IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received February 4, 2020, accepted March 1, 2020, date of publication March 11, 2020, date of current version March 23, 2020.

Digital Object Identifier 10.1109/ACCESS.2020.2980235

A Survey on Privacy Properties for Data Publishing of Relational Data

ATHANASIOS ZIGOMITROS[1], FRAN CASINO[1], (Member, IEEE), AGUSTI SOLANAS[2], (Senior Member, IEEE), AND CONSTANTINOS PATSAKIS[1,3], (Member, IEEE)

[1]Department of Informatics, University of Piraeus, 185 34 Pireas, Greece
[2]Department of Computer Engineering and Mathematics, Rovira i Virgili University, 43007 Tarragona, Spain
[3]Information Management Systems Institute, Athena Research and Innovation Center, 151 25 Marousi, Greece

Corresponding author: Constantinos Patsakis (kpatsak@unipi.gr)

This work was supported in part by the European Commission through the Horizon 2020 Programme (H2020), in part by the OPERANDO Project under Grant 653704, in part by the YAKSHA Project under Grant 780498, and in part by the LOCARD Project under Grant 832735. The work of Agusti Solanas was supported in part by the Government of Catalonia (GC) under Grant 2017-DI-002 and Grant 2017-SGR-896, in part by the Fundació PuntCAT with the Vinton Cerf Distinction, and in part by the Spanish Ministry of Science and Technology under Project RTI2018-095499-B-C32. The content of this article does not reflect the official opinion of the European Union. Responsibility for the information and views expressed therein lies entirely with the authors.

ABSTRACT Recent advances in telecommunications and database systems have allowed the scientific community to efficiently mine vast amounts of information worldwide and to extract new knowledge by discovering hidden patterns and correlations. Nevertheless, all this shared information can be used to invade the privacy of individuals through the use of fusion and mining techniques. Simply removing direct identifiers such as name, SSN, or phone number is not anymore sufficient to prevent against these practices. In numerous cases, other fields, like gender, date of birth and/or zipcode, can be used to re-identify individuals and to expose their sensitive details, e.g. their medical conditions, financial statuses and transactions, or even their private connections. The scope of this work is to provide an in-depth overview of the current state of the art in Privacy-Preserving Data Publishing (PPDP) for relational data. To counter information leakage, a number of data anonymisation methods have been proposed during the past few years, including k-anonymity, ℓ-diversity, t-closeness, to name a few. In this study we analyse these methods providing concrete examples not only to explain how each of them works, but also to facilitate the reader to understand the different usage scenarios in which each of them can be applied. Furthermore, we detail several attacks along with their possible countermeasures, and we discuss open questions and future research directions.

INDEX TERMS Data anonymization, privacy preserving data publishing, data protection, k-anonymity, privacy, review.

I. INTRODUCTION

Knowledge is one of the main keys to innovation. In that respect, the research community continuously retrieves and analyses data to discover new knowledge, whereas corporations are examining patterns of human behaviour to enhance the quality of their provided services and products. Nevertheless, the extreme volume of knowledge that is hidden in the electronic traces of human activity has raised significant challenges which are attributed mainly to the advancements of Big Data. As data recovering techniques can retrieve a lot of sensitive personal information, many concerns about the privacy of individuals have also been emerged [1]–[3]. An adversary/attacker may exploit publicly available information to obtain more details about individuals and to extract knowledge that wouldn't be allowed to have under normal circumstances, thus invading the privacy of individuals [4]. A naïve approach to address this problem is to erase or mask the fields that are explicitly identifying an individual: name, social security number etc. Yet, such measures have been proven to be insufficient since an individual can be uniquely identified even when the explicit identifiers have been discarded from a dataset. For instance, Sweeney [5] was able to link a record to a specific individual by having access to two

The associate editor coordinating the review of this manuscript and approving it for publication was Kuo-Hui Yeh.

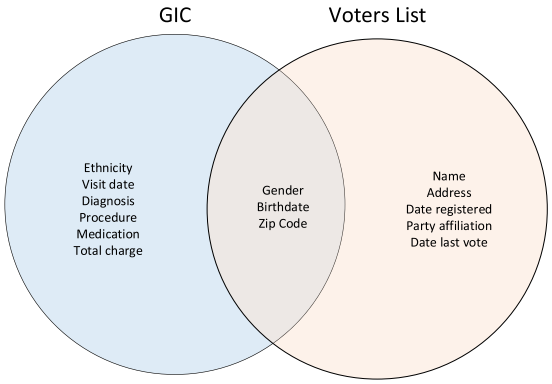


FIGURE 1. Sweeney's Example of Re-identification.

different datasets in which the explicit identifiers had been deleted. The first dataset contained the voter registration list while the other one was a patient dataset disclosed by the Group Insurance Commission (GIC). As shown in Figure 1, in her famous experiment Sweeney linked both datasets using three common fields: Gender, Zip code and birthdate. Based on the 90s census data of the US population [6], Sweeney showcased that the 87% of the population could be uniquely identified through these three specific fields. In a more recent work based on the 2000 census data [7], the percentage of the US population that can be uniquely identifiable by using the same fields decreased to 63%. More studies [8], [9] achieved similar outcomes for other countries.

Legally speaking, personally identifiable information has distinct interpretations depending on the concerned jurisdiction. For example, the California Senate Bill 1386 defines as personal identifying information the Social Security numbers (SSN), driver's license numbers - excluding licence plates, financial accounts, debit/credit card numbers, telephone numbers, and email addresses. However, the European Union uses a broader definition:

> "Any information relating to an identified or identifiable natural person. . . ; an identifiable person is one who can be identified, directly or indirectly, in particular by reference to an identification number or to one or more factors specific to his physical, physiological, mental, economic, cultural or social identity.
> . . . account should be taken of all the means likely reasonably to be used either by the controller or by any other person to identify the said person." [10]

Nonetheless, as Narayanan and Shmatikov [11] point out:

> "Any information that distinguishes one person from another can be used for re-identifying anonymous data."

The recent introduction of the General Data Protection Regulation (GDPR)[1] attempts to protect personal information from misuse and allows citizens to take back control of their data by granting them extensive data protection rights such as the Right to be Forgotten (RtbF) [12]. According to the GDPR (Article 4(1)), personal data are defined as follows:

> "Personal data" means any information relating to an identified or identifiable natural person ("data subject"); an identifiable natural person is one who can be identified, directly or indirectly, in particular by reference to an identifier such as a name, an identification number, location data, an online identifier or to one or more factors specific to the physical, physiological, genetic, mental, economic, cultural or social identity of that natural person.

Such definition implies that any data that can be used on its own or in combination with other data to identify, contact, or locate an individual has to be referred to as personal data. Practically then, under the EU law, any location information or device ID related to an individual that can uniquely bind a place or a device to a physical person is considered personal information. Clearly, this also applies to the sets of seemingly anonymous data which, when correlated with other information, can identify an individual. Overall, the GDPR enforces many constraints to organisations for collecting, storing, and processing personal data, whereas it also refrain from the unconditionally sharing of personal data [13].

Although nowadays vast amounts of information are produced in a daily basis, the discovery of new knowledge faces great challenges [14]. For example, in the context of healthcare, companies extract knowledge to utilise more accurate diagnoses, patients' treatments, and other health-related tasks. Nevertheless, extreme care should be taken when such sensitive information is to be disclosed or shared to other parties. In that respect, the current state of the art in Privacy-Preserving Data Publishing (PPDP) provides a set of appropriate approaches and technical measures for the protection of personal data. One of the main contributions of the current survey is to analyse and compare these approaches, showcasing their benefits and shortcomings.

Apparently, as illustrated in Table 1, the disclosure of any kind of unprotected non-personal data may entail - depending always on the underlying data and their correlations - a privacy risk for individuals due to the feasibility of their identification. Hence, taking also into account the high imposed sanctions for personal data breaches, it is of data publishers' best interest to guarantee the efficiency of any employed privacy-preserving methods. In this regard, and since the most common type of published datasets correspond to data stored in relational databases, this survey focuses on data anonymisation and privacy-preserving methods when publishing relational data.

A. MOTIVATION AND CONTRIBUTIONS

Undoubtedly, there are already significant PPDP surveys in the relevant literature such as those by Chen *et al.* [15]

[1] http://eur-lex.europa.eu/legal-content/EN/TXT/PDF/?uri=CELEX:32016R0679&from=EN

and Fung *et al.* [16]. While these surveys follow an in-depth approach, as in this work, they do not cover the recent advances in the area. For instance, our survey includes new methods like β-likeness, loose associations and disassociation, to name a few. Moreover, both of those surveys cover other fields as well, such as location and social network privacy. Instead, in our survey we focus on the anonymisation of relational data to prevent re-identification, and therefore we provide a more detailed overview of the current work in the field. Contrary to the most recent literature reviews of di Vimercati *et al.* [17] and Domingo-Ferrer *et al.* [18], we follow a bottom-up approach, by covering the whole anonymisation procedure and by providing several examples to facilitate the reader in understanding the different concepts and possible attacks. Another recent PPDP work focused on Differential Privacy can be found in [19].

Of particular interest to our survey is to present the different approaches of PPDP and to introduce the reader to the various privacy guarantees and attacks in the literature. To this end, having a single running example throughout the survey acting as a reference point to all cases would have sound ideal. Nonetheless, it will become apparent as we proceed with our survey that the ''one size fits all'' approach is not relevant to PPDP. This is because the methods and attacks are highly dependent on the original dataset, enabling different attack vectors and therefore requiring different countermeasures. Therefore, we facilitate the reader by employing small distinct examples which have the required properties for each case. Finally, our survey concludes by discussing open issues in the field and by providing directions for future research.

B. ORGANISATION OF THIS WORK

The rest of this work is organised as follows. Section II introduces the Privacy-Preserving Data Mining (PPDM) and Privacy-Preserving Data Publishing (PPDP) concepts and discuss their main differences. Section III outlines several notions and dimensions of privacy, while it further extends the description of the PPDP defining, among others, the actors and their respective roles, the publication process, and the metrics used to to quantify information loss. Section IV presents the data transformation methods that are used to anonymise datasets. In Section V we present the basic trends in de-anonymisation attacks on published datasets. Then, in Section VI we present the privacy models and dataset properties that allow us to countermeasure the aforesaid de-anonymisation attacks. For each case, we highlight the possible data leakages and the level of privacy exposure of individuals. Finally, we conclude this survey in Section VII by identifying and discussing a number of pertinent research directions and open issues for PPDP.

II. BACKGROUND ON PPDM AND PPDP

At a glance, PPDM focuses on data mining tasks in a privacy-preserving way, while PPDP cares for the usefulness of the data even if they are analysed on record level. Therefore, unlike PPDM, in the PPDP the *truthfulness on record level* is usually a requirement. Furthermore, the PPDP does not make any assumption regarding the type of analysis to which the data would be subjected because the published data can be analysed by means of different techniques and with diverse aims. In other words, one could argue that the difference between PPDM and PPDP is where the query is executed. In PPDM, the query is executed in a controlled environment; therefore, we may install the necessary ''watchdogs'' to control the information flow and preserve the privacy of individuals. However, in the case of PPDP, the dataset has been already released and the adversary can access it and execute any arbitrary query. Therefore, while PPDM enables more dynamic configuration as well as control *on the fly*, for PPDP the utilisation of privacy protection methods is mandatory prior to data release. Yet, data publishers do not usually have the technical background to provide a properly processed dataset that is optimal for data mining algorithms while it does not compromise the privacy of individuals. Moreover, anonymisation and privacy-preserving techniques can be applied more efficiently if the posterior data mining process is known in advance. Better yet, the data mining process could be performed ''in-house'' so that only the anonymised results are published. For more on PPDM, the reader may refer to [36]–[42].

In the rest of this survey we regard that a privacy breach occurs when the prior belief of an adversary about an individual differs significantly from his belief after accessing the anonymised dataset. Based on this definition, we try to showcase through examples how specific methods allow us to prevent these breaches as well as the nature of these breaches which, as we are going to discuss later, is highly dependent on the underlying dataset.

III. FUNDAMENTALS OF PRIVACY-PRESERVING DATA PUBLISHING

Depending on what is to be protected and the assumptions that we make for the knowledge and capacity of the adversary various definitions of privacy exist in the literature [43]. Nonetheless, the compliance with each formal privacy definition introduces different limitations on the sanitised data.

In terms of the dimensions of privacy, several approaches are identified in the literature. For instance, Domingo-Ferrer [44] splits database privacy issues into three dimensions related to the main actors involved: respondents (i.e. re-identification of individuals), users (i.e. guaranteeing the privacy of the queries), and owners (allowing the access only to specific subsets of data). Other authors such as Martinez *et al.* [45], [46] and Solanas *et al.* [47] adopt a broader definition of the different privacy contexts which are classified into five dimensions: (i) identity privacy ([48], [49]), which - similarly to respondents' privacy above- refers to the non-disclosure of individuals identities, (ii) query privacy ([50]–[52]), which is analogous to the aforementioned user privacy and refers to the privacy of formulating queries and retrieving information, (iii) location privacy [53], which focuses on protecting the physical location

TABLE 1. Re-identification attack on real datasets [20].

Reference	Year	Individuals re-identified	Proper de-identification of attacked data ?	Re-identification verified ?
Employment statistics of Germany (register data) and the German Life History Study (survey data) [21]	2001	29 of 273	Factually anonymous	Yes (records containing insurance numbers only)
Comparison of Chicago homicide dataset with records in the Social Security Death Index [22]	2001	75% Of 11,000	Direct identifiers removed	No
Group Insurance Commission patient-specific data with nearly one hundred attributes and voter registration list for Cambridge Massachusetts [5]	2002	1 of 135,000	Removal of names and addresses	Yes
Two British datasets, the first dataset was the 2% individual Sample of anonymised records from the 1991 Census which was compared against a subset of health related variables, drawn from the General Household Survey (GHS) for 1991. [23]	2003	219 unique matches, 112 with 2 possibilities, 8 confirmed	Yes	Verified matches, but not identities
Web search queries collected by AOL [4]	2006	1 of 657,000	No	Yes (with individual)
Maps of Boston with the addresses of patients plotted as individual dots or symbols present medical journals papers [24]	2006	79% of 550	No	Verified (with original data set)
A set of movie ratings and a set of movie opinions from MovieLens [25]	2006	Of 133 users, 60% of those who mention at least 8 movies	Direct identifiers removed	No
Persons re-identified from the Illinois Health and Hazardous Substances Registry [26]	2006	18 of 20	Only type of cancer, zip code and date of diagnosis included in request	Yes (verified by the Department of Health)
Identification of genomic risk for specific family relations [27]	2006	70%	No	Yes
The network of friendship links on the blogging site LiveJournal. [28]	2007	2,400 of 4.4 million	Identifying information removed	Verified using original data
The death report of a 26 year-old student consuming a particular drug, which aired on a Canadian national broadcaster, compared with data from the adverse drug reaction database released by Health Canada lead to the re-identification of the deceased [29]	2007	1	Direct Identifiers removed & possibly other unknown de-id methods used	Yes
Anonymous movie ratings of 500000 subscribers of Netflix and 50 IMDB users [30]	2008	2 of 50	Direct identifiers removed & maybe perturbation	No
Prescription data disclosed by pharmacies without patient consent [31]	2009	1 of 3,510	Direct identifiers removed	Yes
Users have accounts on both Twitter and Flickr, re-identified in the anonymous Twitter graph [32]	2009	30.8% of 150 pairs of nodes	Identifying information removed	Verified using ground-truth mapping of the 2 networks
Admission records of hispanics in a hospital system between 2004 and 2009 [33], [34]	2010	2 of 15,000	Yes - HIPAA Safe Harbor	Yes
A complete dump of historical trip and fare logs from NYC taxis[2]	2014	100%	Medallion and licence numbers obfuscated	Verified using rainbow tables
Complete de-anonymisation of South Korean Resident Registration Numbers [35]	2015	23,163	Encrypted RRN	Yes

of individuals interacting with a system, (iv) footprint privacy ([54], [55]), which guarantees that microdata collected from users' interactions with a specific system (i.e. the footprint of the users) are properly sanitised before publishing so as to control the amount of information extracted or inferred from these microdata sets, and (v) Intelligence/Owner privacy, which refers to the owner privacy described in [44] and focuses on the disclosure of data when different parties collaborate (cross-domain and joint queries).

In relation to the above privacy classifications, PPDP techniques aim to publish useful information while protecting the privacy of the individuals in the dataset. Relevant literature defines two types of information disclosure when data are released/published: identity disclosure and attribute disclosure [56]. The former occurs when a specific respondents' identity is associated with a disseminated/disclosed data record containing confidential or private information [57], while the latter occurs when a specific respondent is associated with either an attribute value in the disseminated/disclosed data or an estimated attribute value based on the disseminated data [57]. In order to avoid information disclosure, the data need to be disguised/transformed in a way that prevents an adversary/attacker from linking records of the dataset with specific individuals. Careless disclosure of data could lead to serious privacy breaches, as already seen in a myriad of cases [4], [5], [21]–[35], [58], many of which are reported by El Emam *et al.* in [20] (see Table 1 for more details). Yet, the authors of the report undermined many of them in terms of their actual impact. In principle, it can be easily noticed that the number of attacks in real datasets is relatively low to justify a thorough research in the field. In addition, as Sweeney highlights [59], research contributions on this area often face issues related to their publications due to the fear of consequent legal implications. Moreover, one can assume attacks performed by adversaries that have never surfaced publicly. Contrary to popular belief, an attack may not always be executed by a ''hacker''. For example, an ad service provider can perform a re-identification attack to provide personalised ad recommendations according to each user profile, obtaining thereby a significant advantage over its competitors. On the other hand, the recent example of the attack on Ashley Madison[2] clearly reflects how hackers can use disclosed data to extort individuals. Nonetheless, such attacks showcase the trade-off between the privacy of individuals and the usability of data. Regardless of how nefarious

[2]http://www.reuters.com/article/us-ashleymadison-cybersecurity-idUSKCN0QN2BN20150819

such attacks can be or the attacker's intentions, other aspects of privacy must be considered as well, such as the dignity-based theory of privacy [60]. Indeed, privacy is inseparably linked to the fundamental right of respecting one's life, to the right of controlling the information flow about oneself [61], and "to selectively reveal oneself to the world" [62].

In what follows, we introduce the basic concepts and definitions of PPDP: attributes, actors, publishing scenarios and methods, and information metrics.

A. ATTRIBUTES

Let us assume a tabular dataset T, comprised of t records, where each record corresponds to a particular entity. We classify the attributes of each record into four categories:

- **Explicit Identifiers** are attributes that can uniquely identify an individual, such as the Social Security Number (SSN).
- **Quasi-Identifiers** ($\mathcal{QI}$) are publicly known attributes/features of individuals that might be used by an attacker. Note that a $\mathcal{QI}$ cannot be used to uniquely identify a person by itself. However, a combination of quasi-identifiers can lead to re-identification by diminishing the possible identities of a specific record. As a result, this increases the confidence of an adversary regarding the real identity behind an anonymised record. Typical examples of $\mathcal{QI}$ attributes are Gender, Zip Code, and Age.
- **Sensitive Attributes** (*SA*s) are the fields that store sensitive/personal information. Therefore, *SA*s store the information that an adversary most probably wishes to know. Well-known examples of *SA*s are the Salary or the Disease of an individual in a financial or medical dataset respectively. While in general there is only one *SA* in a dataset, this is not always the case [63]–[67].
- **Non-Sensitive Attributes** are attributes - other than identifiers, quasi-identifiers and sensitive attributes - which contain non-sensitive information about an individual. Still, this type of attributes cannot be ignored when protecting a dataset, since they can be part of a $\mathcal{QI}$. For example, attributes such as Job and Town may not be considered confidential or private information. However, when the population under examination is low, there may be only few people, or even just one individual, that fit to these specific search criteria. Therefore, under certain conditions the combination of these attributes could lead to identity disclosure with high probability.

It is worth noticing that in some cases the distinction between a $\mathcal{QI}$ and an *SA* is not straightforward. Imagine a supermarket basket where some products may be considered not sensitive but correlated with some sensitive ones can re-identify the buyer in an anonymised table.

An example of a table with Explicit Identifiers *EI*s, $\mathcal{QI}$s and *SA*s is illustrated in Table 2.

TABLE 2. Attribute classification - Example.

EI	QI			SA
SSN	Gender	Age	Zip Code	Disease
988-11-2228	M	31	12000	Flu
988-11-2221	M	35	12500	HIV
988-11-2261	F	37	12500	Cancer
988-11-2222	F	39	15400	HIV

Misclassifying an attribute A_i as *SA* when there is more than one *SAs* allowing an adversary to access to it may compromise the anonymity of the data publishing scheme, as it potentially exposes the other sensitive values. Misclassifying an attribute A_i as *SA*, while there are other unclassified *SAs*, would not only expose the other sensitive values in case an adversary gets access to the dataset, but it may potentially compromise the anonymity of the data publishing scheme. Taking into account that in most anonymisation techniques the *SA* is almost never altered, having access to an external table with this sensitive information could always be beneficial to any adversary. Suppose that the Salary and Disease are classified as SA_1 and SA_2 respectively. If an attacker has access to an external table with information about the SA_1, he could perform an attack on SA_2 by correlating the SA_1 along with other *QI*s. On the other hand, misclassifying a SA_i as *QI* could lead to less qualitative anonymisation results, as we are going to discuss afterwards, due to the curse of dimensionality.

