of generalization-based or group-based approaches. In [30] the authors note that traditional algorithms proposed to guarantee k-anonymity can be easily adapted to guarantee (c, k)-safety, since the maximum disclosure is monotonic with respect to generalization. The authors also note that Anatomy [41] can be easily adapted to guarantee (c, k)-safety, since the disclosure decreases merging the groups of tuples in the fragments.

Privacy skyline [6] introduces a more intuitive and usable approach than (c, k)-safety to measure the external knowledge available to a data recipient. In fact, the definition of an adequate value of k modeling the adversarial knowledge is not intuitive. Also, (c, k)-safety is based on the assumption that each data respondent is associated with a unique sensitive value and that all the values in the sensitive attribute domain are equally sensitive. Privacy skyline tries to overcome these limitations. The basic idea is that external knowledge can be seen as composed of several categories, which need to be quantified adopting different measures. As a consequence, the disclosure risk cannot be a numeric value, but it is decomposed according to the external knowledge components that are considered relevant for the specific scenario. In [6], the authors classify the external knowledge in the following three categories:

- the knowledge about a *target individual* (e.g., *Alice* knows that *Bob* does not suffer from *Cancer*);
- the knowledge about individuals *other* than the target individual (e.g., *Alice* knows that *Carol*, who is also a respondent of the considered microdata table, suffers from *H1N1*);

– the knowledge about *same-value families* that model the fact that a group (or family) of respondents have the same value for the sensitive attribute (e.g., *Alice* knows that *David* and *Hellen* are married and, therefore, if any of them suffers from *Flu*, it is highly probable that also the other person suffers from *Flu*).

It is important to note that the composition of a same-value family is value specific, meaning that the relationship among users may be different depending on the specific sensitive value s considered. For instance, the same-value family for an infectious disease may be composed of colleagues and people leaving in the same house, while the same-value family for HIV may only include a married couple. The external knowledge available to a malicious recipient is expressed as a triple (ℓ,k,m) that quantify, for each of the categories illustrated above, the amount of knowledge held by the recipient for each sensitive value s. A triple (ℓ,k,m) indicates that the recipient knows: ℓ sensitive values that the target respondent does not have; the sensitive value associated with k individuals different from the target respondent; and m individuals in the same-value family of the target respondent. For instance, suppose that, with respect to respondent *Ron* and sensitive value *H1N1*, *Alice* knows that *Ron* does not suffer from *Ovarian Cancer* and *Measles* ($\ell = 2$). Also, suppose that *Alice* already knows the diseases of other three respondents represented in the table where *Ron* is represented ($k = 3$). Finally, suppose that *Alice* knows that other five people work in *Ron*'s office ($m = 6$). As a consequence, the triple expressing such external knowledge of *Alice* is (2,3,6).

The release of a microdata table T is *safe* if the maximum probability that a data recipient can infer that an arbitrary respondent p is associated with sensitive value s in the original dataset is lower than a threshold c fixed by the data holder, assuming the recipient's external knowledge to be bounded by (ℓ, k, m). For instance, assume that the external knowledge of a malicious recipient for an arbitrary target respondent p and sensitive value *H1N1* is represented by the triple (2,5,1) and that c=0.5. A released table is considered safe if the malicious recipient cannot determine with probability higher than 0.5 that p suffers from *H1N1*, provided the malicious user knows: at most 2 sensitive attribute values that p does not have; the sensitive value of at most 5 respondents other than p; and at most 1 member in the same-value family as p. To provide the data holder with a more flexible privacy measure, in [6] the authors propose to bound the external knowledge not through a unique triple (ℓ,k,m), but through a set of triples (representing incomparable points in the three-dimensional space $\langle\ell,k,m\rangle$) that form a *skyline*. The release of a microdata table T, protected applying generalization or group-based techniques, is safe if the probability that the malicious recipient violates the privacy of the respondents is lower than threshold c, assuming that the external knowledge of the recipient is bounded by the skyline (e.g., the grey area in Figure 12, depicted in only two dimensions for simplicity, represents the area in which the external knowledge of the recipient is bounded). In [6] the authors extend their definition of safe release to take into account the fact that each respondent may be associated with more than

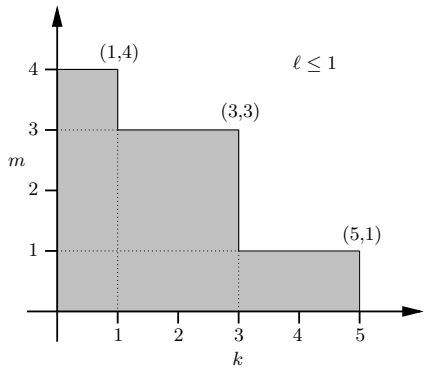

Fig. 12: An example of privacy skyline obtained fixing three points {(1,1,4),(1,3,3),(1,5,1)}

one sensitive attribute value (e.g., a respondent that suffers from more than one disease).

ϵ-Privacy [28] is aimed at defining a realistic adversarial model. Indeed, traditional approaches either assume that the adversary has a very limited knowledge, or an infinite knowledge. Both these assumptions are however unrealistic. To overcome this issue, in [28] the authors propose to measure the privacy degree offered to a respondent p as the difference in the recipient's belief about the respondent's sensitive attribute value when p belongs to the released table and when p does not belong to the same. This measure is based on the observation that looking at the released dataset, a data recipient can learn information about the sensitive value of a respondent p, even if the tuple representing p is not released [22]. Therefore, it is necessary to consider this unavoidable gain of information when evaluating the quality of a released table. In particular, ϵ-privacy states that the release of a table T obtained through generalization exposes the privacy of a respondent if the ratio $\frac{p^{in}}{p^{out}}$ is greater than a threshold ϵ fixed by the data holder, where p^{in} (p^{out}, respectively) is the probability that a recipient infers the sensitive value associated with an arbitrary respondent p when p is not represented in T (belongs to T, respectively). To compute p^{in} and p^{out}, the authors consider two different kinds of external knowledge: *i)* full information on a subset of the tuples in T; and *ii)* information on the distribution of the sensitive attribute values in the dataset. We note that, although both (c, k)-safety and privacy skyline consider the first kind of external knowledge, they do not make any assumption on the second kind of knowledge modeled by ϵ-privacy.

7 Differential privacy

The proposals described so far typically measure the disclosure risk associated with the release of a microdata table as the increase of the probability that an

adversary may correctly guess the identity or the values of sensitive attributes of a respondent represented in the table. Such approaches do not consider that the microdata table can also be exploited for inferring information of respondents that are not represented in the table. In the remainder of this section, we first illustrate in more details such a privacy issue, and then briefly describe differential privacy, a recent privacy notion that is becoming popular in the data protection community.

7.1 Problem

One of the first definitions of *privacy* in data publishing scenarios states that: *anything that can be learned about a respondent from the statistical database should be learnable without access to the database* [13]. Although this definition has been thought for statistical databases, it is also well suited for the microdata publishing scenario. Unfortunately, this definition of ideal privacy cannot be achieved by any privacy-aware microdata publication technique that is aimed at preserving data utility. In fact, as proved in [16], only an empty dataset can guarantee absolute disclosure prevention. This is also due to the fact that the release of a dataset may violate the privacy of any respondent, independently of whether the respondent is represented in the dataset. For instance, suppose that the released dataset permits to compute the average annual income of people living in city A for each ethnic group, and suppose that this information is not publicly available (and therefore a malicious recipient can only gain this information by looking at the released table). Assume also that *Alice* knows that *Bob*'s annual income is 1,000$ more than the average annual income of *Asian* people living in city A. Although this piece of information alone does not permit *Alice* to gain any information about *Bob*'s annual income, if combined with the released dataset, it allows *Alice* to infer *Bob*'s annual income. It is important to note that the disclosure of *Bob*'s annual income does not depend on his representation in the released dataset.

The solutions proposed in the literature for protecting microdata tables implicitly assume that the privacy of individuals not included in the dataset is not at risk. As a consequence, they cannot be adopted to prevent the attack described above.

7.2 Solution

Differential privacy [16] is a novel privacy notion whose goal is to guarantee that the release of a microdata table does not disclose sensitive information about any individual, represented or not by a tuple in the table. In particular, a data release is considered safe if the inclusion in the dataset of tuple t_p, related to respondent p, does not change the probability that a malicious recipient can correctly identify the sensitive attribute value associated with p. More formally, given two datasets T and T' differing only for one tuple t_p, an arbitrary randomized function $\mathcal{K}$ operating on the dataset satisfies ϵ-*differential privacy* if and only if $P(\mathcal{K}(T) \in S) \leq \exp(\epsilon) \cdot P(\mathcal{K}(T') \in S)$, where S is a subset of the

possible outputs of function $\mathcal{K}$ and ϵ is a public privacy parameter. Intuitively, ϵ-differential privacy holds if the removal (insertion, respectively) of one tuple t_p from (into, respectively) the dataset does not significatively affect the result of the evaluation of function $\mathcal{K}$. As an example, consider an insurance company that consults a medical dataset to decide whether an individual p is eligible for an insurance contract. If differential privacy is satisfied, the presence or absence of tuple t_p representing p in the dataset does not significantly affect the final decision by the insurance company. It is also important to note that the external knowledge that an adversary may possess cannot be exploited for breaching the privacy of individuals. In fact, the knowledge that the recipient gains looking at the released dataset is bounded by a multiplicative factor $\exp(\epsilon)$, for any individual either represented or not in the released microdata table.