B. ACTORS

Typically, an anonymisation scenario involves the following actors:

- **Data Holder/Publisher:** The organisation or the person that holds the data that need to be anonymised to avoid privacy breaches. While typically the data holder and publisher roles correspond to the same organisation or person, sometimes the data holder can outsource the anonymisation process to another organisation or person due to the lack of knowledge or resources. In this case, the data holder and the data publisher roles correspond to two distinct actors.
- **Record Owners:** Every entity that pertains to one or more records in the dataset that is going to be released.
- **Data Recipient:** Anyone that has access to the anonymised dataset.
- **Adversary:** A malicious entity whose goal is to obtain knowledge about the sensitive attributes of an individual or a subset of attributes that could lead to re-identification.

C. SINGLE AND MULTIPLE-RELEASE PUBLISHING

There exist three main publishing scenarios each having different privacy requirements dictated by the target audience and the usage of the published data. In each case,

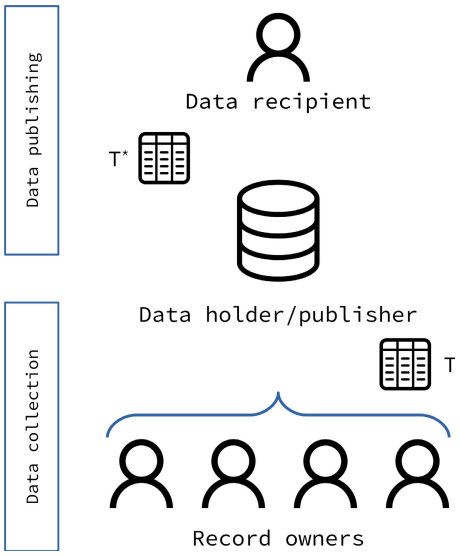


FIGURE 2. Roles in Data Collection/Publishing.

identity, query, location, footprint and owner privacy need to be protected from attackers by guaranteeing non-disclosure of sensitive data. Such scenarios can be categorised as follows:

1) PUBLISHING A SINGLE RELEASE

In this scenario, we assume that the data publisher holds an original table T and the anonymisation process is performed only once, based on the privacy guarantees that the data publisher wants to achieve. The original data T, or a subset of them, have not been previously published and they will not be published again in the future. This is the most frequent scenario in the literature.

2) PUBLISHING RELEASES IN PARALLEL

Parallel releases ([68]–[70]) refer to the case in which the original dataset T is released in several different T_i^* anonymised datasets. Each T_i^* may contain a different subset of the original attributes. For example, from Table 2, two anonymous tables can be released. The first table T_1^* may contain the attributes: Gender, Age, Disease, while the second table T_2^* may contains the attributes: Age, Zip Code, Disease.

The main motivation behind parallel releases of data is to minimise the information loss that stems from publishing a single dataset with the complete set of attributes. In this case, identity and attribute disclosures have to be prevented in the context of the privacy dimensions described in Section III. In addition, parallel releases target disparate recipients who are interested in different attributes. The specific preferences of the recipients result in disparate levels of anonymisation for each release. Nevertheless, the data publisher should take into account that the data recipients may try to collude by combining the released anonymised tables that hold the same *QIs* to obtain additional information.

3) PUBLISHING RELEASES IN SEQUENCE

The scenario of *sequential publishing* ([71]–[91]) examines the incremental release of anonymised data. For instance, consider a company that publishes periodically anonymised data about its clients. Take also into account that each publication may refer to clients whose data were also present in previous publications. Due to the course of time, the data in T are expected to have changed, usually with the addition, alteration or deletion of records. Hence, the data publisher should consider previous data releases, as the privacy of a record owner could be compromised simply by cross-examining them. In [92] the authors study the privacy of released datasets considering also that *SA* values could suffer modifications over time. For instance, the value of a *SA Disease* changed from ''flu'' to ''fever'' in subsequent releases.

D. CENTRALISED VS DECENTRALISED DATA PUBLISHING

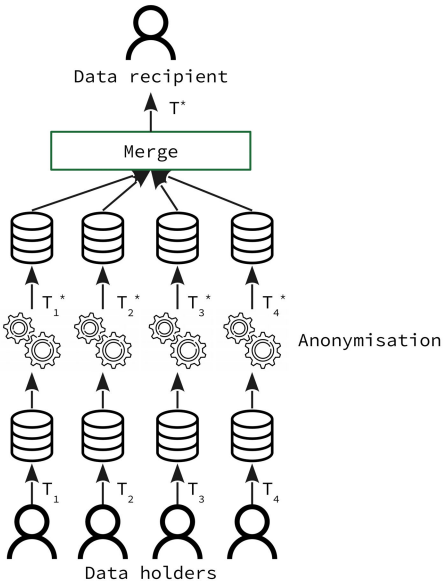
Centralised data publishing, which is performed by a data publisher that holds - or gets from a set of data holders - the entire original dataset, is the focal point of the relevant academic literature [93], [94]. For example, a Ministry of Health could gather, under certain legal conditions, all the datasets from the country's hospitals. Provided always that legal and data protection safeguards are in place, the advantages of this approach derive from the fact that each data holder does not need to anonymise on its own the dataset they wish to provide to the recipients. The data publisher, who has probably more expertise in anonymisation techniques and more computational resources, implements the anonymisation of the entire collection of datasets. Additionally, as the data publisher holds the entire collection of the datasets can handle the trade-off between disclosure risk and data utility better as she has full overview of the data. Clearly, in this scenario, the data publisher must be trusted by all data holders, as risk-wise, it is a single point of failure. Nonetheless, this is not always the case due to the legal, ethical and commercial constraints involved when private information is to be transferred across different data holders/controllers. Therefore, there might be cases of data holders not granting access to their corresponding raw data and as a result the data may be shared among multiple parties [95]–[98]. In this scenario, data might be partitioned between different parties in several ways [99]–[104]:

- Vertical partitioning (VP): In this case, data holders have only disjoint sets of attributes corresponding to the same set of individuals. This situation can be found within third parties of the same data provider, but usually VP is more suitable for obtaining knowledge about specific individuals by crossing large amounts of information of different kinds of datasets.
- Horizontal partitioning (HP): disparate parties hold disjoint sets of individuals with the same set of attributes. International communities and e-commerce companies with related topics are suitable for this kind of data partition model.

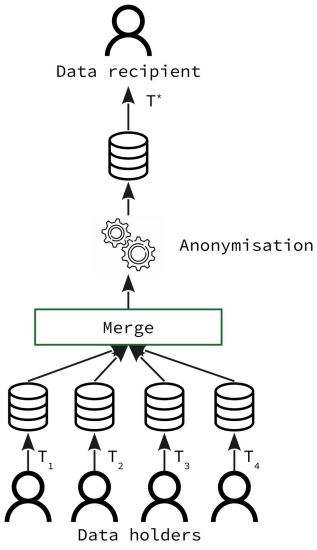
- Arbitrary partitioning (AP): there is no specific pattern of how data are distributed. If the entire set is defined by an $m \times n$ individual-attribute matrix, one party p_1 holds a subset of individuals $m_{p_1} \leq m$ whilst another party p_2 holds the rest $m_{p_2} = m - m_{p_1}$ and the same is applied for attributes as well. Note that VP and HP are specific cases of AP. The AP is the most pragmatic and widely used scheme since commonly datasets contain an undetermined number of coincident attributes and individuals.

Based on the above analysis, the data may be distributed among multiple data holders who wish to release their data in a common anonymised table T^*. The probable solutions for this decentralised scenario, are the following:

- **Anonymise-and-Aggregate:** In this case, data holders anonymise their data independently and then they all aggregate their tables to create a single table [105]. While this approach enables fast data publishing, it may also lead to unnecessary degradation of the data utility.Yet, the cost is relatively manageable.
- **Aggregate-and-Anonymise:** A more proper solution for less data distortion and better privacy guarantees is first all the distributed data to be aggregated and then the anonymisation process to be performed. To achieve this without compromising the privacy of individuals, either a semi-trusted third party need to be introduced or, alternatively due to the legal issues that prohibit the sharing of data between parties, Secure Multi-party Computation protocols should be used [106], [107]. In the latter case, the computations are performed over encrypted data, enabling the functionality without information leakages. Clearly, however, the use of encryption introduces significant computational overhead. An example of such collaborative data publishing scenarios is depicted in Figure 3.



(a) Anonymise and Aggregate.



(b) Aggregate and Anonymise.

FIGURE 3. Collaborative data publishing methods.

E. INFORMATION METRICS

Apparently, every table T can be transformed into an anonymous table T^*, e.g. by generalising all the values to the maximum level. However, while such a table would be safe to publish, it would not have any practical value for the recipients. Therefore, the primary concern when publishing a table is to balance the trade-off between privacy and data utility. In this regard, the claim that privacy is decreased it actually means that an adversary can learn/discover easier the sensitive attributes of an individual. On the other side, the utility of the data refers to their accuracy after their processing, so that one could be able to perform data mining on the sanitised data and receive useful results. This balance is quite biased as even moderate privacy measures result in significant data loss [108]. Nonetheless, the goal of PPDP is to maximise the information of the anonymised dataset, subject to the privacy constraints that were set by the data publisher. There exist a set of information metrics that are used to measure the utility of an anonymised table T^*. Such information preservation metrics can be categorised into three major categories according to their purpose: *general, special*, and *trade-off purpose*.

The data publisher usually ignores how the data recipient will analyse the anonymised table T^*. Different recipients may require different levels of generalisation on different attributes. In such a case, the data publisher wants to publish a table T^* which is as similar as possible to the original table T. With that purpose, the simple *principle of minimal distortion* was introduced in [5], [109], [110]. The *minimal distortion (MD) metric* is issuing a penalty for every value that is generalised, e.g. if a_1 is generalised to A then a $MD(A) = 1$. If, on the other hand, both a_1 and a_2 are generalised to $*$, then $MD(A) = 4$ as there is a penalty of two

from generalising both a_i to A, and another two from A to $*$. The generalisation height metric [109] and the approach of Meyerson and Williams [111] can be considered specific instances of *MD* as they issue a stable penalty for each generalisation regardless of its level. Similar to *MD* is also the Normalized Certainty Penalty introduced by Xu *et al.* [112].

The metric *ILoss* was proposed by Xiao and Tao [113]. Such metric associates each cell of the table with a number between 0 and 1, where 0 is attributed when there is no generalisation and 1 when there is a total suppression. In the rest of the cases, the number is proportional to the extent of the generalisation used for that cell value. This metric allows the data publisher to add weights to the generalised attributes to reflect their relevance.

Contrary to *ILoss* and *MD* metrics, the *discernibility metric (DM)*, introduced by Skowron and Rauszer [114], charges a penalty for each value depending on the values of the rest of the attributes in the release. Therefore, values are modified so that they become indistinguishable from others with respect to the $\mathcal{QI}$. More precisely, if a record belongs to a group of size s, the penalty for the record will be s when it is generalised.

Nergiz and Clifton [115] introduced the *Ambiguity Metric (AM)* which is designed for k-anonymity frameworks. For each record r in the anonymised table T^*, *AM* considers the number of records in T that could have been generalised as r. This number is the ambiguity of r. Thus, the AM for T^* is defined as the average ambiguity of all records in T^*.

When the data publisher knows how the data recipient is going to analyse the anonymised table T^*, then he can exploit this knowledge during the *anonymisation process* to improve utility. Some argue that if the purpose already was known, then the data publisher could simply provide the results of the data mining process. In practice, however, there exist many different ways to perform data mining, whereas the data recipient might further want to extract other information in the future.

Generalisation and suppression methods in a data mining process can affect the results both negatively and positively. By all means, while they might destroy useful classification structures, in some cases this can be perceived as a positive side-effect, e.g. when over-specialisation of the attributes is perceived as noise by the data mining algorithm. Iyengar [116] introduced the *classification metric (CM)* to measure the classification error on training data. The *CM* metric charges a penalty when a record is generalised or suppressed to a group in which the record's class is not the majority class. As a data metric though, *CM* does not solve efficiently the issue of over-specialisation of values.

Trade-off metrics take into consideration both aspects of an anonymisation process, namely the privacy and the information requirements. Let's assume that an anonymisation procedure specialises a general value into child values, iteratively. In every iteration, the general value splits into as many groups as the distinct values of the child nodes. Apparently, each specialisation s gains in terms of information, which is denoted as $IG(s)$, and loses in terms of privacy, $PL(s)$. The metric Information-Gain-to-Privacy-Loss introduced by Fung *et al.* [117] computes the specialisation s that maximises the information gain for each loss of privacy: $IGPL(s) = \frac{IG(s)}{PL(s)+1}$, where the choice of $IG(s)$ and $PL(s)$ depends on the information metric and the privacy model.

The previously described metrics do not consider the distribution of attribute values in the data set. For instance, if an attribute's values follow a uniform distribution, then replacing it with the corresponding range of values would have little effect on the anonymised data because a data analyst would easily assume a uniform distribution of the values within this range. However, if the distribution of values is skewed, then the uniform distribution assumption could lead to false results. Therefore, well-known techniques such as the *Kullback-Leibler divergence* can be used to measure the difference between two probability distributions P and Q, where P represents the ''true'' distribution of data in the original table, and Q is the distribution of attributes in the anonymised table. Nevertheless, since *KL-divergence* cannot be considered as a similarity metric as it does not obey the triangle inequality, other measures such as L_p*-Norm* [118] and *Hellinger Distance* [119] have been introduced. In [120] Ye *et al.* use a search metric to guide each step of the anonymisation process. However, they take a very different approach using rough set theory to introduce a new search metric which is used to find the minimum subset that has the same classification as the attribute they want to anonymise. Kifer and Lin paved the path for an axiomatic justification of the privacy measures. First, in [121] they introduced some axioms for privacy and utility. More precisely, they introduced the axioms of *Transformation Invariance* and *Convexity* for privacy, and the axioms of *Sufficiency*, *Continuity* and *Branching* for utility. Later, Lin and Kifer [122] introduced two more axioms, namely *quasi-convexity* and *quasi-concavity* for information preservation. The core contribution of these works is that the data publisher, depending on the application, can determine which is the most appropriate axiom. Using this axiom, he can select the corresponding metrics that would enable him to find the best balance between privacy and utility of the data. More on privacy metrics can be found in Wagner's *et al.* survey [123].

IV. DATA TRANSFORMATION TECHNIQUES

Anonymisation can be defined as a procedure of transforming a dataset T into a new dataset T^* according to some privacy requirements, as depicted in Figure 2. Therefore, anonymisation methods modify the original data to achieve the desired privacy guarantees in the anonymous dataset T^*. In what follows, we introduce the prevalent data transformation techniques, namely generalisation, suppression, bucketisation, permutation and perturbation.

A. GENERALISATION

Generalisation replaces the value of a $\mathcal{QI}$ with an abstraction of the original value, as shown at the taxonomy tree

TABLE 3. Information metrics.

Metric \ Purpose	General	Special	Trade-off	Notes
Minimal Distortion (MD) [5], [109], [110]	✓			Attribute-oriented
Height metric [109]	✓			Attribute-oriented
Normalized Certainty Penalty (NCP) [112]	✓			Attribute-oriented
ILoss [113]	✓			Cell-oriented
Discernibility [114]	✓			Record-oriented
Ambiguity metric (AM) [115]				Record-oriented
Classification metric (CM) [116]		✓		Record-oriented
Information-Gain-to-Privacy-Loss [117]			✓	Generic
L_p-Norm [118]	✓			Distribution-oriented
Hellinger distance [119]	✓			Distribution-oriented
HCE-TDR search metric [120]		✓		Rough-set based
Sufficiency, Continuity, Branching [121]	✓			Axiom-Based
Transformation invariance, Convexity [121]	✓			Axiom-Based
Quasi-convexity and Quasi-concavity [122]	✓			Axiom-Based

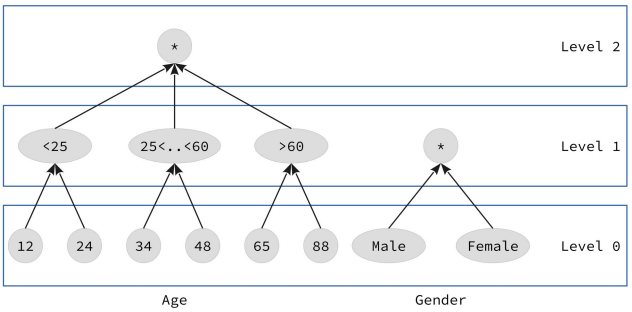


FIGURE 4. Taxonomy tree.

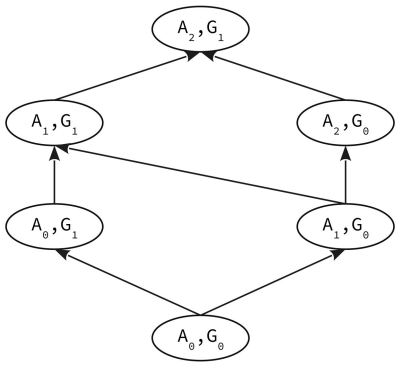


FIGURE 5. Domain generalisation hierarchy of attributes age and gender.

in Figure 4 for the attributes Age and Gender. This can be performed through the *Domain Generalisation Hierarchy* (DGH), a lattice or a graph which is the solution space for the anonymisation problem (cf. Figure 5). Each node of the lattice represents a different combination of generalisation levels of the attributes. Nevertheless, in quantitative data the DGH can be defined dynamically as in [124], [125].

In [126] values are replaced following a predefined generalisation hierarchy in which original values are substituted with a more generic representation of their domain. Generalisation transformations can be classified into two main categories:

1) **Global recoding** refers to the replacement of a value with a more generic one, e.g. if a value a_1 is replaced by A, then all the occurrences of a_1 in T will be replaced by A. Three classes of global recording can be further identified:
   - **Full domain generalisation:** In this case, all values of an attribute are generalised to the same level of the generalisation hierarchy [5], [109], [110], [127]–[130]. The main advantage of this approach is that the anonymous dataset T^* stores all its data with the same granularity level, enhancing its readability. However, substantial information may be lost due to excessive generalisation.
   - **Subtree generalisation:** This method requires all sibling nodes to be generalised to the same value. Nodes of other subtrees can be generalised independently if deemed necessary [116], [117], [131]–[133].
   - **Sibling generalisation:** In this case, some of the sibling's nodes are generalised, while others remain intact. Sibling generalisation further reduces information loss since it grants higher flexibility to the anonymisation algorithm. Nevertheless, it implies that the solution's search space is increased [127].
2) **Local recoding**: This method enables more flexibility than global recoding as it allows only specific appearances of a value to be generalised in the anonymous dataset. In other words, $\mathcal{QI}s$ values can be generalised to different levels to form groups of records with the same $\mathcal{QI}s$ values. For example, the value 13500 of the attribute Zip Code in one group can be generalised to 125**, while in other groups it can have any value of the hierarchy, e.g. 12*** or 1255*.
   - **Cell generalisation:** Contrary to the previous techniques of global recoding, cell generalisation allows the generalisation of an instance of a value

while leaving the rest instances unmodified. If we assume that in a specific record the value a_1 needs to be generalised to A, then every other occurrence of the value a_1 can either remain unchanged or be modified to the same or another generalisation level, depending on the anonymisation procedure [112].

- **Multidimensional generalisation:** A multidimensional generalisation can be obtained by applying a single function to the relation that generalises a $QI = (v_1, \ldots, v_n)$ to $QI^* = (u_1, \ldots, u_n)$ such that $v_i = u_i$ or v_i is a descendant node of u_i in the taxonomy of attribute i [134], [135]. Therefore, contrary to cell generalisation, multidimensional generalisation considers the multiple dimensions of the tuples.

When selecting a generalisation operator, two factors have to be considered: a) the quality of the generalisation; the more flexible an operator is, the lower the information loss will be, and b) the computational cost of the algorithm. More flexible operators offer a greater solution space but the search for a solution might be significantly harder. The importance of selecting the right DGH has been studied in [126], [136]

B. SUPPRESSION

Suppression is the erasure of specific values from the original dataset. Different levels of suppression can be considered. For example, in *Tuple or Record Suppression* [128], [131], the entire record is suppressed, while *Value Suppression* [137] entails the suppression of a given value throughout the entire table. Finally, *Cell suppression* [111] deletes only a subset of instances in the table.

C. BUCKETISATION

Another transformation for data anonymisation is *bucketisation*. The association between $\mathcal{QI}s$ and the *SA* breaks simply by publishing them in separate tables. A common attribute, the $group - id$, in the published tables lets the data recipient to form groups from the *SA* table where any *SA* value with $group - id = i$ can be linked to any individual with $group - id = i$ at the $\mathcal{QI}s$ table. Bucketisation succeeds to break the connection between $\mathcal{QI}s$ and the *SA* without modifying them.

Bucketisation, often referred to as Anatomisation [138] as well, simply de-associates the relationship between the $\mathcal{QI}$ and the *SA*, without modifying them. This approach releases the $\mathcal{QI}$s and the *SA*s in separate tables conserving only a common attribute, the ID of the group. Therefore, the records with the same group ID in the $\mathcal{QI}$ table are linked to the corresponding values in the *SA*. Compared to the generalisation method, the anatomised tables grant a more accurate answer to aggregation queries that involve $\mathcal{QI}$s since the values remain intact. Therefore, bucketisation enables better preservation of the original terms, compared with other methods.

One of the main drawbacks of bucketisation methods relies on the fact that an adversary could infer more easily whether his target participates or not in a released dataset. This privacy breach is specially relevant when the participation in a table is considered sensitive. For instance, an attacker could infer whether his victim took a test for a sexually transmitted disease, without knowing though the result of the test.

1) SLICING

The basic concept of slicing is to disassociate the cross-column relations while preserving the association within the context of each column. Anatomisation can be seen as a special case of Slicing [139], where there are only two columns, one containing all the $\mathcal{QI}$s and another containing only the *SA*. In slicing, columns can be formed with one or more $\mathcal{QI}$s, *SA* or both. Grouping highly correlated attributes preserves the utility, whereas breaking the associations between uncorrelated attributes increases privacy protection.

2) DISASSOCIATION

Terrovitis *et al.* proposed an anonymisation transformation termed disassociation [140]. This technique focuses on identity disclosure protection in sparse multidimensional data. The main advantage of such method is that it maintains the original terms without suppressing or generalising. The applied transformation partitions the original records into smaller and disassociated subrecords. The main aim is to disguise infrequent term combinations in the original records by scattering terms in disassociated subrecords.

3) LOOSE ASSOCIATIONS

Loose associations [141] is a similar idea but provides a more flexible solution than Anatomisation to cater for privacy without using generalisation. The goal of Loose associations is to protect sensitive associations among the attributes in a dataset. The data publisher can define a set of sensitive associations among selected attributes from the original table and then break these associations by publishing the attributes in different fragments. These sensitive associations are modelled through confidentiality constraints. An extension of the method was proposed in [142], [143] to handle more than a pair of fragments.

D. PERMUTATION

Zhang *et al.* [144] proposed the permutation method based on the concept of anatomisation. Permutation disassociates a $\mathcal{QI}$ with a numerical *SA* by separating the data records into groups and then shuffling *SA* values inside each group. Therefore, permutation enables accurate answering of aggregate queries compared with other methods such as generalisation-based approaches. Nevertheless, while data permutation seems an efficient method, it has several drawbacks. For instance, if logical links exist between the different attributes, the random permutation of the *SA* values may result in poor/low privacy guarantees.

E. PERTURBATION

Perturbation modifies/disguises records in a way that they do not correspond to the original values anymore. This method keeps the distortion of statistical information values to acceptable levels, while a synthetic dataset replaces the original one. Therefore, record attribute linkage attacks are not useful in this scenario. However, this does not imply that they are immune to other attacks [145]. Perturbation methods can be categorised as follows:

- **Noise addition** In the case of noise addition, we may perturb the numerical values of a dataset using a Gaussian distribution with zero mean and standard deviation σ (*i.e.* $\mathcal{N}(0, \sigma)$). The higher the σ value, the greater the range of the generated values (*i.e.* it is more likely to generate values close to the boundaries of the value range). We may also use a discrete uniform distribution $\mathcal{U}(S)$, where S is the set of actual values present in the category which is being evaluated (*i.e.* we may substitute rare values or values with too few observations by other real values present in the dataset, such as age or weight). Laplace distributions are also widely used because of their interesting properties [146]. A simple way to hide a number a is to add a random number r to it. Although that we cannot retrieve the original value of a, as it is disguised, we can perform certain computations if we are interested in the aggregated data rather than in individual data [147]. The main idea of random noise addition is to perturb/obfuscate the data so that certain computations can be performed while preserving users' privacy. Despite that each individual's information is disguised, if the number of participants is significantly large, the aggregate information of such participants can be estimated with decent accuracy. For instance, the scalar product and random sum are widely used methods that can benefit from the aforementioned property [148]. Therefore, we can estimate the required information by using disguised data, and thereby we obtain meaningful outcome without knowing the exact values of individual data items.

  Random Sum Let O be the original vector with n values, where $O = (o_1, o_2, \ldots, o_n)$. O is disguised by $R = (r_1, r_2, \ldots, r_n)$, where r_i's are values generated by a Gaussian distribution with 0 mean and standard deviation σ. Let $O' = O + R$ be the disguised data that is known. Since r_i's are uniformly distributed in domain $[-\sigma, \sigma]$, their contribution to the actual sum of the values of vector O is close to zero. Thus, in the long run, the relative error will converge to zero. Hence, we have:

$$\sum_{i=1}^{n}(o_i + r_i) = \sum_{i=1}^{n} o_i + \sum_{i=1}^{n} r_i \approx \sum_{i=1}^{n} o_i \tag{1}$$

  Scalar Product Let A and B be the original vectors, where $A = (a_1, a_2, \ldots, a_n)$ and $B = (b_1, b_2, \ldots, b_n)$. Let R and V be two vectors with values generated by a Gaussian distribution with 0 mean and standard deviation σ, so that $R = (r_1, r_2, \ldots, r_n)$, $V = (v_1, v_2, \ldots, v_n)$, and the sum of the values of R and V are equal to 0. A is disguised by R and B is disguised by V. Let $A' = A + R$ and $B' = B + V$ be the disguised data that are known. The scalar product of A and B can be estimated from A' and B' as follows:

$$A' \cdot B' = \sum_{i=1}^{n}(a_i b_i + a_i v_i + r_i b_i + r_i v_i) \tag{2}$$

  Because R and B are independent, we have:

$$\sum_{i=1}^{n} r_i b_i \approx 0 \text{ and similarly} \sum_{i=1}^{n} a_i v_i \approx 0 \text{ and} \sum_{i=1}^{n} r_i v_i \approx 0 \tag{3}$$

  Therefore, we have:

$$\begin{aligned}\sum_{i=1}^{n}(a_i + r_i)(b_i + v_i) &= \sum_{i=1}^{n}(a_i b_i + a_i v_i + r_i b_i + r_i v_i) \\ &\approx \sum_{i=1}^{n} a_i b_i \end{aligned} \tag{4}$$

- **Data swapping** is not constrained by the type of *SA* (i.e. it can be used for both numerical and categorical values). This model anonymises the original table by exchanging the *SA* values among the records [149]–[153]. In several occasions, *SA* values may have interdependencies which, when broken, might undermine the utility of the data. A typical example can be the case of gender-specific diseases in medical data. For instance, assigning *Prostate cancer* to a woman or *Mastitis* to a man are impossible real pairs; nonetheless, they may be a result of random swaps.
- **Synthetic data generation** builds a mathematical model based on the original data so that basic statistical measures or relationships are preserved. Therefore, it uses the mathematical model to generate the anonymised table with synthetic records [154]–[159]. The main drawback of synthetic data is that they are no longer useful for analysis on random subdomains. To overcome this issue, two approaches were proposed, namely the Partially synthetic approach [160], and the Hybrid data approach [161]–[163].
- **Microaggregation** is a perturbation method which consists on the aggregation of the attribute's values to reduce re-identification risk. This method is implemented in two different phases, *data partitioning* and *partition aggregation* [164]. The first phase partitions the dataset T in subset $T_{s_1}, T_{s_2}, \ldots T_{s_n}$ in such way that for $i \neq j$, $T_{s_i} \cap T_{s_j} = \emptyset$ and $T_{s_1} \cup T_{s_2} \cup \ldots \cup T_{s_n} = T$. In the second phase, a representative value for each cluster is selected (e.g. the median or the mean value are widely used as representative values) to replace the original values. A cardinality parameter k controls the minimum size of clusters.

TABLE 4. Microaggregation example.

(a) Original Data.

Hours paid for	Wage rate	Wage sum	Total hours
7	32	224	11
23	25	575	35
30	50	1500	41
21	60	1260	38
8	40	320	13
11	60	660	15
26	75	1950	32
65	30	1950	72
5	32	160	9
11	75	825	19
51	60	3060	57
9	50	450	12

(b) Microaggregation.

Hours paid for	Wage rate	Wage sum	Total hours
6.667	34.667	234.667	11
23.333	53.333	1261.667	35
48.667	46.667	2170	56.667
23.333	53.333	1261.667	35
6.667	34.667	234.667	11
10.333	61.667	645	15.333
23.333	53.333	1261.667	35
48.667	46.667	2170	56.667
6.667	34.667	234.667	11
10.333	61.667	645	15.333
48.667	46.667	2170	56.667
10.333	61.667	645	15.333

TABLE 5. Data transformation: data privacy vs utility.

Data Transformation	Data Privacy	Utility
Bucketisation	Low	High
Generalisation	Average	Average
Permutation	Average	Average
Perturbation	High	Low
Suppression	High	Low

Although microaggregation was initially designed for numerical attributes, it was later extended to cover categorical attributes [165]. In the example of microaggregation in Table 4, the multiplication of the value of the attribute "Hours paid for" with "Wage Rate" should result in the "Wage Sum". While this relation is valid in the original data, in the microaggregated version, such relation is violated. In such cases, the data publisher should use certain constraints [166], [167] which the microaggregation algorithm should not violate. Moreover, variations of microaggregation can use variable group sizes to increase utility and decrease information loss [168], [169].

A very important principle when applying anonymisation is the so-called *Minimality principle*.

Definition 1: Minimality principle Assume that an algorithm A is used to produce an anonymous table T^* that satisfies the requirements R. For any EC in T^* there are no specialization of any QI that can result another table $T^{\#}$ which satisfies the requirements R

In essence, the minimality principle mandates that an anonymisation algorithm should not generalise, suppress, or distort the original data on T more than it is necessary to achieve, e.g. k-anonymity.

Beyond any doubt, the relation between privacy and utility of an anonymised table is crucial [108], [170], [171]. In that respect, data publishers have to carefully balance between an anonymised table that has no practical use and a useful table of microdata that has not any strong privacy guarantees. In this section, we have described the most widely-used data transformation techniques that are used to satisfy the data holders' requirements. Moreover, we have shown with practical examples in which cases each of them could be efficiently used to provide for privacy. With this in mind, Table 5 summarises in a three-scale level the impact of each data transformation technique on privacy and utility.

V. DE-ANONYMISATION ATTACKS

In the following paragraphs, we discuss the main methods an adversary would use to attack anonymised datasets and infer sensitive information about individuals.

Data publication opens the door to a wide range of possible attacks. Factors related to adversaries' capacities in terms of processing power, technical knowledge, as well as goals and motives for accomplishing their attacks (e.g. identifying complete records of individuals or determining whether one is included in as published dataset) affect greatly the variety of possible attacks. We can classify them into the following categories:

- Record Linkage: In the *record linkage* attack the adversary, by using his background knowledge of the $\mathcal{QI}$s, tries to link one or more records of the anonymised dataset to an individual
- Attribute Linkage: In this attack an adversary attempts to link a specific attribute value to an individual. Even if the data are anonymised and there exist several occurrences of the same $\mathcal{QI}$ to prevent a record linkage attack, the adversary might still be able to associate a specific individual with an SA value. This attack can be effectively executed if the diversity of SAs in each group of records sharing the same $\mathcal{QI}$ is insufficient. Therefore, in the case of groups formatted by means of $\mathcal{QI}$s, the adversary could still infer the sensitive value of an individual.

  The Attribute Linkage attack can be manifested in various ways. In [172]the authors describe the *Homogeneity Attack* in which a unique or a very common SA value among individuals with the same $\mathcal{QI}$s can lead the adversary to infer the victim's SA. To achieve his goal the adversary can also use his background knowledge about his target. For example, let us assume that Alice has a Korean friend Uneko and an SA attribute about Uneko's health status could take only two values, either

Viral Infection or Stroke. Since it is well-known that Koreans have an extremely low incidence of strokes, Alice can conclude with high confidence that Uneko has an infection. This is an example of *Background Knowledge Attack* [172].

Probabilistic attacks are special cases of attribute linkage attack and they focus on how the attacker's belief about the *SA* value of an individual would be modified after accessing the anonymised table T^*. The two main types are the *Skewness Attack* which exploits a skewed distribution of an *SA* and the *Similarity Attack* which take into account the semantic similarity of the *SA* values. More details on Skewness and Similarity attack will be given in VI-B9.

- Table Linkage Attack: There might be cases in which only the presence of an individual could lead to privacy breaches. Such an example could be the presence of a specific person in a published cancer dataset. This kind of attacks that reveal whether someone's record exists or not in an anonymised table are called *table linkage* attacks [173], [174].
- Attack on Continuous Data Publications: The adversary can use previously released anonymous publications of the same dataset to perform any of the aforementioned linkage attacks.
- Algorithm exploitation: In this attack the adversary is aware of the algorithm that has been used in the anonymisation process and therefore he can use reverse engineering to reason about the decision the algorithm made and extract a lot of knowledge about anonymised records.

VI. PRIVACY MODELS AND COUNTERMEASURES

In what follows, we illustrate the methods a data publisher would apply to protect the published data against the attacks discussed in the previous section, and what privacy guarantees each method provides.

A. COUNTERMEASURES TO RECORD LINKAGE

In this section, we present the most well-known countermeasures against record linkage. Of specific interest is k-anonymity, one of the first and most widely-used methods in the field. Moreover, we present some of its variations and discuss their drawbacks and limitations in specific attack scenarios.

1) k-ANONYMITY

Samarati and Sweeney [175] and Sweeney [5] presented the notion of k-anonymity as a countermeasure to record linkage. A dataset is k-anonymous if it includes k records for any set of $\mathcal{QI}$ values. Therefore, a record must be indistinguishable from at least $k-1$ other records with respect to $\mathcal{QI}$s. The group of records sharing the same $\mathcal{QI}$ form an equivalence class (EC). From an attacker's perspective, the probability to successfully link his target record is never greater than $\frac{1}{k}$. This is the probability that the adversary knows a specific individual is present in the dataset as well as the possible values of $\mathcal{QI}$ of the target.

TABLE 6. Example of 4-anonymity.

(a) T_1

Non-Sensitive			Sensitive
Zip Code	Age	Nationality	Condition
13053	28	Russian	Heart Disease
13068	29	American	Heart Disease
13068	21	Japanese	Viral Infection
13053	23	American	Viral Infection
14853	50	Indian	Cancer
14853	55	Russian	Heart Disease
14850	47	American	Viral Infection
14840	49	American	Viral Infection
13053	31	American	Cancer
13053	37	Indian	Cancer
13068	36	Japanese	Cancer
13068	35	American	Cancer

(b) T_1^* 4-anonymous

Non-Sensitive			Sensitive
Zip Code	Age	Nationality	Condition
130**	< 30	*	Heart Disease
130**	< 30	*	Heart Disease
130**	< 30	*	Viral Infection
130**	< 30	*	Viral Infection
148**	> 40	*	Cancer
148**	> 40	*	Heart Disease
148**	> 40	*	Viral Infection
148**	> 40	*	Viral Infection
130**	30-40	*	Cancer
130**	30-40	*	Cancer
130**	30-40	*	Cancer
130**	30-40	*	Cancer

A formal definition of k-anonymity, as given by Machanavajjhala *et al.* in [172], is the following:

Definition 2 (k-anonymity): A table T is k-anonymous if for every record (tuple) $t \in T$ there exist $k-1$ other records $t_{i_1}, t_{i_2}, \ldots, t_{i_{k-1}} \in T$ such that $t[\mathcal{C}] = t_{i_1}[\mathcal{C}] = t_{i_2}[\mathcal{C}] = \ldots = t_{i_{k-1}}[\mathcal{C}],\ \forall \mathcal{C} \in \mathcal{QI}$

Example 1: The original table T_1 in Table 6a, is transformed into T_1^* (see Table 6b) with the generalisation of the $\mathcal{QI}$ Age and Zip Code, and with generalisation to the maximum level of Nationality, which is equivalent to the suppression of this $\mathcal{QI}$ attribute. Obviously, the anonymised table is 4-anonymous, since for every record there exist at least three others with the same $\mathcal{QI}$ values.

The *Curse of dimensionality* plays a crucial role in anonymisation, as observed by Aggarwal [176]. He showed that if the set of $\mathcal{QI}s$ becomes large enough, the k-anonymity property can only be guaranteed if most of the records are deleted, as seen in Table 6. In this scenario, it is clear enough that the higher the dimensionality of data, the greater the information loss. To overcome this problem, practitioners often anonymise data by using only a subset of the $\mathcal{QI}s$ according to their purpose. Moreover, they release data in parallel tables with different subsets of $\mathcal{QI}$s.

a: SELECTING k IN k-ANONYMITY

The parameter k in k-anonymity is selected according to each data publisher. Nevertheless, Dewri *et al.* [177] argue that k

must be selected in a more informative and objective manner when suppression is not permitted. In that regard, they propose a multi-objective optimisation problem to analyse the trade-off between different k values and the information gain.

b: k-ANONYMITY THOUGH MICROAGGREGATION

The use of microaggregation to satisfy k-anonymisation was first studied in [178]–[180]. When numeric attributes are microaggregated independently - a case called *univariate microaggregation* - k-anonymity cannot be guaranteed since each set of $\mathcal{QI}s$ may have less than $k-1$ instances. On the contrary, *multivariate microaggregation* considers the full set of $\mathcal{QI}s$ and microaggregates these attributes together so the k-anonymity is guaranteed. Nevertheless, *multivariate microaggregation* suffers from larger information loss than *univariate microaggregation*.

c: k-MAP

Prior to k-anonymity, Sweeney [5] introduced the k-map property. Let us assume that T_{id} is an identification database which is k-anonymised to produce T^*_{id}. The k-map property states that each record in the disclosed k-map dataset T^* can be linked to at least k records in T^*_{id} (and not in the original dataset T as in k-anonymity).

This way the data publisher guarantees that the re-identification risk is the same as k-anonymity and simultaneously the information loss is reduced. The main drawback of this approach is that the combination of external data with those the data holder wants to release is rarely feasible.

d: (X,Y)-ANONYMITY

As previously stated, k-anonymity assumes that each record holder has only one record in the dataset. Nevertheless, as can be seen in the medical context this is not always the case. For example, let us assume a dataset with a set of $\mathcal{QI}$s, namely Age, Gender and Zip Code, an *SA* Disease and the Social Security Number (SSN) as *EI*. Since a record holder may suffer from more than one diseases, there may be more than one records representing the same individual in the dataset. Therefore, even when removing the *EI* SSN, each *EC* may not contain k distinct individuals. In the extreme case in which a record holder has k records, an *EC* could contain only records from the same individual. In [71] Wang and Fung presented the notion of *(X,Y)-Anonymity*. *(X,Y)-Anonymity* requires that each value on X must be linked to at least k distinct values on Y. In our example, $X = \{\text{Age, Gender, Zip Code}\}$ and $Y = \{\text{SSN}\}$. Note that Y may also be set to the *SA* disease so that each group is associated with a diverse set of *SA* values, enhancing the protection of the *SA* value.

2) (1,k)-ANONYMISATION, (k,1)-ANONYMISATION, (k,k)-ANONYMISATION

In [181], Gionis *et al.* proposed a relaxation of k-anonymity by introducing the notions of $(1,k)$-anonymity and $(k,1)$-anonymity.

$(1,k)$-anonymity In the case that adversaries only possess knowledge of public datasets T_{pub}, the generalisation of the table entries in such way that the public data T_{pub} of every individual are consistent with at least k records of the released table T^* may be sufficient. It is worth noting that every k-anonymous table is also a $(1,k)$-anonymised table but the opposite is not necessarily true.

$(k,1)$-anonymity A table satisfies $(k,1)$-anonymity if all records in the released table are consistent with at least k records on the original table T. As in the previous case, a k-anonymous table is also $(k,1)$-anonymous.

Obviously, since both these methods are relaxed adoptions of k-anonymity, a combination of the two could be used to increase the privacy guarantees.

(k,k)-anonymity If an anonymous table satisfies both $(k,1)$-anonymity and $(1,k)$-anonymity, then this table also guarantees (k,k)-anonymity. In this case, the privacy protection level is similar to k-anonymity when the attack scenario considers an adversary who has knowledge of a subset of the individuals in the table. By employing (k,k)-anonymity a data publisher may offer more data utility than when using k-anonymity.

3) NON-HOMOGENEOUS GENERALISATION

To enhance the baseline approach that considers the same generalised values for each $\mathcal{QI}$ within an *EC*, some researchers [182]–[184] have explored the idea of further reducing the information loss in *ECs* with more than k members by using non-homogeneous generalisation. In this regard, tuples within a partition can now take different generalised $\mathcal{QI}s$ values inside the *EC*.