Differential privacy is applicable to both the *non-interactive* publishing scenario (i.e., public release of a dataset) and the *interactive* publishing scenario (i.e., evaluation of queries over a private dataset). The techniques proposed in the literature to guarantee ϵ-differential privacy are based on the addition of noise, and therefore do not preserve data truthfulness. To achieve differential privacy, by definition, it is necessary to hide the presence (or absence) of the tuple associated with an individual. Therefore, considering simple count queries, the query result returned to the requesting recipient may differ by at most one (0 if the tuple is not removed; 1 otherwise) from the result computed on the original dataset. To compute the query result, the solution proposed in [17] consists in adding random noise to the query result evaluated on the original dataset. The distribution considered for the random noise is the *Laplace distribution* $Lap(\Delta(f)/\epsilon)$ with probability density function $P(x|b) = exp(-|x|/b)/2b$, where $b = \Delta(f)/\epsilon$ and $\Delta(f)$ is the maximum difference between the query result evaluated over T and over T' (which is equal to 1 for count queries, since T and T' differ for one tuple). The addition of independently generated noise, with distribution $Lap(\Delta(f)/\epsilon)$, to the query result guarantees ϵ-differential privacy [17]. This strategy can also be adopted in a non-interactive scenario, where the data holder releases the frequency matrix representing the dataset and each cell in the matrix is the result of a count query. The frequency matrix has a dimension for each attribute in the table and the entries in each dimension are labeled with the values in the attribute domain. Each cell in the matrix reports the number of tuples in the table with value, for each attribute, equal to the label of the corresponding entry in the frequency matrix.

8 Conclusions

The public and semi-public release of large microdata collections is becoming increasingly popular, thanks to the high availability of computational power at low prices, which makes data analysis an easy task for most data recipients. Although microdata collections represent a valuable resource for data recipients, the release of fine-grained information related to single individuals may put the privacy of respondents at risk. In this chapter, we first illustrated the traditional

approaches designed for preventing identity and attribute disclosure in microdata publishing. We then discussed the privacy risks that may arise when changing the underlying assumptions, and described some techniques recently proposed in the literature to overcome these privacy risks.

Acknowledgments

This work was supported in part by the EU within the 7FP project "PrimeLife" under grant agreement 216483, by the Italian Ministry of Research within the PRIN 2008 project "PEPPER" (2008SY2PH4), and by the Università degli Studi di Milano within the "UNIMI per il Futuro - 5 per Mille" project "PREVIOUS".

References

1. Aggarwal, G., Feder, T., Kenthapadi, K., Motwani, R., Panigrahy, R., Thomas, D., Zhu, A.: Approximation algorithms for k-anonymity. Journal of Privacy Technology (November 2005)
2. Azzini, A., Marrara, S., Sassi, R., Scotti, F.: A fuzzy approach to multimodal biometric continuous authentication. Fuzzy Optimization and Decision Making 7(3), 215–302 (November 2008)
3. Barbaro, M., Zeller, T.: A face is exposed for AOL searcher no. 4417749. New York Times (August 9 2006)
4. Bayardo, R.J., Agrawal, R.: Data privacy through optimal k-anonymization. In: Proc. of the 21st IEEE International Conference on Data Engineering (ICDE 2005). Tokyo, Japan (April 2005)
5. Bezzi, M., De Capitani di Vimercati, S., Livraga, G., Samarati, P.: Protecting privacy of sensitive value distributions in data release. In: Proc. of the 6th Workshop on Security and Trust Management (STM 2010). Athens, Greece (September 2010)
6. Chen, B.C., LeFevre, K., Ramakrishnan, R.: Privacy skyline: Privacy with multidimensional adversarial knowledge. In: Proc. of the 33rd International Conference on Very Large Data Bases (VLDB 2007) (September 2007)
7. Cimato, S., Gamassi, M., Piuri, V., Sassi, R., Scotti, F.: Privacy-aware biometrics: Design and implementation of a multimodal verification system. In: Proc. of the 24th Annual Computer Security Applications Conference (ACSAC 2008). Anaheim, CA, USA (December 2008)
8. Ciriani, V., De Capitani di Vimercati, S., Foresti, S., Jajodia, S., Paraboschi, S., Samarati, P.: Fragmentation design for efficient query execution over sensitive distributed databases. In: Proc. of the 29th International Conference on Distributed Computing Systems (ICDCS 2009). Montreal, Canada (June 2009)
9. Ciriani, V., De Capitani di Vimercati, S., Foresti, S., Jajodia, S., Paraboschi, S., Samarati, P.: Combining fragmentation and encryption to protect privacy in data storage. ACM Transactions on Information and System Security (TISSEC) 13(3), 22:1–22:33 (July 2010)
10. Ciriani, V., De Capitani di Vimercati, S., Foresti, S., Samarati, P.: k-Anonymity. In: Yu, T., Jajodia, S. (eds.) Secure Data Management in Decentralized Systems. Springer-Verlag (2007)