Example 2: The original data are shown in Table 7a while Table 7b is the 2-anonymous table with homogeneous generalisation applied. Now consider the possible publication of Table 7a, as shown in Table 7c. The first 3 records, out of 5, have different generalised $\mathcal{QI}s$. In this example, we assume that the adversary has knowledge of all the $\mathcal{QI}s$ of all record holders in Table 7a. In Tables 7b and 7c the adversary has a 50% chance to perform a successful record linkage attack since both of them are 2-anonymous. Someone can easily observe that for each record and $\mathcal{QI}$ attribute of Table 7c, the generalised range is either smaller or equal to the corresponding range for that record and $\mathcal{QI}$ attribute in Table 7b. The latter improves utility by means of non-homogenous generalisation regardless of the information metric used.

a: k-CONCEALMENT

Based on (k,k)-anonymisation, Tassa *et al.* [183] proposed the notion of *k-concealment* to achieve anonymity. In contrast to k-anonymity where *EC*s with identical $\mathcal{QI}$ are required, in k-concealment the generalisation is made so that each record becomes computationally-indistinguishable from $k-1$ others.

Example 3: Consider the Table 8a with $\mathcal{QI}s$ Age and Zip Code and *SA* Disease. Table 8b corresponds to 2-anonymised version of Table 8a, where there are two *EC* with two and

TABLE 7. Homogeneous vs non-homogeneous generalisation.

(a) Original Table

Age	Gender	Zip code	Disease
30	M	10152	Viral Infection
28	F	10157	Diabetes
15	M	10118	Cancer
48	M	10500	Heart Disease
20	M	10511	Flu

(b) 2-anonymous using homogeneous generalisation

Age	Gender	Zip code	Disease
15 - 30	*	10***	Viral Infection
15 - 30	*	10***	Diabetes
15 - 30	*	10***	Cancer
20 - 48	M	105**	Heart Disease
20 - 48	M	105**	Flu

(c) 2-anonymous using non-homogeneous generalisation

Age	Gender	Zip code	Disease
28 - 30	*	1015*	Viral Infection
15 - 28	*	10***	Diabetes
15 - 30	M	10***	Cancer
20 - 48	M	105**	Heart Disease
20 - 48	M	105**	Flu

three records. Table 8c corresponds to 2-concealment version of Table 8a, where an adversary who knows the $\mathcal{QI}s$ of all records cannot link a specific record to less than two records. Assuming that the adversary knows the $\mathcal{QI}s$ of Alice (Age and Zip Code), he cannot tell which one of the two records (the first or the third one) belongs to Alice. Although the first record is more likely to belong to Alice, the authors [183] showed that proving that is computationally expensive.

b: n-CONFUSION

***n*-Confusion** [184] is another relaxation of k-anonymity that resembles k-concealment. The main idea behind it is to modify the records so that they become indistinguishable with respect to the re-identification process. The re-identification process is a function that given a collection of entries in the anonymised table and some auxiliary additional information, returns the probability that there are entries from the original table.

4) MultiRelational k-ANONYMITY

The vast majority of the algorithms assume that each individual's record corresponds to one row in a table. Nevertheless, information about an individual can be disseminated across multiple tables in a database scheme. In this regard, k-anonymity offers protection at a record level but not at an owner's record level. Negriz *et al.* [185] showed that algorithms designed for a single table were insufficient, even when the database tables were transformed into a single

TABLE 8. 2-anonymous vs 2-concealment [183].

(a) Original Table [183]

Name	Age	Zip Code	Disease
Alice	30	10055	Measles
Bob	21	10055	Flu
Carol	21	10023	Angina
David	55	10165	Flu
Eve	47	10224	Diabetes

(b) 2-anonymous

Age	Zip Code	Disease
21-30	100**	Measles
21-30	100**	Flu
21-30	100**	Angina
47-55	10***	Flu
47-55	10***	Diabetes

(c) 2-concealment

Age	Zip Code	Disease
21-30	**10055**	Measles
21	100**	Flu
21-30	100**	Angina
47-55	100**	Flu
47-55	100**	Diabetes

table. Hence, Nergiz *et al.* proposed [185] the *Multirelational k-Anonymity*, which assumes that a database contains a person-specific table with a unique key, and such table is linked with other tables which have a foreign key, some $\mathcal{QI}s$ and *SAs*. In this scenario, MultiRelational k-Anonymity is satisfied if, after we join any person-specific table with all the others, there exist at least $k-1$ record owners having the same $\mathcal{QI}s$.

5) k^m-ANONYMITY

The notion of k^m-anonymity was proposed by Terrovitis *et al.* [186] for transaction databases. Formally, it is defined as:

Definition 3 (k^m-anonymity): A table T is k^m-anonymous if any adversary with background knowledge of up to m items of a transaction $t \in T$, cannot use these items to re-identify less than k records from T.

Hence, any subset query of size m or less should return either zero results or more than k records. Note that queries not returning an answer are also considered secure since they indicate that background information cannot be associated with any transaction. k^m-anonymity relaxes the guarantee of k-anonymity and does not take into account the distinction between $\mathcal{QI}s$ and *SA* since, in the case of transactional data, there are no $\mathcal{QI}s$ and all items are considered sensitive.

Example 4: An adversary may know that Alice has purchased milk, beer and diapers from a store. This background knowledge is easy to acquire by just observing the top of the shopping bags of Alice or by just looking at a picture on a social network where these items are depicted

in Alice's house. The adversary now can examine the released transactional data to find the records with these 3 items (beer, milk and diapers). By exploiting this knowledge over the released data the adversary can limit the results to a reduced set of transactions or even uniquely associate a transaction with an individual. Since the transaction may include sensitive items, if the adversary knows its content, Alice's privacy may be compromised. If the dataset were anonymised using 5^3-anonymity, then for such 3 items there would exist at least 4 other transactions containing beer, milk and diapers, so that the adversary could not distinguish which one is Alice's.

B. COUNTERMEASURES TO ATTRIBUTE LINKAGE AND TABLE LINKAGE

To protect data from Attribute Linkage & Table Linkage attacks, several privacy notions have been introduced in the literature. The most well-known are ℓ-diversity and t-closeness, which are analysed in the next paragraphs. Nonetheless, we extend the discussion to other approaches as well and we provide specific scenarios in which they can be applied.

1) CONFIDENCE BOUNDING

In [137], [187] Wang *et al.* introduced the idea of bounding the adversary's confidence of inferring a *SA* value from a set of $\mathcal{QI}s$ by specifying privacy templates. Such templates define which *SA* value to protect with a threshold h of a given set of $\mathcal{QI}s$. A table satisfies the privacy template if for the given $\mathcal{QI}$s the confidence of inferring the *SA* value is lower than h. A key point of confidence bounding is that the data publisher can set different templates to protect different values of *SA* rather than a unique policy protection for the dataset.

2) p-SENSITIVE k-ANONYMITY

The (p)-sensitive k-anonymity has been proposed by Truta and Vinay [188]:

Definition 4 ((p)-sensitive k-anonymity): An anonymised table T^* satisfies (p)-sensitive k-anonymity property if it satisfies k-anonymity, and for each *EC* in T^*, the number of distinct values for each *SA* is at least p within the same *EC*. Table 9b shows an example of (p)-sensitive k-anonymity.

Someone can easily observe that while in the first *EC* (*ID*s from 1 to 4) the *SA* values are different, an adversary can still conclude that his target suffers from a severe and incurable disease.

The authors in [189] proposed two extensions of (p)-sensitive k-anonymity.

a: (p^+)-SENSITIVE k-ANONYMITY

Definition 5 ((p^+)-sensitive k-anonymity): An anonymised table T^* satisfies (p^+)-sensitive k-anonymity property if it satisfies k-anonymity, and for each *EC* in T^*, the number of distinct categories for each *SA* is at least p within the same *EC*.

TABLE 9. (p)-sensitive k-anonymity.

(a) Original Dataset.

ID	Age	Country	Zip Code	Disease
1	37	Brazil	24248	HIV
2	38	Mexico	24207	HIV
3	36	Brazil	24206	Cancer
4	35	Mexico	24249	Cancer
5	51	Italy	23053	Diabetes
6	58	Spain	23074	Pneumonia
7	55	Germany	23064	Bronchitis
8	52	Germany	23062	Gastritis
9	43	Brazil	24248	Zika fever
10	47	Mexico	24204	Zika fever
11	46	Mexico	24205	Zika fever
12	45	Brazil	24248	Colitis

(b) (p)-sensitive k-anonymity.

ID	Age	Country	Zip Code	Disease
1	<40	America	242**	HIV
2	<40	America	242**	HIV
3	<40	America	242**	Cancer
4	<40	America	242**	Cancer
5	>50	Europe	230**	Diabetes
6	>50	Europe	230**	Pneumonia
7	>50	Europe	230**	Bronchitis
8	>50	Europe	230**	Gastritis
9	4*	America	242**	Zika fever
10	4*	America	242**	Zika fever
11	4*	America	242**	Zika fever
12	4*	America	242**	Colitis

TABLE 10. Grouping *SA* values.

CategoryID	Sensitive Values	Sensitivity
One	Cancer, HIV	Very High
Two	Pneumonia, Diabetes	High
Three	Bronchitis,Gastritis	Medium
Four	Colitis, Zika fever	Low

Table 10 shows an example of how *SA* values can be grouped to achieve the preferred (p^+)-sensitive k-anonymity.

b: (p, α)-SENSITIVE k-ANONYMITY

Definition 6 ((p, α)-sensitive k-anonymity): An anonymised table T^* satisfies (p, α)-sensitive k-anonymity if it satisfies k-anonymity, and each *EC* has at least p distinct *SA* values with its total weight being at least α. An example of this property is depicted in Table 11.

3) ℓ-DIVERSITY

One of the first attempts to counter Attribute Linkage was made by Machanavajjhala *et al.* [172]. To illustrate their approach we provide the following example.

Example 5: Even if anonymisation of $\mathcal{QI}s$ has been applied in table T_1 (see Table 6a), and a 4-anonymous table T_1^* has been produced, someone can easily observe that there is a privacy leakage. Let us assume that we have an adversary, Malory, who has some background knowledge about the

TABLE 11. (*p*)-sensitive *k*-anonymity.

(a) (2^+)-sensitive 4-anonymity.

Age	Country	Zip	Disease	Category
<50	America	2424*	HIV	One
<50	America	2424*	Cancer	One
<50	America	2424*	Zika fever	Four
<50	America	2424*	Colitis	Four
>50	Europe	230**	Diabetes	Two
>50	Europe	230**	Pneumonia	Two
>50	Europe	230**	Bronchitis	Three
>50	Europe	230**	Gastritis	Three
<50	America	2420*	HIV	One
<50	America	2420*	Cancer	One
<50	America	2420*	Zika fever	Four
<50	America	2420*	Zika fever	Four

(b) $(3, 1)$-sensitive 4-anonymity.

Age	Country	Zip	Disease	Weight	Total
<40	America	242**	HIV	0	1
<40	America	242**	HIV	0	
<40	America	242**	Cancer	0	
<40	America	242**	Zika fever	1	
>40	Europe	230**	Diabetes	1/3	2
>40	Europe	230**	Pneumonia	1/3	
>40	Europe	230**	Bronchitis	2/3	
>40	Europe	230**	Gastritis	2/3	
<40	America	24***	Cancer	0	3
<40	America	24***	Zika fever	1	
<40	America	24***	Zika fever	1	
<40	America	24***	Colitis	1	

TABLE 12. 4-anonymous 3-diverse.

	Non-Sensitive			Sensitive
	Zip Code	Age	Nationality	Condition
1	1305*	≤ 40	*	Heart Disease
4	1305*	≤ 40	*	Viral Infection
9	1305*	≤ 40	*	Cancer
10	1305*	≤ 40	*	Cancer
5	1485*	> 40	*	Cancer
6	1485*	> 40	*	Heart Disease
7	1485*	> 40	*	Viral Infection
8	1485*	> 40	*	Viral Infection
2	1306*	≤ 40	*	Heart Disease
3	1306*	≤ 40	*	Viral Infection
11	1306*	≤ 40	*	Cancer
12	1306*	≤ 40	*	Cancer

victim Bob. In this case, we assume that Malory is Bob's neighbour that she knows that Bob is a 31 year old American and, since they are neighbours, she also knows Bob's zip code. Malory knows that Bob recently visited the hospital and she is curious to find out why. From Table 6b Malory can infer, even if she cannot point which is Bob's record, that Bob has cancer.

To counter such attacks, the ℓ-diversity model requires that each *EC* contains at least ℓ different attribute values. A more formal definition of ℓ-diversity given by Machanavajjhala *et al.* [172] is the following:

Definition 7 (ℓ-diversity): An *EC* is ℓ-diverse if there are at least ℓ "well-represented" values for the sensitive attribute. A table T is ℓ-diverse if every $EC \in T$ is ℓ-diverse.

The term "well-represented" denotes that there are at least ℓ distinct values for the *SA* in each *EC*, which is identical to (*p*)-sensitive *k*-anonymity property.

When a table T has more than one *SA* then the use of Multi-Attribute ℓ-diversity provides the required privacy (see Table 12):

Definition 8 (Multi-Attribute ℓ-diversity): Let T a table with quasi-identifiers $Q_1, Q_2, \ldots, Q_{m_1}$ and sensitive attributes $S_1, S_2, \ldots, S_{m_2}$. T is called ℓ-diverse if for all $i = 1 \ldots m_2$, the table is ℓ-diverse when S_i is treated as the sole *SA* and $\{Q_1, Q_2, \ldots, Q_{m_1}, S_1, \ldots, S_{i-1}, S_{i+1}, \ldots, S_{m_2}\}$ is treated as the $\mathcal{QI}$.

Nonetheless, an adversary can have a lot of background information about specific individuals due to, e.g. acquaintance or inference [190]. Martin *et al.* [191] provided a formal language to express the background knowledge of the adversary into individual units. Therefore, prior to publication, one could quantify the disclosure risk of the anonymized dataset for different background knowledge scenarios. Moreover, the authors provided a method to generate an anonymized table whose maximum disclosure is well bounded.

4) ℓ^+-DIVERSITY

The ℓ^+-diversity notion was presented by Liu and Wang [192]. Instead of granting a global protection for all *SA* values, it sets a different privacy threshold for each *SA* value to decrease the distortion of the original data while offering user defined value-based privacy protection.

5) (X,Y) - PRIVACY

The main weakness of *(X,Y) - Anonymity* is that when a value on Y occurs more often than others, then the probability of inferring the *SA* value can be greater than $\frac{1}{k}$. To overcome this issue, Wang and Fung introduced the notion of *(X, Y)-Privacy* [71] which extends *(X,Y) - Anonymity* by inserting the constraints of confidence bounding. To satisfy *(X, Y)-Privacy* each group x on X has to contain at least k records and for each *SA* value s on Y the confidence to infer s from x is less than h, where h is the value for the confidence bounding.

6) (α, k)-ANONYMITY

Wong *et al.* [193] proposed the notion of (α, k)-anonymity which acts as an extension of k-anonymity. (α, k)-anonymity limits the confidence of the disclosure between a $\mathcal{QI}$ and *SA* value within a threshold α, to enhance the protection of sensitive information. The extension of k-anonymity relies on the following requirement.

Definition 9 (α - Deassociation Requirement): Given a table T, an attribute set $\mathcal{QI}$ and an SA value s. Let (E, s) be the set of tuples in EC E containing s for SA and α be a user-specified threshold, where $0 < \alpha < 1$. Dataset T is a α -deassociated with respect to attribute set $\mathcal{QI}$ and the sensitive value s if the relative frequency of s in every EC is less than or equal to α. That is $|(E, s)|/|E| \leq \alpha$ for all ECs E

Therefore, (α, k)-anonymity is defined as follows:

Definition 10 ((α, k)-anonymity): A table T^* is an (α, k)-anonymisation of the table T if it satisfies both k-anonymity and the α-deassociation properties with respect to $\mathcal{QI}$.

7) PERSONALISED PRIVACY

The concept of *Personalised Privacy* [113] was introduced by Xiao and Tao for categorical SA with a taxonomy. In this case, each individual defines the desired level of privacy, rather than applying the same level to all of them. This is achieved by means of *guarding nodes*. A guarding node is a node in the SA taxonomy which a user can reveal. The personalised privacy level is achieved by limiting the breach probability of any leaf value under the guarding node within a user-defined threshold. Depending on the selected privacy level, some individuals may hinder the utility of a subset of the released data.

8) (k, e) AND (ϵ, m) ANONYMITY

(k, e)-Anonymity is a pertubation-based method proposed by Zhang *et al.* [144]. This approach aims to cover the gap in protecting numerical SA, given that ℓ-diversity is designed only for categorical SA. (k, e)-Anonymity requires that each EC has at least k different sensitive values, while the range of sensitive values in the EC is no less than a threshold e.

However, (k, e)-Anonymity has a major drawback: it ignores the distribution of sensitive values within the group. This could open the door to proximity attacks which allow an attacker to discover with high confidence that a numeric sensitive value falls within a short interval, even if the re-identification of its exact value has low confidence.

Example 6: Let us assume that in an EC we have 4 records and the salary corresponds to the SA. If an adversary knows that his victim is in this EC and the possible salary values are {1000, 1030, 1050, 4000}, the adversary has 25% chance to discover the real salary of his victim but he can also infer with 75% probability that his victim's salary is in the interval $[1000 - 1050]$.

As a countermeasure to proximity attacks, Li *et al.* [194] extended (k, e)-Anonymity by introducing (ϵ,m)-Anonymity. (ϵ,m)-Anonymity requires that for every SA value in an EC, at most $\frac{1}{m}$ of the records will have similar values. The similarity is controlled by ϵ and can take different values, such as the absolute difference $| y-x | \leq \epsilon$ or a relative one $| y-x | \leq \epsilon x$.

Another countermeasure to proximity attacks is the Worst Group Protection (WGP) introduced by Loukides and Shao [195]. This approach prevents range disclosure and can be applied without generalising SA values. WGP measures the probability of disclosing any range in the least protected group of a table, and captures the way SA values form ranges in a group, based on their frequency and similarity. WGP handles both numerical and categorical attributes and also considers the possible background knowledge of an attacker.

9) t-CLOSENESS

Even though ℓ-diversity solves many of the weaknesses of k-anonymity in protecting against attribute linkage, it presents several shortcomings. Apart from the fact that ℓ-diversity may be difficult to be achieved in lots of cases, below we provide an example to demonstrate that privacy protection offered by ℓ-diversity may be insufficient as well.

Example 7: Suppose that the original data on table T has only one SA: the test result of a rare virus infection represented by a boolean value "Positive" (True) or "Negative" (False). Let us assume that T contains $10K$ records, 99% of which are negative, and thus only 1% are positive. Clearly, the two values have different degrees of sensitivity. For instance, if we assume that the virus is the HIV, a possible infection (positive value) may have a social impact on the patient, whereas the disclosure of the negative result would have a low impact as 99% of the population in our example has the same result. Unavoidably, an individual having a positive result would like to avoid its disclosure. In this case, to have a distinct 2-diverse table, there can exist at most $10000 \times 1\% = 100$ ECs. Hence, the information loss in such case would be prohibitive.

Based on the aforementioned example, we explain clearly below both *Skewness Attack* and *Similarity Attack* in action.

- **Skewness Attack:** In the case of a skewed distribution, satisfying ℓ-diversity does not prevent attribute disclosure. For instance, considering the previous example and assuming that one EC has an equal number of positive and negative records, the table T satisfies distinct 2-diversity and its variations, entropy 2-diversity, and any recursive (c,2)-diversity [172] requirement that can be imposed. However, it is exposed to a serious privacy risk since anyone in the class would have 50% chances of being positive, as compared with the 1% of the overall population. Another relevant issue in terms of privacy risks occurs when an EC is 2-diverse by having 49 positive records and 1 negative. While the overall possibility of being positive is 1%, in this specific EC this chance is raised to 98%, hindering the privacy of the individuals.
- **Similarity Attack:** Since ℓ-diversity does not consider the semantical closeness of the values, a similarity attack can be triggered. For instance, let us assume that an adversary finds the EC of his target in an anonymous medical publication which is 3-diverse and the three values of this class are *(gastric ulcer, gastritis, stomach cancer)*. In this case, it is obvious that the adversary can deduce that an individual has a stomach-related issue, regardless of the specific disease of his target.

With the aim to overcome these attacks, a new privacy notion, t-closeness, was proposed by Li et al [196], [197]. t-closeness requires the distribution of an SA in any $\mathcal{QI}$ group to be close to the distribution of the SA in the original table T.

Definition 11 (t-closeness): An EC satisfies the t-closeness requirement if the distance of an SA distribution in this class compared to the distribution of that attribute in the whole table is not greater than a threshold t. A table satisfies the t-closeness requirement, if all ECs satisfy the t-closeness requirement.

t-closeness uses the *Earth Mover Distance (EMD)* [198] function to measure the closeness between two distributions of sensitive values and requires this closeness to be within t. Notably, as proved in [199], the complexity of t-closeness for every constant t such that $0 \leq t < 1$, it is NP-hard to find the optimal t-closeness generalisation of T.

To give an insight of the EMD function, let us assume a field with dug holes and the corresponding quantity of soil to fill these holes to be dispersed in different points on that field. EMD would calculate the least amount of work that someone would need to fill in the holes. Let us assume that a unit of work corresponds to transporting a unit of soil by a unit of (ground) distance. Formally, EMD evaluates the dissimilarity between two multi-dimensional distributions in a feature space for a given distance measure, which in this case would be the ground distance. Therefore, EMD lifts this distance from individual features to full distributions.

Cao *et al.* [200] proposed the SABRE framework which was based on t-closeness principle for both categorical and numerical attributes. In this framework, the data are first partitioned into a set of buckets and then form ECs by selecting the appropriate number of records, a task that is performed considering the t-closeness requirement from each bucket. The authors argue that algorithms for t-closeness that are built on top of k-anonymisation [196], [197] fail in terms of efficiency, whereas their experimental evaluation shows that SABRE achieves higher information quality. In [201] the authors use and evaluate three microaggregation-based methods to achieve k-anonymous t-closeness datasets.

10) β-LIKENESS

The t-closeness notion exhibits several limitations and weaknesses. First, it lacks the flexibility of specifying different protection levels for different sensitive values. Second, the EMD function is not suitable for preventing attribute linkage on numerical SAs. Moreover, enforcing t-closeness would substantially hinder the data utility as it requires the distribution of sensitive values to be the same in all QI groups.

Example 8: Assume a dataset T with SA values HIV and Flu. If the overall SA distribution between them is $P = (0.4, 0.6)$ and their distribution in an EC is $Q = (0.5, 0.5)$, then $EMD(P, Q) = 0.1$. Still, if their overall distribution is $P' = (0.01, 0.99)$ and their distribution in an EC is $Q' = (0.11, 0.89)$, then $EMD(P', Q') = 0.1$ again. Clearly, both cases satisfy 0.1-closeness. However, the information gain in the latter is larger than in the former one, because in the first case the probability of HIV is increased by 25%, from 0.4 to 0.5, while in the second by 1000% from 0.01 to 0.11. In effect, the two cases do not afford the same privacy. Unfortunately, any function that aggregates absolute differences (as in the case of EMD) faces the same problem.

Likewise t-closeness, β-likeness [202] is a privacy model for categorical data.

Definition 12 (basic β-likeness): Given a table T with a sensitive attribute SA_1, let $V = \{v_1, v_2, \ldots, v_m\}$ and $P = (p_1, p_2, \ldots, p_m)$ the overall SA_1 distribution in T. An EC G with SA_1 distribution $Q = (q_1, q_2, \ldots, q_m)$ is said to satisfy basic β-likeness, if and only if $max\{D(p_i, q_i)|p_i \in P, p_i < q_i\} \leq \beta$, where $\beta > 0$ is a threshold.

For an anonymised table T^* from T to satisfy the β-likeness, all ECs $G \subset T^*$ have to comply with the β-likeness requirement.

11) δ-PRESENCE

δ-presence [173], [174] is a metric to evaluate the risk of identifying an individual in a table based on the generalisation of publicly known data.

Definition 13 (δ-presence): Given an external table T_p and a private table T, we say that δ-presence holds for a generalisation T^* of T, with $\delta = (\delta_{min}, \delta_{max})$ if:

$$\delta_{min} \leq P(t \in T|T^*) \leq \delta_{max}, \quad \forall t \in P$$

In such datasets, we say that each tuple $t \in P$ is δ-present in T. Therefore, $\delta = (\delta_{min}, \delta_{max})$ is a range of acceptable probabilities. The parameters δ_{min} and δ_{max} define the trade-off between the utility and the privacy of the anonymised table T^*. The increase in δ_{min} leads to better privacy protection, as more information is hidden. Similarly, when δ_{max} decreases the utility rises, but at the cost of lowering the level of privacy. The data publisher should select the maximal δ_{min} and the minimal δ_{max} value that guarantee her desired thresholds on privacy and usability of the data. For more information on selecting the appropriate δ_{min} and δ_{max} the interested reader may refer to [174].

A drawback of δ-presence is that it requires all the available publicly known data to be in the form of a table. The c-confident δ-presence [174], an extension of δ-presence, was introduced to address this issue by relaxing the assumption on the availability of a public table to the publisher. The c-confident δ-presence assumes that data publisher has some knowledge, for example statistics and count queries, about the world from which the table T was drawn from. Such information is not sensitive and more likely to be publicly available. On the other hand, the assumption for the adversary is that she has access to the whole world knowledge in the form of a public table. The only thing she does not know is the presence or the absence of individuals in the private table.

C. OTHER COUNTERMEASURES TO ATTACKS

In addition to the well-known countermeasures described in Sections VI-A and VI-B, several techniques take into

account scenarios that are not usually considered in PPDP. In the next subsections, we describe the m-invariance and m-confidentiality concepts. Later in Section VI-D we introduce the notion of ϵ-Differential Privacy and its relevance to the PPDP field.

1) m-INVARIANCE

A scenario that most data protection methods do not take into consideration is the future re-publication of the anonymous table T^*. Insertions and deletions in the original table T happening through the pass of time make the anonymisation procedure to generate different T^* anonymous tables over time. An adversary can use multiple time releases of T^* tables to infer sensitive information just by comparing them. On the assumption that deletions in T are not allowed, a naïve approach would be to anonymise the new records separately from the already anonymous published table. However, if there is a small amount of new data, then this would lead to severe information loss. Moreover, it is difficult to analyse a collection of datasets with different levels of generalisation for each. For instance, if in a previous release the required generalisation reached to country level (e.g. Greece) but in a second release at the city level (e.g. Athens), aggregation queries such as those for counting the people in the city of Athens would be rendered almost useless.

Another solution is to require the latest release to be no more specialised than the previous [5]. The problem here is that, although the new data in general lead to a better anonymisation, each subsequent release gets increasingly distorted. This drawback led Xiao and Tao to propose the m-invariance method [72]. Yet, before proceeding with the description of m-invariance we introduce below some more necessary definitions.

Definition 14 (Historical Union): At time $n \geq 1$, the historical union $U(n)$ contains all the tuples in T at timestamps $1, 2, \ldots, n$, respectively. Formally:

$$U(n) = \bigcup_{j=1}^{n} T(j)$$

Definition 15 (Lifespan): Each tuple $t \in U(n)$ is implicitly associated with a lifespan $[x, y]$, where x is the smallest and y is the largest integer j such that t appears in $T(j)$.

Using the above, m-invariance can be defined as follows:

Definition 16 (m-invariance): A sequence release of $T_1, T_2, \ldots, T_p$ is m-invariant if the following properties are met:

- every $\mathcal{QI}$ group in any T_i has at least m records and all records in a $\mathcal{QI}$ group have different values on the sensitive attribute.
- for any record r with published lifespan $[x, y]$ where $1 \leq x, y \leq p$, $QI_x, \ldots, QI_y$ have the same set of sensitive values, where $QI_x, \ldots, QI_y$ are the generalised $\mathcal{QI}$ groups containing r in $T_x, \ldots, T_y$.

The rationale behind m-invariance is that if a record r has been published in different anonymous releases $T_x, \ldots, T_y$, then all the ECs containing that record r in all $T_x, .., T_y$ are required to have the same set of SA values. This is done to ensure that the intersection of SA values over all such ECs does not reduce the set of SA values. One of the drawbacks of this method is that, in order to achieve the m-invariance, Xiao's and Tao's algorithm adds the minimum required counterfeit data records, which results in the loss of truthfulness at the record level.

2) m-CONFIDENTIALITY

Before defining m-confidentiality [203] we need to define *Credibility*.

Definition 17 (Credibility): Let T^* be a published anonymous table generated from T. Consider an individual $o \in O$ and a sensitive value set s in the sensitive attribute. $Credibility(o, s, K_{ad})$ is the probability that an adversary can infer from T^* and background knowledge K_{ad} that o is associated with s.

The background knowledge mentioned here refers to the *minimality principle*, as a mean for *Algorithms' exploitation*.

Definition 18 (m-confidentiality): A table T is said to satisfy m-confidentiality if for any individual o and any sensitive value set s, $Credibility(o, s, K_{ad})$ does not exceed $\frac{1}{m}$.

m-confidentiality restricts the probability that an adversary can infer from T^* the association between any individual and a record in T to $\frac{1}{m}$ by taking into account the adversary's background knowledge.

D. DIFFERENTIAL PRIVACY

In 2006, Dwork [204] proposed the notion of Differential Privacy which is commonly used for PPDM and interactive query answering. Differential privacy states that the risk to one's privacy should not substantially (as bounded by a parameter ϵ) increase as a result of participating in a statistical database. Thus, an attacker should not be able to learn any information about any participant that they could not learn if the participant had opted out of the database. In this paper, we adopt the definition presented in [205].

Definition 19 (Differential Privacy): A privacy mechanism A gives ϵ-differential privacy if for any dataset T_1 and T_2 differing at most one record, and for any possible anonymous $T^* \in Range(A)$,

$$Pr[A(T_1) = T^*] \leq e^{\epsilon} \times Pr[A(T_2) = T^*]$$

where the probability is taken over the randomness of A.

However, this definition of ϵ-differential privacy is too restrictive to be satisfiable in some scenarios. Therefore, to increase the functionality of differential privacy with respect to more particular and sensitive queries, several relaxations have been developed, with the most widely-adopted being (ϵ, δ)-differential privacy [206]. In our case, this relaxation is adopted by allowing the output T^* to violate the inequality of Definition 19 with a small error probability δ. Thus, (ϵ, δ)-differential privacy ensures that for all adjacent queries, the absolute value of the privacy loss will be bounded by ϵ with probability at least $1 - \delta$.

TABLE 13. Attacks each Privacy Model prevents.

Privacy Model	Record Linkage	Attribute Linkage	Table Linkage	Skewness / Similarity Attacks	Attacks on Continuous Data Publishing	Algorithm exploitation
k-anonymity [5], [175]	✓					
(X,Y)-Anonymity [71]	✓					
(k,k)-anonymity [181]	✓					
n-Confusion [184]	✓					
MultiRelational k-Anonymity [185]	✓					
k^m-anonymity [186]	✓					
ϵ-Differential Privacy [204]			✓	✓		
k-concealment [183]	✓					
Confidence bounding [137], [187]		✓				
ℓ-diversity [172]	✓	✓				
(X,Y) - Privacy [71]		✓				
(α, k)-anonymity [193]	✓	✓				
Personalized Privacy [113]		✓				
(k, e)-Anonymity [144]		✓				
$(\epsilon$, m)-Anonymity [194]		✓				
t-closeness [196], [197]		✓		✓		
β-likeness [202]		✓		✓		
δ-presence [173], [174]			✓			
m-invariance [72]					✓	
m-confidentiality [203]						✓

Among the multiple mechanisms used to obtain differential privacy for real-valued queries [206] the most widely-used is the Laplace mechanism [146]. Given a function f, the Laplace mechanism will perturb each coordinate with noise drawn from the Laplace distribution, defined as:

$$Lap(x|b) = \frac{1}{2b} e^{-\frac{|x|}{b}} \tag{5}$$

where b denotes the scale. Hence, the variance of this distribution is $\sigma^2 = 2b^2$. Note that this distribution is centered at 0. The scale of the noise would be calibrated to the sensitivity of f (divided by ϵ). Also, using Gaussian noise with variance calibrated to $\Delta f ln(1/\delta)/\epsilon$, one can achieve (ϵ, δ)-differential privacy [206].

As previously stated, the differential privacy notion is not novel. A survey of the first works towards this breakthrough in privacy-preserving data analysis can be found in [207]. However, the vast majority of research on differential privacy are for answering statistical queries, rather than publishing microdata [208], [209]. More precisely, differential privacy is based on adding noise to query results [146], [210], [211] (*i.e.* output perturbation). Therefore, although some PPDP works based on differential privacy can be found [19], [205], [208], [212]–[219] this approach remains useful as far as statistical results are concerned.

Microdata publishing in the context of differential privacy has been widely studied [208]. Machanavajjhala *et al.* [220] introduced a variant of (ϵ, δ)-differential privacy named (ϵ, δ)-probabilistic differential privacy. The authors used a synthetic data generation method to release privately commuting patterns of the population in the US privately. Differential privacy for secure release of search queries has also attracted the attention of many researchers such as Korolova *et al.* [221] and Gotz *et al.* [222].

While most privacy concepts are easily to be understood by data publishers, data recipients, and record holders, differential privacy is quite theoretical in nature and thus difficult to be explained in terms of its level of anonymisation guarantee [223], [224]. Moreover, as claimed by several authors, differential privacy has not yet reached the maturity to replace other existing models of PPDP [225]. Nevertheless, the introduction of Local Differential Privacy (LDP) partially solved some of these drawbacks, enhancing the trade-off between privacy and efficiency [226]–[228]. LDP is a data collection framework in which each contributor locally perturbs data using a mechanism that guarantees differential privacy (e.g. Laplace noise). Next, this perturbed data are sent to the data collector, avoiding, for instance, repeated query attacks, since all responses would be based on this perturbed version of the data. This method is usually combined with a randomized response [226]. Therefore, each response would be randomly obfuscated (according to a predefined level of privacy) in a way in which statistical indicators are preserved, considering a large enough number of responses. Moreover, the use of randomized response can be extended to complex data types and sophisticated statistics at the data collector, so that information can be extracted from repeated queries without hindering the privacy of individuals [228], [229].

Finally, we observe that current definitions of (ϵ, δ)-differential privacy require δ to be very small to provide sufficient privacy protection when publishing microdata, hindering the efficiency of such method. This is particularly challenging in some scenarios such as high dimensional datasets [230]–[232]. Therefore, realistic attacks should be analysed more profoundly to provide for insights about what ϵ-differential privacy (and its variants) means in practice and how its drawbacks can be minimised [233], especially in terms of the new era of big data [231], [234], [235].

A summary of all the privacy models described in the paper and the attacks that each one of them prevents is depicted in Table 13. Note that all the countermeasures in Table 13, except differential privacy, offer "truthfulness on the record level". That means each record on the anonymised table T^*, no matter how much generalised, corresponds to a record on the original table T. On the contrary, similar to randomisation, differential privacy does not guarantee such property [236], a fact that can affect the quality of the released data, rendering them useless (e.g. in the case of existing mutually exclusive output queries) for some applications [220].

VII. OPEN QUESTIONS AND FUTURE DIRECTIONS

The recent advances in IoT, machine learning and big data analytics have significantly increased the requests for data resources. In view of these countless requests, the use of privacy-preserving methods to provide data without exposing users' privacy is mandatory. Examples of such practices can be already seen in companies such as Apple which embraced the use of differential privacy in its products [237]. Moreover, there is a wide range of companies and startups, such as Aircloak,[3] whose primary service is to provide dataset anonymisation, thus paving the way for the greater adoption of PPDP. However, regardless of the currently powerful mechanisms and anonymisation procedures implemented in the PPDP context, several open issues need to be addressed in the upcoming years. An overview of these open questions and challenges in the PPDP research area is provided in Figure 6.

One of the biggest challenges for the wider dissemination of PPDP methods is the management of Big Data, as its three "V"s, namely Volume, Variety and Velocity, impose many constraints to their adoption. Volume, an inherent characteristic of Big Data, affects substantially the PPDP algorithms since data processing and sanitisation demand high computational power. In this regard, microaggregation methods offer an attractive computational trade-off among computational complexity, privacy, and information loss. Besides microaggregation, differential privacy attracts also a significant research interest in the context of Big Data [238], [239]. Big Data Variety adds further constraints to anonymization algorithms since methods such as Generalisation are highly dependant on data variations. Notwithstanding the two aforesaid "V"s, the third "V", Velocity, appears to be the most challenging one as Big Data are being generated and stored at an unprecedented rate. It appears that for high velocity data that need to be anonymised on the fly the only thus far viable solutions rely on (Local) differential privacy. However, in the context of sequential data publications, the security of the data that are anonymised through differential privacy methods has not been studied thoroughly. Yet, as practice has shown, this might be subject to the underlying implementation of each particular application.

While, as already discussed, differential privacy has great potential for PPDP, it is far from being considered a panacea or immune to attacks. For instance, Clifton and Tassa [225] have already criticised the widespread belief that differential privacy is resistant to attacks. As a matter of fact, as shown by Cormode in [240], even under differential privacy the accurate inference of private attributes from realistic data is possible. What's more, the noise addition is often subject to the queries that are expected to be performed on the dataset, or to whether there exist sequential data publications. In that regard, some recent research studies explore the generation of incremental ϵ-differentially private releases of data so as to address the demand for up-to-date information [241], [242]. However, in such cases the incremental privacy risk that arises when different anonymised versions of the same dataset are released, has to be also taken into consideration. While several approaches to this challenge can be found in the literature [74]–[76], [243], yet there is still a big gap to be filled. Furthermore, it has been shown that machine learning algorithms can be exploited to extract further knowledge from the anonymised dataset [244]. Based on the above, it is clear that although differential privacy might be very useful in PPDP, its privacy guarantees as well as its implementations and use case scenarios have to be further investigated.

Another very active field in PPDP is the study of high-dimensional data [245]. As previously stated, the curse of dimensionality has a significant impact on the k-anonymity model [176]. These days, due to the prevailing vast collection of data from mobile sensors and ubiquitous computing devices [246]–[249], high-dimensional data are not only found in healthcare anymore - in which traditionally PPDP is applied - but to other application domains as well. Hence, the development of proper anonymisation mechanisms that can be extended to many and diverse application areas becomes compulsory [250]–[253]. While there are several proposals for the treatment of such high-dimensional

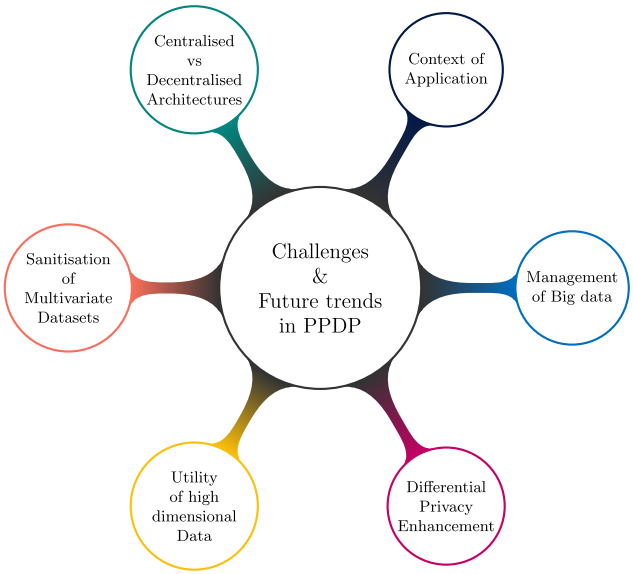


FIGURE 6. Mindmap representation of the challenges and future trends in PPDP.

[3]https://www.aircloak.com/

datasets [254]–[257], more generic approaches try to exploit either the fact that in real datasets the actual background information cannot span to many QIs, or the inherent correlation of many QIs [258], [259]. In general, the current state of the art is rather promising and leads the way in fruitful future research extensions [260]. Nevertheless, while the proposed methods provide enough privacy guarantees, they fail to provide adequate levels of data utility. Therefore, this unbalanced trade-off hinders their adoption.

Closely related to the above, multivariate datasets (i.e. which if they are high-dimensional face the challenges described in the previous paragraph as well) present several well-known challenges. Amongst others, we consider the proper detection of outlier users (i.e. *outliers*) as a significant issue since the presence of outliers in datasets may lead to the disclosure of information about the data distribution, thereby hindering the quality of the obfuscation. For instance, clustering algorithms may incorrectly select the proper users to form a group, creating inaccurate data representations. In this regard, several approaches to deal with outliers in multivariate datasets have been proposed in the literature [261]–[269].

Nowadays, due to the need for massive amounts of real-time data as well as seamless and continuous user interactions, centralised data publication approaches are shifting towards more decentralised solutions. As already discussed, novel cryptographic solutions have been developed to face such novel approaches. Nevertheless, these primitives imply a significant computational cost, especially for large-scale distributed networks [270]. To overcome this challenge, one strategy would be to enhance secure multiparty protocols in terms of computational and communication costs. On top of that, application-oriented mechanisms may also be implemented to deal with specific scenarios and requirements.

Beyond all the above challenges, the models discussed so far fail to take into account some mining patterns which, depending on the context of data, may infer sensitive information. In healthcare, for example, a link between hospitalisation costs and a particular ZIP area may disclose information that could be abused by insurance companies. In this context, the goal is not to identify individuals or their information but to infer knowledge about specific group of individuals [271]. To protect against this challenge, similar sensitive knowledge patterns must be identified before applying further operations on data or sharing them. Against this background, the active field of research in group privacy is steadily growing [272].