11. Ciriani, V., De Capitani di Vimercati, S., Foresti, S., Samarati, P.: Microdata protection. In: Yu, T., Jajodia, S. (eds.) Secure Data Management in Decentralized Systems. Springer-Verlag (2007)
12. Ciriani, V., De Capitani di Vimercati, S., Foresti, S., Samarati, P.: Theory of privacy and anonymity. In: Atallah, M., Blanton, M. (eds.) Algorithms and Theory of Computation Handbook (2nd edition). CRC Press (2009)
13. Dalenius, T.: Towards a methodology for statistical disclosure control. Statistik Tidskrift 15, 429–444 (1977)
14. De Capitani di Vimercati, S., Foresti, S., Jajodia, S., Paraboschi, S., Samarati, P.: Encryption policies for regulating access to outsourced data. ACM Transactions on Database Systems (TODS) 35(2), 12:1–12:46 (April 2010)
15. De Capitani di Vimercati, S., Foresti, S., Jajodia, S., Paraboschi, S., Samarati, P.: Fragments and loose associations: Respecting privacy in data publishing. Proc. of the VLDB Endowment 3(1), 1370–1381 (September 2010)
16. Dwork, C.: Differential privacy. In: Proc. of the 33rd International Colloquium on Automata, Languages and Programming (ICALP 2006). Venice, Italy (2006)
17. Dwork, C., Mcsherry, F., Nissim, K., Smith, A.: Calibrating Noise to Sensitivity in Private Data Analysis. In: Proc. of the 3rd Theory of Cryptography Conference (TCC 2006). New York, NY, USA (March 2006)
18. Frikken, K.B., Zhang, Y.: Yet another privacy metric for publishing micro-data. In: Proc. of the 7th Workshop on Privacy in Electronic Society (WPES 2008). Alexandria, VA, USA (October 2008)
19. Gamassi, M., Lazzaroni, M., Misino, M., Piuri, V., Sana, D., Scotti, F.: Accuracy and performance of biometric systems. In: Proc. of the 2004 IEEE Instrumentation & Measurement Technology Conference (IMTC 2004). Como, Italy (May 2004)
20. Gamassi, M., Piuri, V., Sana, D., Scotti, F.: Robust fingerprint detection for access control. In: Proc. of the 2nd RoboCare Workshop (RoboCare 2005). Rome, Italy (May 2005)
21. Golle, P.: Revisiting the uniqueness of simple demographics in the US population. In: Proc. of the 5th Workshop on Privacy in the Electronic Society (WPES 2006). Alexandria, VA, USA (October 2006)
22. Kifer, D.: Attacks on privacy and deFinetti's theorem. In: Proc. of the 2009 ACM SIGMOD International Conference on Management of Data (SIGMOD 2009). Providence, RI, USA (June 2009)
23. LeFevre, K., DeWitt, D.J., Ramakrishnan, R.: Incognito: Efficient full-domain k-anonymity. In: Proc. of the 2005 ACM SIGMOD International Conference on Management of Data (SIGMOD 2005). Baltimore, MD, USA (June 2005)
24. LeFevre, K., DeWitt, D.J., Ramakrishnan, R.: Mondrian multidimensional k-anonymity. In: Proc. of the 22nd IEEE International Conference on Data Engineering (ICDE 2006). Atlanta, GA, USA (April 2006)
25. Li, F., Sun, J., Papadimitriou, S., Mihaila, G., Stanoi, I.: Hiding in the crowd: Privacy preservation on evolving streams through correlation tracking. In: Proc. of the 23rd IEEE International Conference on Data Engineering (ICDE 2007). Istanbul, Turkey (April 2007)
26. Li, N., Li, T., Venkatasubramanian, S.: t-closeness: Privacy beyond k-anonymity and ℓ-diversity. In: Proc.of the 23rd IEEE International Conference on Data Engineering (ICDE 2007). Istanbul, Turkey (April 2007)
27. Li, T., Li, N.: On the tradeoff between privacy and utility in data publishing. In: Proc. of the 15th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD 2009). Paris, France (June - July 2009)