Overall, it may be said that the type and context of published data are very relevant to the way they can be protected since all attacks are crafted to the context of information. Hence, prior to de-anonymization, all known external data sources that may be used for attacking the anonymized data can be used in our advantage to selectively reinforce data anonymisation procedures where deemed appropriate [273]. Above all, we must identify the specific weaknesses of each application domain so as to protect data efficiently.

REFERENCES

[1] H. Ye, X. Cheng, M. Yuan, L. Xu, J. Gao, and C. Cheng, "A survey of security and privacy in big data," in *Proc. 16th Int. Symp. Commun. Inf. Technol. (ISCIT)*, Sep. 2016, pp. 268–272.

[2] D. Vatsalan, Z. Sehili, P. Christen, and E. Rahm, *Privacy-Preserving Record Linkage for Big Data: Current Approaches and Research Challenges*. Cham, Switzerland: Springer, 2017, pp. 851–895.

[3] A. Mehmood, I. Natgunanathan, Y. Xiang, G. Hua, and S. Guo, "Protection of big data privacy," *IEEE Access*, vol. 4, pp. 1821–1834, 2016.

[4] M. Barbaro, T. Zeller, and S. Hansell, "A face is exposed for AOL searcher, no. 4417749," New York Times, 2006. [Online]. Available: http://www.nytimes.com/2006/08/09/technology/09aol.html

[5] L. Sweeney, "K-Anonymity: A model for protecting privacy," *Int. J. Uncertainty, Fuzziness Knowl.-Based Syst.*, vol. 10, no. 5, pp. 557–570, Oct. 2002.

[6] L. Sweeney, "Simple demographics often identify people uniquely," *Health (San Francisco)*, vol. 671, pp. 1–34, Jan. 2000.

[7] P. Golle, "Revisiting the uniqueness of simple demographics in the US population," in *Proc. 5th ACM Workshop Privacy Electron. Soc. (WPES)*, 2006, pp. 77–80.

[8] M. R. Koot, G. van't Noordende, and C. de Laat, "A study on the re-identifiability of Dutch citizens," in *Proc. HotPETs*, 2010, pp. 1–16.

[9] K. El Emam, D. Buckeridge, R. Tamblyn, A. Neisa, E. Jonker, and A. Verma, "The re-identification risk of canadians from longitudinal demographics," *BMC Med. Informat. Decis. Making*, vol. 11, no. 1, p. 46, Dec. 2011.

[10] European Commission. *European Union Data Protection Directive 95/46/EC*. Accessed: Sep. 10, 2019. [Online]. Available: http://eurlex.europa.eu/LexUriServ/LexUriServ.do?uri=CELEX:31995L0046:EN:HTML

[11] A. Narayanan and V. Shmatikov, "Myths and fallacies of 'personally identifiable information,'" *Commun. ACM*, vol. 53, no. 6, pp. 24–26, 2010.

[12] E. Politou, E. Alepis, and C. Patsakis, "Forgetting personal data and revoking consent under the GDPR: Challenges and proposed solutions," *J. Cybersecur.*, vol. 4, no. 1, Jan. 2018, Art. no. tyy001, doi: 10.1093/cybsec/tyy001.

[13] A. Zigomitros, "Content-based information retrieval and anonymisation in data and multimedia streams," Ph.D. dissertation, Piraeus Univ., Pireas, Greece, 2018. [Online]. Available: http://dione.lib.unipi.gr/xmlui/handle/unipi/11465

[14] *10 Key Marketing Trends for 2017*. (2017). [Online]. Available: https://www-01.ibm.com/common/ssi/cgi-bin/ssialias?htmlfid=WRL12345USEN

[15] B.-C. Chen, D. Kifer, K. LeFevre, and A. Machanavajjhala, "Privacy-preserving data publishing," *Found. Trends Databases*, vol. 2, nos. 1–2, pp. 1–167, 2009.

[16] B. Fung, K. Wang, R. Chen, and P. S. Yu, "Privacy-preserving data publishing: A survey of recent developments," *ACM Comput. Surv.*, vol. 42, no. 4, pp. 1–53, 2010.

[17] S. D. C. di Vimercati, S. Foresti, G. Livraga, and P. Samarati, "Data privacy: Definitions and techniques," *Int. J. Uncertainty, Fuzziness Knowl.-Based Syst.*, vol. 20, no. 06, pp. 793–817, 2012.

[18] J. Domingo-Ferrer, D. Sánchez, and S. Hajian, "Database Privacy," in *Privacy in a Digital, Networked World* (Computer Communications and Networks), S. Zeadally and M. Badra, Eds. Springer, 2015, pp. 9–35.

[19] T. Zhu, G. Li, W. Zhou, and P. S. Yu, "Differentially private data publishing and analysis: A survey," *IEEE Trans. Knowl. Data Eng.*, vol. 29, no. 8, pp. 1619–1638, Aug. 2017.

[20] K. El Emam, E. Jonker, L. Arbuckle, and B. Malin, "A systematic review of re-identification attacks on health data," *PLoS ONE*, vol. 6, no. 12, 2011, Art. no. e28071.

[21] S. Bender, R. Brand, and J. Bacher, "Re-identifying register data by survey data: An empirical study," *Stat. J. United Nations Econ. Commission Eur.*, vol. 18, no. 4, pp. 373–381, Dec. 2001.

[22] S. Ochoa, J. Rasmussen, C. Robson, and M. Salib, "Reidentification of individuals in Chicago's homicide database: A technical and legal study," Massachusetts Inst. Technol., Cambridge, MA, USA, Tech. Rep., 2001.

[23] M. Elliot and K. Purdam, "The evaluation of risk from identification attempts," Univ. Manchester, Manchester, U.K., CASC Project Deliverable 5D3, 2003

[24] J. S. Brownstein, C. A. Cassa, and K. D. Mandl, "No place to hide—Reverse identification of patients from published maps," *New England J. Med.*, vol. 355, no. 16, pp. 1741–1742, 2006.

[25] D. Frankowski, D. Cosley, S. Sen, L. Terveen, and J. Riedl, "You are what you say: Privacy risks of public mentions," in *Proc. 29th Annu. Int. ACM SIGIR Conf. Res. Develop. Inf. Retr.*, 2006, pp. 565–572.
[26] *The Supreme Court of the State of Illinois (2006) Southern Illinoisan vs. The Illinois Department of Public Health*, docket no. 98712, Supreme Court of the State of Illinois, Springfield, IL, USA, 2006. [Online]. Available: http://www.state.il.us/
[27] B. Malin, "Re-identification of familial database records," in *Proc. AMIA Annu. Symp.*, 2006, p. 524.
[28] L. Backstrom, C. Dwork, and J. Kleinberg, "Wherefore art thou r3579x?: Anonymized social networks, hidden patterns, and structural steganography," in *Proc. 16th Int. Conf. World Wide Web*, 2007, pp. 181–190.
[29] *Federal Court: Canada (2007) Mike Gordon v. the Minister of Health and the Privacy Commissioner of Canada: Memorandum of Fact and Law of the Privacy Commissioner of Canada. Federal Court.*, Federal Court of Canada,Ottawa, ON, Canada, 2007.
[30] A. Narayanan and V. Shmatikov, "Robust de-anonymization of large sparse datasets," in *Proc. IEEE Symp. Secur. Privacy (SP)*, May 2008, pp. 111–125.
[31] K. El Emam and P. Kosseim, "Privacy interests in prescription data, part 2: Patient privacy," *IEEE Secur. Privacy Mag.*, vol. 7, no. 2, pp. 75–78, Mar. 2009.
[32] A. Narayanan and V. Shmatikov, "De-anonymizing social networks," in *Proc. 30th IEEE Symp. Secur. Privacy*, May 2009, pp. 173–187.
[33] P. Kwok, M. Davern, E. Hair, and D. Lafky, "Harder than you think: A case study of re-identification risk of HIPAA-compliant records," Chicago: NORC Univ., Chicago, IL, USA, Abstract Tech. Rep. 302255, 2011.
[34] D. Lafky, "The safe harbor method of de-identification: An empirical test," Dept. Health Hum. Services, Office Nat. Coordinator Health Inf. Technol., Washington, DC, USA, Oct. 2009, pp. 1–24.
[35] L. Sweeney and J. Yoo, "De-anonymizing South Korean resident registration numbers shared in prescription data," *Technol. Sci.*, pp. 1–27, Sep. 2015, Art. no. 2015092901.
[36] V. S. Verykios, E. Bertino, I. N. Fovino, L. P. Provenza, Y. Saygin, and Y. Theodoridis, "State-of-the-art in privacy preserving data mining," *ACM SIGMOD Rec.*, vol. 33, no. 1, pp. 50–57, 2004.
[37] R. Agrawal and R. Srikant, "Privacy-preserving data mining," in *Proc. ACM SIGMOD Int. Conf. Manage. Data (SIGMOD)*. New York, NY, USA: Association for Computing Machinery, 2000, pp. 439–450. doi: 10.1145/342009.335438.
[38] M. B. Malik, M. A. Ghazi, and R. Ali, "Privacy preserving data mining techniques: Current scenario and future prospects," in *Proc. 3rd Int. Conf. Comput. Commun. Technol.*, Nov. 2012, pp. 26–32.
[39] L. Xu, C. Jiang, J. Wang, J. Yuan, and Y. Ren, "Information security in big data: Privacy and data mining," *IEEE Access*, vol. 2, pp. 1149–1176, 2014.
[40] K. Kenthapadi, I. Mironov, and A. Thakurta, "Privacy-preserving data mining in industry," in *Proc. Companion World Wide Web Conf. (WWW)*, 2019, pp. 840–841.
[41] R. Mendes and J. P. Vilela, "Privacy-preserving data mining: Methods, metrics, and applications," *IEEE Access*, vol. 5, pp. 10562–10582, 2017.
[42] Z. Feng and Y. Zhu, "A survey on trajectory data mining: Techniques and applications," *IEEE Access*, vol. 4, pp. 2056–2067, 2016.
[43] A. Machanavajjhala and D. Kifer, "Designing statistical privacy for your data," *Commun. ACM*, vol. 58, no. 3, pp. 58–67, Feb. 2015.
[44] J. Domingo-Ferrer, "A three-dimensional conceptual framework for database privacy," in *Proc. Workshop Secure Data Manage.*, 2007, pp. 193–202.
[45] A. Martinez-Balleste, P. Perez-martinez, and A. Solanas, "The pursuit of citizens' privacy: A privacy-aware smart city is possible," *IEEE Commun. Mag.*, vol. 51, no. 6, pp. 136–141, Jun. 2013.
[46] P. A. Pérez-Martínez and A. Solanas, "W3-privacy: The three dimensions of user privacy in LBS," in *Proc. 12th ACM Int. Symp. Mobile Ad Hoc Netw. Comput.*, 2011, pp. 1–2.
[47] A. Solanas, E. Batista, F. Casino, A. Papageorgiou, and C. Patsakis, "Privacy-oriented analysis of ubiquitous computing systems: A 5-D approach," in *Security of Ubiquitous Computing Systems*. Cham, Switzerland: Springer, 2021, ch. 12, pp. 181–194.
[48] A. R. Beresford and F. Stajano, "Mix zones: User privacy in location-aware services," in *Proc. 2nd IEEE Annu. Conf. Pervas. Comput. Commun. Workshops*, Mar. 2004, pp. 127–131.
[49] C. Bettini, X. S. Wang, and S. Jajodia, "Protecting privacy against location-based personal identification," in *Secure Data Management*. Berlin, Germany: Springer, 2005, pp. 185–199.
[50] F. Olumofin, P. K. Tysowski, I. Goldberg, and U. Hengartner, "Achieving efficient query privacy for location based services," in *Privacy Enhancing Technologies*. Berlin, Germany: Springer, 2010, pp. 93–110.
[51] A. Pingley, N. Zhang, X. Fu, H.-A. Choi, S. Subramaniam, and W. Zhao, "Protection of query privacy for continuous location based services," in *Proc. IEEE INFOCOM*, Apr. 2011, pp. 1710–1718.
[52] J. Wang, Z. Cai, Y. Li, D. Yang, J. Li, and H. Gao, "Protecting query privacy with differentially private k-anonymity in location-based services," *Pers. Ubiquitous Comput.*, vol. 22, no. 3, pp. 453–469, Jun. 2018.
[53] M. Ataei and C. Kray, "Ephemerality is the new black: A novel perspective on location data management and location privacy in LBS," in *Progress in Location-Based Services 2016* (Lecture Notes in Geoinformation and Cartography), G. Gartner and H. Huang, Eds. Cham, Switzerland: Springer, 2017.
[54] O. Abul and C. Bayrak, "From location to location pattern privacy in location-based services," *Knowl. Inf. Syst.*, vol. 56, no. 3, pp. 533–557, Sep. 2018.
[55] Y. Cai and G. Xu, "Cloaking with footprints to provide location privacy protection in location-based services," U.S. Patent 9 736 685, Aug. 15, 2017,
[56] A. Hundepool, J. Domingo-Ferrer, L. Franconi, S. Giessing, E. S. Nordholt, K. Spicer, and P.-P. De Wolf, *Statistical Disclosure Control*. Hoboken, NJ, USA: Wiley, 2012.
[57] G. T. Duncan and S. L. Stokes, "Disclosure risk vs. Data utility: The R-U confidentiality map as applied to topcoding," *Chance*, vol. 17, no. 3, pp. 16–20, Jun. 2004.
[58] F. Casino, E. Politou, E. Alepis, and C. Patsakis, "Immutability and decentralized storage: An analysis of emerging threats," *IEEE Access*, vol. 8, pp. 4737–4744, 2020.
[59] L. Sweeney, "Only you, your doctor, and many others may know," *Technol. Sci.*, pp. 1–22, Sep. 2015, Art. no. 2015092903.
[60] E. J. Bloustein, "Privacy as an aspect of human dignity: An answer to Dean Prosser," *NYUL Rev.*, vol. 39, p. 962, Feb. 1964.
[61] W. A. Parent, "Privacy, morality, and the law," in *Philosophy & Public Affairs*. Evanston, IL, USA: Routledge, 1983, pp. 269–288.
[62] E. Hughes. (1993). *A Cypherpunk's Manifesto*. [Online]. Available: https://w2.eff.org/Privacy/Crypto/Crypto_misc/cypherpunk.manifesto
[63] Z. Li and X. Ye, "Privacy protection on multiple sensitive attributes," in *Information and Communications Security (ICICS)* (Lecture Notes in Computer Science), vol. 4861, S. Qing, H. Imai and G. Wang, Eds. Berlin, Germany: Springer, 2007.
[64] Y. Ye, Y. Liu, C. Wang, D. Lv, and J. Feng, "Decomposition: Privacy preservation for multiple sensitive attributes," in *Database Systems for Advanced Applications (DASFAA)* (Lecture Notes in Computer Science), vol. 5463, X. Zhou, H. Yokota, K. Deng, and Q. Liu, Eds. Berlin, Germany: Springer, 2009.
[65] Y. Wu, X. Ruan, S. Liao, and X. Wang, "P-cover k-anonymity model for protecting multiple sensitive attributes," in *Proc. 5th Int. Conf. Comput. Sci. Edu.*, Aug. 2010, pp. 179–183.
[66] J. Liu, J. Luo, and J. Z. Huang, "Rating: Privacy preservation for multiple attributes with different sensitivity requirements," in *Proc. IEEE 11th Int. Conf. Data Mining Workshops*, Dec. 2011, pp. 666–673.
[67] N. Maheshwarkar, K. Pathak, and N. S. Choudhari, "K-anonymity model for multiple sensitive attributes," *Int. J. Comput. Appl., Optim. On-Chip Commun.*, no. 1, pp. 51–56, Feb. 2012.
[68] C. Yao, X. S. Wang, and S. Jajodia, "Checking for k-anonymity violation by views," in *Proc. 31st Int. Conf. Very Large Data Bases*, 2005, pp. 910–921.
[69] B. Barak, K. Chaudhuri, C. Dwork, S. Kale, F. McSherry, and K. Talwar, "Privacy, accuracy, and consistency too: A holistic solution to contingency table release," in *Proc. 26th ACM SIGMOD-SIGACT-SIGART Symp. Princ. Database Syst.*, 2007, pp. 273–282.
[70] D. Kifer and J. Gehrke, "Injecting utility into anonymized datasets," in *Proc. ACM SIGMOD Int. Conf. Manage. Data (SIGMOD)*, 2006, pp. 217–228.
[71] K. Wang and B. C. M. Fung, "Anonymizing sequential releases," in *Proc. 12th ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining (KDD)*, 2006, pp. 414–423.
[72] X. Xiao and Y. Tao, "M-invariance: Towards privacy preserving re-publication of dynamic datasets," in *Proc. ACM SIGMOD Int. Conf. Manage. Data*, 2007, pp. 689–700.
[73] Y. Bu, A. W. C. Fu, R. C. W. Wong, L. Chen, and J. Li, "Privacy preserving serial data publishing by role composition," *Proc. VLDB Endowment*, vol. 1, no. 1, pp. 845–856, Aug. 2008.

[74] J. Pei, J. Xu, Z. Wang, W. Wang, and K. Wang, "Maintaining K-Anonymity against incremental updates," in *Proc. 19th Int. Conf. Sci. Stat. Database Manage (SSDBM)*, Jul. 2007.

[75] J. W. Byun, Y. Sohn, E. Bertino, and N. Li, "Secure anonymization for incremental datasets," in *Secure Data Management (SDM)* (Lecture Notes in Computer Science), vol. 4165, W. Jonker and M. Petković, Eds. Berlin, Germany: Springer, 2006.

[76] J.-W. Byun, T. Li, E. Bertino, N. Li, and Y. Sohn, "Privacy-preserving incremental data dissemination," *J. Comput. Secur.*, vol. 17, no. 1, pp. 43–68, Mar. 2009.

[77] M. Barbosa, A. Pinto, and B. Gomes, "Generically extending anonymization algorithms to deal with successive queries," in *Proc. 21st ACM Int. Conf. Inf. Knowl. Manage. (CIKM)*, 2012, pp. 1362–1371.

[78] Y. Tao, Y. Tong, S. Tan, S. Tang, and D. Yang, "T-rotation: Multiple publications of privacy preserving data sequence," in *Advanced Data Mining and Applications (ADMA)* (Lecture Notes in Computer Science), vol. 5139, C. Tang, C. X. Ling, X. Zhou, N. J. Cercone, and X. Li, Eds. Berlin, Germany: Springer, 2008.

[79] G. Wang, Z. Zhu, W. Du, and Z. Teng, "Inference analysis in privacy-preserving data re-publishing," in *Proc. 8th IEEE Int. Conf. Data Mining*, Dec. 2008, pp. 1079–1084.

[80] E. Shmueli, T. Tassa, R. Wasserstein, B. Shapira, and L. Rokach, "Limiting disclosure of sensitive data in sequential releases of databases," *Inf. Sci.*, vol. 191, pp. 98–127, May 2012.

[81] E. Shmueli and T. Tassa, "Privacy by diversity in sequential releases of databases," *Inf. Sci.*, vol. 298, pp. 344–372, Mar. 2015, doi: 10.1016/j.ins.2014.11.005.

[82] X. Yang, T. Wang, X. Ren, and W. Yu, "Survey on improving data utility in differentially private sequential data publishing," *IEEE Trans. Big Data*, to be published.

[83] B. Krawczyk, L. L. Minku, J. Gama, J. Stefanowski, and M. Woźniak, "Ensemble learning for data stream analysis: A survey," *Inf. Fusion*, vol. 37, pp. 132–156, Sep. 2017.

[84] S. A. Abdelhameed, S. M. Moussa, and M. E. Khalifa, "Restricted sensitive attributes-based sequential anonymization (RSA-SA) approach for privacy-preserving data stream publishing," *Knowl.-Based Syst.*, vol. 164, pp. 1–20, Jan. 2019.

[85] A. S. M. T. Hasan and Q. Jiang, "A general framework for privacy preserving sequential data publishing," in *Proc. 31st Int. Conf. Adv. Inf. Netw. Appl. Workshops (WAINA)*, Mar. 2017, pp. 519–524.

[86] H. Zhu, H. Liang, L. Zhao, D. Peng, and L. Xiong, "τ-Safe (l, k)-diversity privacy model for sequential publication with high utility," *IEEE Access*, vol. 7, pp. 687–701, 2019.

[87] A. Anjum, G. Raschia, M. Gelgon, A. Khan, N. Ahmad, M. Ahmed, S. Suhail, and M. M. Alam, "τ-safety: A privacy model for sequential publication with arbitrary updates," *Comput. Secur.*, vol. 66, pp. 20–39, May 2017.

[88] J. Salas and V. Torra, "A general algorithm for k-anonymity on dynamic databases," in *Data Privacy Management, Cryptocurrencies and Blockchain Technology (DPM CBT)* (Lecture Notes in Computer Science), vol. 11025, J. Garcia-Alfaro, J. Herrera-Joancomartí, G. Livraga, and R. Rios, Eds. Cham, Switzerland: Springer, 2018, pp. 407–414.

[89] S. Kabou, S. M. Benslimane, and M. Mosteghanemi, "A survey on privacy preserving dynamic data publishing," *Int. J. Org. Collective Intell.*, vol. 8, no. 4, pp. 1–20, Oct. 2018.

[90] M. Bewong, J. Liu, L. Liu, and J. Li, "Privacy preserving serial publication of transactional data," *Inf. Syst.*, vol. 82, pp. 53–70, May 2019.

[91] O. Temuujin, J. Ahn, and D.-H. Im, "Efficient L-Diversity algorithm for preserving privacy of dynamically published datasets," *IEEE Access*, vol. 7, pp. 122878–122888, 2019.

[92] R. C.-W. Wong, A. W.-C. Fu, J. Liu, K. Wang, and Y. Xu, "Global privacy guarantee in serial data publishing," in *Proc. IEEE 26th Int. Conf. Data Eng. (ICDE)*, Mar. 2010, pp. 956–959.

[93] A. Siddiqa, A. Karim, and A. Gani, "Big data storage technologies: A survey," *Frontiers Inf. Technol. Electron. Eng.*, vol. 18, no. 8, pp. 1040–1070, 2017.

[94] S. Phansalkar and S. Ahirrao, "Survey of data partitioning algorithms for big data stores," in *Proc. 4th Int. Conf. Parallel, Distrib. Grid Comput. (PDGC)*, Dec. 2016, pp. 163–168.

[95] K. Wang, B. C. M. Fung, and G. Dong, "Integrating private databases for data analysis," in *Intelligence and Security Informatics (ISI)* (Lecture Notes in Computer Science), vol. 3495, P. Kantor *et al.*, Eds. Berlin, Germany: Springer, 2005.

[96] S. Goryczka, L. Xiong, and B. C. Fung, "m-Privacy for collaborative data publishing," in *Proc. 7th Int. Conf. Collaborative Comput., Netw., Appl. Worksharing (CollaborateCom)*, Oct. 2011, pp. 1–10.

[97] N. Mohammed, B. C. M. Fung, and M. Debbabi, "Anonymity meets game theory: Secure data integration with malicious participants," *VLDB J.*, vol. 20, no. 4, pp. 567–588, Aug. 2011.

[98] J. Soria-Comas and J. Domingo-Ferrer, "Co-utile collaborative anonymization of microdata," in *Modeling Decisions for Artificial Intelligence (MDAI)* (Lecture Notes in Computer Science), vol. 9321, V. Torra and T. Narukawa, Eds. Cham, Switzerland: Springer, 2015.

[99] H. Polat and W. Du, "Privacy-preserving top-N recommendation on distributed data," *J. Amer. Soc. Inf. Sci. Technol.*, vol. 59, no. 7, pp. 1093–1108, May 2008.

[100] C.-L.-A. Hsieh, J. Zhan, D. Zeng, and F. Wang, "Preserving privacy in joining recommender systems," in *Proc. Int. Conf. Inf. Secur. Assurance (ISA)*, Apr. 2008, pp. 561–566.

[101] C. Kaleli and H. Polat, "P2P collaborative filtering with privacy," *Turkish J. Elect. Eng. Comput. Sci.*, vol. 18, no. 1, pp. 101–116, 2010.

[102] I. Yakut and H. Polat, "Privacy-preserving SVD-based collaborative filtering on partitioned data," *Int. J. Inf. Technol. Decis. Making*, vol. 09, no. 03, pp. 473–502, May 2010.

[103] I. Yakut and H. Polat, "Estimating NBC-based recommendations on arbitrarily partitioned data with privacy," *Knowl.-Based Syst.*, vol. 36, pp. 353–362, Dec. 2012.

[104] A. Kumar, M. Gyanchandani, and P. Jain, "A comparative review of privacy preservation techniques in data publishing," in *Proc. 2nd Int. Conf. Inventive Syst. Control (ICISC)*, Jan. 2018, pp. 1027–1032.

[105] A. Solanas, A. Martinez-Balleste, and J. M. Mateo-Sanz, "Distributed architecture with double-phase microaggregation for the private sharing of biomedical data in mobile health," *IEEE Trans. Inf. Forensics Secur.*, vol. 8, no. 6, pp. 901–910, Jun. 2013.

[106] F. Kohlmayer, F. Prasser, C. Eckert, and K. A. Kuhn, "A flexible approach to distributed data anonymization," *J. Biomed. Informat.*, vol. 50, pp. 62–76, Aug. 2014.

[107] D. Karapiperis and V. S. Verykios, "An LSH-based blocking approach with a homomorphic matching technique for privacy-preserving record linkage," *IEEE Trans. Knowl. Data Eng.*, vol. 27, no. 4, pp. 909–921, Apr. 2015.

[108] J. Brickell and V. Shmatikov, "The cost of privacy: Destruction of data-mining utility in anonymized data publishing," in *Proc. 14th ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*, 2008, pp. 70–78.

[109] P. Samarati, "Protecting respondents identities in microdata release," *IEEE Trans. Knowl. Data Eng.*, vol. 13, no. 6, pp. 1010–1027, Nov./Dec. 2001.

[110] L. Sweeney, "Datafly: A system for providing anonymity in medical data," in *Proc. Int. Conf. Database Secur.*, 1998, pp. 356–381.

[111] A. Meyerson and R. Williams, "On the complexity of optimal K-anonymity," in *Proc. 23rd ACM SIGMOD-SIGACT-SIGART Symp. Princ. Database Syst. (PODS)*, 2004, pp. 223–228.

[112] J. Xu, W. Wang, J. Pei, X. Wang, B. Shi, and A. W.-C. Fu, "Utility-based anonymization using local recoding," in *Proc. 12th ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining (KDD)*, 2006, pp. 785–790.

[113] X. Xiao and Y. Tao, "Personalized privacy preservation," in *Proc. ACM SIGMOD Int. Conf. Manage. Data*, 2006, pp. 229–240.

[114] A. Skowron and C. Rauszer, "The discernibility matrices and functions in information systems," in *Intelligent Decision Support. Theory and Decision Library* (Series D: System Theory, Knowledge Engineering and Problem Solving), vol. 11, R. Słowiński, Ed. Dordrecht, The Netherlands: Springer, 1992.

[115] M. E. Nergiz and C. Clifton, "Thoughts on k-anonymization," *Data Knowl. Eng.*, vol. 63, no. 3, pp. 622–645, 2007.

[116] V. S. Iyengar, "Transforming data to satisfy privacy constraints," in *Proc. 8th ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining (KDD)*, 2002, pp. 279–288.

[117] B. C. M. Fung, K. Wang, and P. S. Yu, "Top-down specialization for information and privacy preservation," in *Proc. 21st Int. Conf. Data Eng. (ICDE)*, Apr. 2005, pp. 205–216.

[118] D. Agrawal and C. C. Aggarwal, "On the design and quantification of privacy preserving data mining algorithms," in *Proc. 20th ACM SIGMOD-SIGACT-SIGART Symp. Princ. Database Syst. (PODS)*, 2001, pp. 247–255.

[119] E. Hellinger, "Neue begründung der theorie quadratischer formen von unendlichvielen veränderlichen," *J. für die reine und Angewandte Math.*, vol. 1909, no. 136, pp. 210–271, 1909.

[120] M. Ye, X. Wu, X. Hu, and D. Hu, "Anonymizing classification data using rough set theory," *Knowl.-Based Syst.*, vol. 43, pp. 82–94, May 2013.
[121] D. Kifer and B.-R. Lin, "An axiomatic view of statistical privacy and utility," *J. Privacy Confidentiality*, vol. 4, no. 1, p. 2, 2012.
[122] B.-R. Lin and D. Kifer, "Information measures in statistical privacy and data processing applications," *ACM Trans. Knowl. Discovery Data*, vol. 9, no. 4, pp. 1–29, Jun. 2015.
[123] I. Wagner and D. Eckhoff, "Technical privacy metrics: A systematic survey," *ACM Comput. Surv.*, vol. 51, no. 3, p. 57, Jun. 2018.
[124] A. Campan, N. Cooper, and T. M. Truta, "On-the-fly generalization hierarchies for numerical attributes revisited," in *Secure Data Management (SDM)* (Lecture Notes in Computer Science), vol. 6933, W. Jonker and M. Petković, Eds. Berlin, Germany: Springer, 2011.
[125] O. Gkountouna, S. Angeli, A. Zigomitros, M. Terrovitis, and Y. Vassiliou, "k^m-Anonymity for continuous data using dynamic hierarchies," in *Privacy in Statistical Databases (PSD)* (Lecture Notes in Computer Science), vol. 8744, J. Domingo-Ferrer, Ed. Cham, Switzerland: Springer, 2014.
[126] S. Yaseen, S. M. A. Abbas, A. Anjum, T. Saba, A. Khan, S. U. R. Malik, N. Ahmad, B. Shahzad, and A. K. Bashir, "Improved generalization for secure data publishing," *IEEE Access*, vol. 6, pp. 27156–27165, 2018.
[127] K. LeFevre, D. J. DeWitt, and R. Ramakrishnan, "Incognito: Efficient full-domain k-anonymity," in *Proc. ACM SIGMOD Int. Conf. Manage. Data*, 2005, pp. 49–60.
[128] L. Sweeney, "Achieving k-anonymity privacy protection using generalization and suppression," *Int. J. Uncertainty, Fuzziness Knowl.-Based Syst.*, vol. 10, no. 5, pp. 571–588, Oct. 2002.
[129] F. Kohlmayer, F. Prasser, C. Eckert, A. Kemper, and K. A. Kuhn, "Flash: Efficient, stable and optimal K-Anonymity," in *Proc. Int. Conf. Privacy, Secur., Risk Trust Int. Conf. Social Comput.*, Sep. 2012, pp. 708–717.
[130] K. El Emam, F. K. Dankar, R. Issa, E. Jonker, D. Amyot, E. Cogo, J.-P. Corriveau, M. Walker, S. Chowdhury, R. Vaillancourt, T. Roffey, and J. Bottomley, "A globally optimal k-Anonymity method for the de-identification of health data," *J. Amer. Med. Inform. Assoc.*, vol. 16, no. 5, pp. 670–682, Sep. 2009.
[131] R. J. Bayardo and R. Agrawal, "Data privacy through optimal k-Anonymization," in *Proc. 21st Int. Conf. Data Eng. (ICDE)*, Apr. 2005, pp. 217–228.
[132] B. C. M. Fung, K. Wang, and P. S. Yu, "Anonymizing classification data for privacy preservation," *IEEE Trans. Knowl. Data Eng.*, vol. 19, no. 5, pp. 711–725, May 2007.
[133] J. Cao, P. Karras, C. Raïssi, and K.-L. Tan, "ρ-uncertainty: Inference-proof transaction anonymization," *Proc. VLDB Endowment*, vol. 3, nos. 1–2, pp. 1033–1044, 2010.
[134] K. LeFevre, D. J. DeWitt, and R. Ramakrishnan, "Mondrian multidimensional K-Anonymity," in *Proc. 22nd Int. Conf. Data Eng. (ICDE)*, 2006, p. 25.
[135] G. Ghinita, P. Karras, P. Kalnis, and N. Mamoulis, "Fast data anonymization with low information loss," in *Proc. 33rd Int. Conf. Very Large Data Bases*, 2007, pp. 758–769.
[136] V. Ayala-Rivera, P. McDonagh, T. Cerqueus, L. Murphy, and C. Thorpe, "Enhancing the utility of anonymized data by improving the quality of generalization hierarchies," *Trans. Data Privacy*, vol. 10, no. 1, pp. 27–59, 2017.
[137] K. Wang, B. C. Fung, and P. S. Yu, "Template-based privacy preservation in classification problems," in *Proc. 5th IEEE Int. Conf. Data Mining*, Nov. 2005, p. 8.
[138] X. Xiao and Y. Tao, "Anatomy: Simple and effective privacy preservation," in *Proc. 32nd Int. Conf. Very Large Data Bases*, 2006, pp. 139–150.
[139] T. Li, N. Li, J. Zhang, and I. Molloy, "Slicing: A new approach for privacy preserving data publishing," *IEEE Trans. Knowl. Data Eng.*, vol. 24, no. 3, pp. 561–574, Mar. 2012.
[140] M. Terrovitis, N. Mamoulis, J. Liagouris, and S. Skiadopoulos, "Privacy preservation by disassociation," *Proc. VLDB Endowment*, vol. 5, no. 10, pp. 944–955, Jun. 2012.
[141] S. De Capitani di Vimercati, S. Foresti, S. Jajodia, S. Paraboschi, and P. Samarati, "Fragments and loose associations: Respecting privacy in data publishing," *Proc. VLDB Endowment*, vol. 3, nos. 1–2, pp. 1370–1381, 2010.
[142] S. De Capitani di Vimercati, S. Foresti, S. Jajodia, G. Livraga, S. Paraboschi, and P. Samarati, "Extending loose associations to multiple fragments," in *Data and Applications Security and Privacy XXVII (DBSec)* (Lecture Notes in Computer Science), vol. 7964, L. Wang and B. Shafiq, Eds. Springer, 2013.
[143] S. De Capitani di Vimercati, S. Foresti, S. Jajodia, G. Livraga, S. Paraboschi, and P. Samarati, "Loose associations to increase utility in data publishing 1," *J. Comput. Secur.*, vol. 23, no. 1, pp. 59–88, Jan. 2015. [Online]. Available: http://dl.acm.org/citation.cfm?id=2746188.2746191
[144] Q. Zhang, N. Koudas, D. Srivastava, and T. Yu, "Aggregate query answering on anonymized tables," in *Proc. IEEE 23rd Int. Conf. Data Eng.*, Apr. 2007, pp. 116–125.
[145] K. Liu, C. Giannella, and H. Kargupta, "A survey of attack techniques on privacy-preserving data perturbation methods," in *Privacy-Preserving Data Mining* (Advances in Database Systems), vol. 34, C. C. Aggarwal and P. S. Yu, Eds. Boston, MA, USA: Springer, 2008.
[146] C. Dwork, F. McSherry, K. Nissim, and A. Smith, *Calibrating Noise to Sensitivity in Private Data Analysis*. Berlin, Germany: Springer, 2006, pp. 265–284.
[147] H. Polat and W. Du, "Privacy-preserving collaborative filtering using randomized perturbation techniques," in *Proc. 3rd IEEE Int. Conf. Data Mining*, Nov. 2003, pp. 625–628.
[148] C. Clifton, M. Kantarcioglu, J. Vaidya, X. Lin, and M. Y. Zhu, "Tools for privacy preserving distributed data mining," *ACM SIGKDD Explor. Newslett.*, vol. 4, no. 2, pp. 28–34, Dec. 2002.
[149] V. S. Verykios, A. K. Elmagarmid, E. Bertino, Y. Saygin, and E. Dasseni, "Association rule hiding," *IEEE Trans. Knowl. Data Eng.*, vol. 16, no. 4, pp. 434–447, Apr. 2004.
[150] A. Evfimievski, J. Gehrke, and R. Srikant, "Limiting privacy breaches in privacy preserving data mining," in *Proc. 22nd ACM SIGMOD-SIGACT-SIGART Symp. Princ. Database Syst. (PODS)*, 2003, pp. 211–222.
[151] M. E. Nergiz and M. Z. Gök, "Hybrid k-Anonymity," *Comput. Secur.*, vol. 44, pp. 51–63, Jul. 2014.
[152] A. S. M. T. Hasan, Q. Jiang, J. Luo, C. Li, and L. Chen, "An effective value swapping method for privacy preserving data publishing," *Secur. Commun. Netw.*, vol. 9, no. 16, pp. 3219–3228, Nov. 2016.
[153] M. Rodriguez-Garcia, M. Batet, and D. Sánchez, "Utility-preserving privacy protection of nominal data sets via semantic rank swapping," *Inf. Fusion*, vol. 45, pp. 282–295, Jan. 2019.
[154] D. B. Rubin, "Statistical disclosure limitation," *J. Off. Statist.*, vol. 9, no. 2, pp. 461–468, 1993.
[155] J. M. Mateo-Sanz, A. Martínez-Ballesté, and J. Domingo-Ferrer, "Fast generation of accurate synthetic microdata," in *Privacy in Statistical Databases (PSD)* (Lecture Notes in Computer Science), vol. 3050, V. Torra, Ed. Berlin, Germany: Springer, 2004.
[156] J. M. Abowd and J. Lane, "New approaches to confidentiality protection: Synthetic data, remote access and research data centers," in *Privacy in Statistical Databases (PSD)* (Lecture Notes in Computer Science), vol. 3050, J. Domingo-Ferrer and V. Torra, Eds. Berlin, Germany: Springer, 2004.
[157] V. Ayala-Rivera, A. O. Portillo-Dominguez, L. Murphy, and C. Thorpe, *COCOA: A Synthetic Data Generator for Testing Anonymization Techniques*. Cham, Switzerland: Springer, 2016, pp. 163–177, doi: 10.1007/978-3-319-45381-1_13.
[158] V. Bindschaedler, R. Shokri, and C. A. Gunter, "Plausible deniability for privacy-preserving data synthesis," *Proc. VLDB Endowment*, vol. 10, no. 5, pp. 481–492, Jan. 2017.
[159] N. C. Abay, Y. Zhou, M. Kantarcioglu, B. Thuraisingham, and L. Sweeney, "Privacy preserving synthetic data release using deep learning," in *Machine Learning and Knowledge Discovery in Databases (ECML PKDD)* (Lecture Notes in Computer Science), vol. 11051, M. Berlingerio, F. Bonchi, T. Gärtner, N. Hurley, and G. Ifrim, Eds. Cham, Switzerland: Springer, 2018.
[160] J. P. Reiter, "Inference for partially synthetic, public use microdata sets," *Surv. Methodol.*, vol. 29, no. 2, pp. 181–188, 2003.
[161] J. Domingo-Ferrer and Ú. González-Nicolás, "Hybrid microdata using microaggregation," *Inf. Sci.*, vol. 180, no. 15, pp. 2834–2844, Aug. 2010.
[162] K. Muralidhar and R. Sarathy, "Generating sufficiency-based non-synthetic perturbed data," *Trans. Data Privacy*, vol. 1, no. 1, pp. 17–33, 2008.
[163] A. Oganian and J. Domingo-Ferrer, "Hybrid microdata via model-based clustering," in *Privacy in Statistical Databases (PSD)* (Lecture Notes in Computer Science), vol. 7556, J. Domingo-Ferrer and I. Tinnirello, Eds. Berlin, Germany: Springer, 2012.
[164] J. Domingo-Ferrer and J. M. Mateo-Sanz, "Practical data-oriented microaggregation for statistical disclosure control," *IEEE Trans. Knowl. Data Eng.*, vol. 14, no. 1, pp. 189–201, Jan./Feb. 2002.
[165] V. Torra, "Microaggregation for categorical variables: A median based approach," in *Privacy in Statistical Databases (PSD)* (Lecture Notes in Computer Science), vol. 3050, J. Domingo-Ferrer and V. Torra, Eds. Berlin, Germany: Springer, 2004.