28. Machanavajjhala, A., Gehrke, J., Götz, M.: Data publishing against realistic adversaries. Proc. of the VLDB Endowment 2(1), 790–801 (August 2009)
29. Machanavajjhala, A., Kifer, D., Gehrke, J., Venkitasubramaniam, M.: ℓ-diversity: Privacy beyond k-anonymity. ACM Transactions on Knowledge Discovery from Data (TKDD) 1(1), 3:1–3:52 (March 2007)
30. Martin, D.J., Kifer, D., Machanavajjhala, A., Gehrke, J., Halpern, J.Y.: Worst-case background knowledge for privacy-preserving data publishing. In: Proc. of the 23rd IEEE International Conference on Data Engineering (ICDE 2007). Istanbul, Turkey (April 2007)
31. Meyerson, A., Williams, R.: On the complexity of optimal k-anonymity. In: Proc. of the 23rd ACM-SIGMOD-SIGACT-SIGART Symposium on the Principles of Database Systems (PODS 2004). Paris, France (June 2004)
32. Narayanan, A., Shmatikov, V.: Robust de-anonymization of large sparse datasets. In: Proc. of the 2008 IEEE Symposium on Security and Privacy (SP 2008). Berkeley/Oakland, CA, USA (May 2008)
33. Narayanan, A., Shmatikov, V.: Myths and fallacies of "personally identifiable information". Communications of the ACM (CACM) 53, 24–26 (June 2010)
34. Nergiz, M.E., Atzori, M., Clifton, C.: Hiding the presence of individuals from shared databases. In: Proc. of the 2007 ACM SIGMOD International Conference on Management of Data (SIGMOD 2007). Beijing, China (June 2007)
35. Nergiz, M., Clifton, C., Nergiz, A.: Multirelational k-anonymity. In: Proc. of the 23rd IEEE International Conference on Data Engineering (ICDE 2007). Istanbul, Turkey (April 2007)
36. Pei, J., Tao, Y., Li, J., Xiao, X.: Privacy preserving publishing on multiple quasi-identifiers. In: Proc. of the 25th IEEE International Conference on Data Engineering (ICDE 2009). Shanghai, China (March - April 2009)
37. Samarati, P.: Protecting respondents' identities in microdata release. IEEE Transactions on Knowledge and Data Engineering (TKDE) 13(6), 1010–1027 (November 2001)
38. Terrovitis, M., Mamoulis, N., Kalnis, P.: Privacy-preserving anonymization of set-valued data. Proc. of the VLDB Endowment 1, 115–125 (August 2008)
39. Wang, K., Fung, B.C.M.: Anonymizing sequential releases. In: Proc. of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD 2006). Philadelphia, PA, USA (August 2006)
40. Wang, K., Xu, Y., Wong, R., Fu, A.: Anonymizing temporal data. In: Proc. of the 2010 IEEE International Conference on Data Mining (ICDM 2010). Sydney, Australia (December 2010)
41. Xiao, X., Tao, Y.: Anatomy: Simple and effective privacy preservation. In: Proc. of the 32nd International Conference on Very Large Data Bases (VLDB 2006). Seoul, Korea (September 2006)
42. Xiao, X., Tao, Y.: Personalized privacy preservation. In: Proc. of the 2006 ACM SIGMOD International Conference on Management of Data (SIGMOD 2006). Chicago, IL, USA (June 2006)
43. Xiao, X., Tao, Y.: m-invariance: Towards privacy preserving re-publication of dynamic datasets. In: Proc. of the 2007 ACM SIGMOD International Conference on Management of Data (SIGMOD 2007). Beijing, China (June 2007)
44. Xiao, X., Yi, K., Tao, Y.: The hardness and approximation algorithms for ℓ-diversity. In: Proc. of the 13th International Conference on Extending Database Technology (EDBT 2010). Lausanne, Switzerland (March 2010)

45. Zhou, B., Han, Y., Pei, J., Jiang, B., Tao, Y., Jia, Y.: Continuous privacy preserving publishing of data streams. In: Proc. of the 12th International Conference on Extending Database Technology (EDBT 2009). Saint Petersburg, Russia (March 2009)