[166] V. Torra, "Constrained microaggregation: Adding constraints for data editing," *Trans. Data Privacy*, vol. 1, no. 2, pp. 86–104, 2008.
[167] I. Cano, G. Navarro-Arribas, and V. Torra, "A new framework to automate constrained microaggregation," in *Proc. Proc. ACM 1st Int. Workshop Privacy Anonymity Very Large Databases (PAVLAD)*, 2009, pp. 1–8.
[168] A. Solanas and A. Martinez-Balleste, "V-MDAV: A multivariate microaggregation with variable group size," in *17th COMPSTAT Symp. (IASC)*, Rome, Italy, 2006, pp. 1–8.
[169] F. Casino, C. Patsakis, and A. Solanas, "Privacy-preserving collaborative filtering: A new approach based on variable-group-size microaggregation," *Electron. Commerce Res. Appl.*, vol. 38, Nov./Dec. 2019, Art. no. 100895.
[170] T. Li and N. Li, "On the tradeoff between privacy and utility in data publishing," in *Proc. 15th ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining (KDD)*, 2009, pp. 517–526.
[171] L. Sankar, S. R. Rajagopalan, and H. V. Poor, "Utility-privacy tradeoffs in databases: An information-theoretic approach," *IEEE Trans. Inf. Forensics Secur.*, vol. 8, no. 6, pp. 838–852, Jun. 2013.
[172] A. Machanavajjhala, D. Kifer, J. Gehrke, and M. Venkitasubramaniam, "*L*-diversity: Privacy beyond *k*-anonymity," *ACM Trans. Knowl. Discovery Data*, vol. 1, no. 1, p. 3, 2007.
[173] M. E. Nergiz, M. Atzori, and C. Clifton, "Hiding the presence of individuals from shared databases," in *Proc. ACM SIGMOD Int. Conf. Manage. Data (SIGMOD)*, 2007, pp. 665–676.
[174] M. E. Nergiz and C. Clifton, "δ-presence without complete world knowledge," *IEEE Trans. Knowl. Data Eng.*, vol. 22, no. 6, pp. 868–883, Jun. 2010.
[175] P. Samarati and L. Sweeney, "Protecting privacy when disclosing information: K-anonymity and its enforcement through generalization and suppression," SRI Int., Menlo Park, CA, USA, Tech. Rep. CSL-98-04, 1998.
[176] C. C. Aggarwal, "On k-anonymity and the curse of dimensionality," in *Proc. 31st Int. Conf. Very Large Data Bases*, 2005, pp. 901–909.
[177] R. Dewri, I. Ray, I. Ray, and D. Whitley, "On the optimal selection of k in the k-Anonymity problem," in *Proc. IEEE 24th Int. Conf. Data Eng.*, Apr. 2008, pp. 1364–1366.
[178] J. Domingo-Ferrer and V. Torra, "Ordinal, continuous and heterogeneous k-Anonymity through microaggregation," *Data Mining Knowl. Discovery*, vol. 11, no. 2, pp. 195–212, Sep. 2005.
[179] J. Domingo-Ferrer, A. Solanas, and A. Martinez-Balleste, "Privacy in statistical databases: K-Anonymity through microaggregation," in *Proc. IEEE Int. Conf. Granular Comput.*, 2006, pp. 774–777.
[180] J. Domingo-Ferrer, "Microaggregation: Achieving k-anonymity with quasi-optimal data quality," in *Proc. Eur. Conf. Qual. Survey Statist.*, Cardiff, U.K., 2006.
[181] A. Gionis, A. Mazza, and T. Tassa, "K-anonymization revisited," in *Proc. IEEE 24th Int. Conf. Data Eng.*, Apr. 2008, pp. 744–753.
[182] W. K. Wong, N. Mamoulis, and D. W. L. Cheung, "Non-homogeneous generalization in privacy preserving data publishing," in *Proc. Int. Conf. Manage. Data (SIGMOD)*, 2010, pp. 747–758.
[183] T. Tassa, A. Mazza, and A. Gionis, "K-concealment: An alternative model of k-type anonymity," *Trans. Data Privacy*, vol. 5, no. 1, pp. 189–222, 2012.
[184] K. Stokes and V. Torra, "n-Confusion: A generalization of k-anonymity," in *Proc. Joint EDBT/ICDT Workshops*, 2012, pp. 211–215.
[185] M. E. Nergiz, C. Clifton, and A. E. Nergiz, "Multirelational k-Anonymity," *IEEE Trans. Knowl. Data Eng.*, vol. 21, no. 8, pp. 1104–1117, Aug. 2009.
[186] M. Terrovitis, N. Mamoulis, and P. Kalnis, "Privacy-preserving anonymization of set-valued data," *Proc. VLDB Endowment*, vol. 1, no. 1, pp. 115–125, Aug. 2008.
[187] K. Wang, B. C. M. Fung, and P. S. Yu, "Handicapping attacker's confidence: An alternative to k-anonymization," *Knowl. Inf. Syst.*, vol. 11, no. 3, pp. 345–368, Apr. 2007.
[188] T. M. Truta and B. Vinay, "Privacy protection: P-Sensitive k-Anonymity property," in *Proc. 22nd Int. Conf. Data Eng. Workshops (ICDEW)*, 2006, p. 94.
[189] X. Sun, L. Sun, and H. Wang, "Extended k-anonymity models against sensitive attribute disclosure," *Comput. Commun.*, vol. 34, no. 4, pp. 526–535, Apr. 2011.
[190] A. Zigomitros, A. Solanas, and C. Patsakis, "The role of inference in the anonymization of medical records," in *Proc. IEEE 27th Int. Symp. Comput.-Based Med. Syst.*, May 2014, pp. 83–93.
[191] D. J. Martin, D. Kifer, A. Machanavajjhala, J. Gehrke, and J. Y. Halpern, "Worst-case background knowledge for privacy-preserving data publishing," in *Proc. IEEE 23rd Int. Conf. Data Eng.*, Apr. 2007, pp. 126–135.
[192] J. Liu and K. Wang, "On optimal anonymization for l^+-diversity," in *Proc. IEEE 26th Int. Conf. Data Eng.*, Mar. 2010, pp. 213–224.
[193] R. C.-W. Wong, J. Li, A. W.-C. Fu, and K. Wang, "(α, k)-anonymity: An enhanced k-anonymity model for privacy preserving data publishing," in *Proc. 12th ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*, 2006, pp. 754–759.
[194] J. Li, Y. Tao, and X. Xiao, "Preservation of proximity privacy in publishing numerical sensitive data," in *Proc. ACM SIGMOD Int. Conf. Manage. Data (SIGMOD)*, 2008, pp. 473–486.
[195] G. Loukides and J. Shao, "Preventing range disclosure in *k*-anonymised data," *Expert Syst. Appl.*, vol. 38, no. 4, pp. 4559–4574, 2011.
[196] N. Li, T. Li, and S. Venkatasubramanian, "T-closeness: Privacy beyond k-Anonymity and l-Diversity," in *Proc. IEEE 23rd Int. Conf. Data Eng.*, Apr. 2007, pp. 106–115.
[197] N. Li, T. Li, and S. Venkatasubramanian, "Closeness: A new privacy measure for data publishing," *IEEE Trans. Knowl. Data Eng.*, vol. 22, no. 7, pp. 943–956, Jul. 2010.
[198] Y. Rubner, C. Tomasi, and L. J. Guibas, "The earth mover's distance as a metric for image retrieval," *Int. J. Comput. Vis.*, vol. 40, no. 2, pp. 99–121, Nov. 2000.
[199] H. Liang and H. Yuan, "On the complexity of *t*-closeness anonymization and related problems," in *Database Systems for Advanced Applications (DASFAA)* (Lecture Notes in Computer Science), vol. 7825, W. Meng, L. Feng, S. Bressan, W. Winiwarter, and W. Song, Eds. Berlin, Germany: Springer, 2013.
[200] J. Cao, P. Karras, P. Kalnis, and K.-L. Tan, "SABRE: A sensitive attribute bucketization and REdistribution framework for t-closeness," *VLDB J.*, vol. 20, no. 1, pp. 59–81, Feb. 2011.
[201] J. Soria-Comas, J. Domingo-Ferrer, D. Sanchez, and S. Martinez, "T-closeness through microaggregation: Strict privacy with enhanced utility preservation," *IEEE Trans. Knowl. Data Eng.*, vol. 27, no. 11, pp. 3098–3110, Nov. 2015.
[202] J. Cao and P. Karras, "Publishing microdata with a robust privacy guarantee," *Proc. VLDB Endowment*, vol. 5, no. 11, pp. 1388–1399, Jul. 2012.
[203] R. C.-W. Wong, A. W.-C. Fu, K. Wang, and J. Pei, "Minimality attack in privacy preserving data publishing," in *Proc. 33rd Int. Conf. Very Large Data Bases*, 2007, pp. 543–554.
[204] C. Dwork, "Differential privacy," in *Automata, Languages and Programming (ICALP)* (Lecture Notes in Computer Science), vol. 4052, M. Bugliesi, B. Preneel, V. Sassone, and I. Wegener, Eds. Springer, 2006.
[205] R. Chen, B. Fung, B. C. Desai, and N. M. Sossou, "Differentially private transit data publication: A case study on the montreal transportation system," in *Proc. 18th ACM SIGKDD Int. Conf. Knowl. Discovery Data*, 2012, pp. 213–221.
[206] C. Dwork and A. Roth, "The algorithmic foundations of differential privacy," *Found. Trends Theor. Comput. Sci.*, vol. 9, nos. 3–4, pp. 211–407, 2013.
[207] C. Dwork, "Differential privacy: A survey of results," in *Theory and Applications of Models of Computation (TAMC)* (Lecture Notes in Computer Science), vol. 4978, M. Agrawal, D. Du, Z. Duan, and A. Li, Eds. Berlin, Germany: Springer, 2008.
[208] N. Li, W. Qardaji, and D. Su, "On sampling, anonymization, and differential privacy or, *k*-anonymization meets differential privacy," in *Proc. 7th ACM Symp. Inf., Comput. Commun. Secur.*, 2012, pp. 32–33.
[209] N. Johnson, J. P. Near, and D. Song, "Towards practical differential privacy for SQL queries," *Proc. VLDB Endowment*, vol. 11, no. 5, pp. 526–539, Jan. 2018.
[210] I. Dinur and K. Nissim, "Revealing information while preserving privacy," in *Proc. 22nd ACM SIGMOD-SIGACT-SIGART Symp. Princ. Database Syst. (PODS)*, 2003, pp. 202–210.
[211] A. Blum, C. Dwork, F. McSherry, and K. Nissim, "Practical privacy: The SuLQ framework," in *Proc. 24th ACM SIGMOD-SIGACT-SIGART Symp. Princ. Database Syst.*, Baltimore, MD, USA, Jun. 2005. [Online]. Available: https://www.microsoft.com/en-us/research/publication/practical-privacy-the-sulq-framework/
[212] C. Dwork, M. Naor, O. Reingold, G. N. Rothblum, and S. Vadhan, "On the complexity of differentially private data release: Efficient algorithms and hardness results," in *Proc. 41st Annu. ACM Symp. Theory Comput.*, 2009, pp. 381–390.
[213] M. Hardt, K. Ligett, and F. McSherry, "A simple and practical algorithm for differentially private data release," in *Proc. NIPS*, 2012, pp. 2348–2356.
[214] A. Blum, K. Ligett, and A. Roth, "A learning theory approach to noninteractive database privacy," *J. ACM*, vol. 60, no. 2, pp. 1–25, Apr. 2013.

[215] R. Chen, N. Mohammed, B. C. M. Fung, B. C. Desai, and L. Xiong, "Publishing set-valued data via differential privacy," *Proc. VLDB Endowment*, vol. 4, no. 11, pp. 1087–1098, 2011.
[216] J. Soria-Comas, J. Domingo-Ferrer, D. Sánchez, and S. Martínez, "Enhancing data utility in differential privacy via microaggregation-based *k*-anonymity," *VLDB J.*, vol. 23, pp. 771–794, Feb. 2014.
[217] D. Su, J. Cao, N. Li, and M. Lyu, "PrivPfC: Differentially private data publication for classification," *VLDB J.*, vol. 27, no. 2, pp. 201–223, Apr. 2018.
[218] X. Cheng, P. Tang, S. Su, R. Chen, Z. Wu, and B. Zhu, "Multi-party high-dimensional data publishing under differential privacy," *IEEE Trans. Knowl. Data Eng.*, to be published.
[219] C. Piao, Y. Shi, J. Yan, C. Zhang, and L. Liu, "Privacy-preserving governmental data publishing: A fog-computing-based differential privacy approach," *Future Gener. Comput. Syst.*, vol. 90, pp. 158–174, Jan. 2019.
[220] A. Machanavajjhala, D. Kifer, J. Abowd, J. Gehrke, and L. Vilhuber, "Privacy: Theory meets practice on the map," in *Proc. IEEE 24th Int. Conf. Data Eng.*, Washington, DC, USA, Apr. 2008, pp. 277–286, doi: 10.1109/ICDE.2008.4497436.
[221] A. Korolova, K. Kenthapadi, N. Mishra, and A. Ntoulas, "Releasing search queries and clicks privately," in *Proc. 18th Int. Conf. World Wide Web (WWW)*, New York, NY, USA, 2009, pp. 171–180, doi: 10.1145/1526709.1526733.
[222] M. Gotz, A. Machanavajjhala, G. Wang, X. Xiao, and J. Gehrke, "Publishing search logs—A comparative study of privacy guarantees," *IEEE Trans. Knowl. Data Eng.*, vol. 24, no. 3, pp. 520–532, Mar. 2012.
[223] D. Leoni, "Non-interactive differential privacy: A survey," in *Proc. 1st Int. Workshop Open Data*, 2012, pp. 40–52.
[224] C. Dwork, F. McSherry, K. Nissim, and A. Smith, "Differential privacy—A primer for the perplexed," *Proc. Joint UNECE/Eurostat Work Session Stat. Data Confidentiality*, vol. 11, 2011, pp. 1–8.
[225] C. Clifton and T. Tassa, "On syntactic anonymity and differential privacy," in *Proc. IEEE 29th Int. Conf. Data Eng. Workshops (ICDEW)*, Apr. 2013, pp. 161–183.
[226] Ú. Erlingsson, V. Pihur, and A. Korolova, "RAPPOR: Randomized aggregatable privacy-preserving ordinal response," 2014, *arXiv:1407.6981*. [Online]. Available: http://arxiv.org/abs/1407.6981
[227] G. Cormode, S. Jha, T. Kulkarni, N. Li, D. Srivastava, and T. Wang, "Privacy at scale: Local differential privacy in practice," in *Proc. Int. Conf. Manage. Data*, New York, NY, USA, 2018, pp. 1655–1658.
[228] Z. Qin, Y. Yang, T. Yu, I. Khalil, X. Xiao, and K. Ren, "Heavy hitter estimation over set-valued data with local differential privacy," in *Proc. ACM SIGSAC Conf. Comput. Commun. Secur. (CCS)*, 2016, pp. 192–203.
[229] A. Gupta, M. Hardt, A. Roth, and J. Ullman, "Privately releasing conjunctions and the statistical query barrier," in *Proc. 43rd Annu. ACM Symp. Theory Comput.*, New York, NY, USA, 2011, pp. 803–812.
[230] N. Wang, X. Xiao, Y. Yang, J. Zhao, S. C. Hui, H. Shin, J. Shin, and G. Yu, "Collecting and analyzing multidimensional data with local differential privacy," in *Proc. IEEE 35th Int. Conf. Data Eng. (ICDE)*, Apr. 2019, pp. 638–649.
[231] X. Ren, C. Yu, W. Yu, S. Yang, X. Yang, J. A. McCann, and P. S. Yu, "*LoPub*: High-dimensional crowdsourced data publication with local differential privacy," *IEEE Trans. Inf. Forensics Security*, vol. 13, no. 9, pp. 2151–2166, Sep. 2018.
[232] W.-Y. Day and N. Li, "Differentially private publishing of high-dimensional data using sensitivity control," in *Proc. 10th ACM Symp. Inf., Comput. Commun. Secur. (ASIA CCS)*, 2015, pp. 451–462.
[233] R. Shokri, "Privacy games: Optimal user-centric data obfuscation," *Proc. Privacy Enhancing Technol.*, vol. 2015, no. 2, pp. 299–315, 2015.
[234] Q. Geng and P. Viswanath, "Optimal noise adding mechanisms for approximate differential privacy," *IEEE Trans. Inf. Theory*, vol. 62, no. 2, pp. 952–969, Feb. 2016.
[235] P. Jain, M. Gyanchandani, and N. Khare, "Big data privacy: A technological perspective and review," *J. Big Data*, vol. 3, no. 1, p. 25, Dec. 2016.
[236] A. Sharma, G. Singh, and S. Rehman, "A review of big data challenges and preserving privacy in big data," in *Advances in Data and Information Sciences*, M. L. Kolhe, S. Tiwari, M. C. Trivedi, and K. K. Mishra, Eds. Singapore: Springer, 2020, pp. 57–65.
[237] J. Pease and J. Freudiger. (2016). *Engineering Privacy for Your Users*. [Online]. Available: http://devstreaming.apple.com/videos/wwdc/2016/709tvxadw201avg5v7n/709/709_engineering_privacy_for_your_users.pdf
[238] C. Lin, Z. Song, H. Song, Y. Zhou, Y. Wang, and G. Wu, "Differential privacy preserving in big data analytics for connected health," *J. Med. Syst.*, vol. 40, no. 4, pp. 1–9, Apr. 2016.
[239] L. Fan and H. Jin, "A practical framework for privacy-preserving data analytics," in *Proc. 24th Int. Conf. World Wide Web (WWW)*, 2015, pp. 311–321.
[240] G. Cormode, "Personal privacy vs population privacy: Learning to attack anonymization," in *Proc. 17th ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*, 2011, pp. 1253–1261.
[241] D. Riboni and C. Bettini, "Incremental release of differentially-private check-in data," *Pervas. Mobile Comput.*, vol. 16, pp. 220–238, Jan. 2015, doi: 10.1016/j.pmcj.2014.11.007.
[242] X. Zhang, X. Meng, and R. Chen, *Differentially Private Set-Valued Data Release against Incremental Updates*. Berlin, Germany: Springer, 2013, pp. 392–406.
[243] B. Srisungsittisunti and J. Natwichai, "An incremental privacy-preservation algorithm for the (k, e)-Anonymous model," *Comput. Electr. Eng.*, vol. 41, pp. 126–141, Jan. 2015.
[244] Z. Ji, Z. C. Lipton, and C. Elkan, "Differential privacy and machine learning: A survey and review," 2014, *arXiv:1412.7584*. [Online]. Available: https://arxiv.org/abs/1412.7584
[245] G. Ghinita, Y. Tao, and P. Kalnis, "On the anonymization of sparse high-dimensional data," in *Proc. IEEE 24th Int. Conf. Data Eng.*, Apr. 2008, pp. 715–724.
[246] F. Casino, C. Patsakis, E. Batista, F. Borras, and A. Martinez-Balleste, "Healthy routes in the smart city: A context-aware mobile recommender," *IEEE Softw.*, vol. 34, no. 6, pp. 42–47, Nov. 2017.
[247] K. Sehgal, A. Jain, P. Nagrath, and A. Kumar, "Recent advances in networks and data security survey on various mobile operating systems," in *Proc. Int. Conf. Innov. Comput. Commun.* Springer, 2019, pp. 181–190.
[248] E. Politou, E. Alepis, and C. Patsakis, "A survey on mobile affective computing," *Comput. Sci. Rev.*, vol. 25, pp. 79–100, Aug. 2017.
[249] F. Casino, C. Patsakis, D. Puig, and A. Solanas, "On privacy preserving collaborative filtering: Current trends, open problems, and new issues," in *Proc. IEEE 10th Int. Conf. E-Bus. Eng.*, Sep. 2013, pp. 244–249.
[250] M. Sookhak, H. Tang, Y. He, and F. R. Yu, "Security and privacy of smart cities: A survey, research issues and challenges," *IEEE Commun. Surveys Tutr.*, vol. 21, no. 2, pp. 1718–1743, 2nd Quart., 2019.
[251] T. Braun, B. C. M. Fung, F. Iqbal, and B. Shah, "Security and privacy challenges in smart cities," *Sustain. Cities Soc.*, vol. 39, pp. 499–507, May 2018.
[252] M. Ammar, G. Russello, and B. Crispo, "Internet of Things: A survey on the security of IoT frameworks," *J. Inf. Secur. Appl.*, vol. 38, pp. 8–27, Feb. 2018. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S2214212617302934
[253] F. Casino and C. Patsakis, "An efficient blockchain-based privacy-preserving collaborative filtering architecture," *IEEE Trans. Eng. Manag.*, to be published.
[254] N. Mohammed, B. C. M. Fung, P. C. K. Hung, and C.-K. Lee, "Centralized and distributed anonymization for high-dimensional healthcare data," *ACM Trans. Knowl. Discovery Data*, vol. 4, no. 4, pp. 1–33, Oct. 2010.
[255] B. C. M. Fung, T. Trojer, P. C. K. Hung, L. Xiong, K. Al-Hussaeni, and R. Dssouli, "Service-oriented architecture for high-dimensional private data mashup," *IEEE Trans. Services Comput.*, vol. 5, no. 3, pp. 373–386, 3rd Quart., 2012.
[256] W. Wang, L. Chen, and Q. Zhang, "Outsourcing high-dimensional healthcare data to cloud with personalized privacy preservation," *Comput. Netw.*, vol. 88, pp. 136–148, Sep. 2015.
[257] F. Prasser, R. Bild, J. Eicher, H. Spengler, F. Kohlmayer, and K. A. Kuhn, "Lightning: Utility-driven anonymization of high-dimensional data," *Trans. Data Privacy*, vol. 9, no. 2, pp. 161–185, Aug. 2016. [Online]. Available: http://dl.acm.org/citation.cfm?id=2993206.2993209
[258] H. Zakerzadeh, C. C. Aggrawal, and K. Barker, "Towards breaking the curse of dimensionality for high-dimensional privacy: An extended version," 2014, *arXiv:1401.1174*. [Online]. Available: http://arxiv.org/abs/1401.1174
[259] H. Zakerzadeh, C. C. Aggarwal, and K. Barker, "Managing dimensionality in data privacy anonymization," *Knowl. Inf. Syst.*, vol. 49, no. 1, pp. 341–373, Oct. 2016.
[260] L. Chen, W.-K. Lee, C.-C. Chang, K.-K.-R. Choo, and N. Zhang, "Blockchain based searchable encryption for electronic health record sharing," *Future Gener. Comput. Syst.*, vol. 95, pp. 420–429, Jun. 2019.
[261] S. Ramaswamy, R. Rastogi, and K. Shim, "Efficient algorithms for mining outliers from large data sets," *ACM SIGMOD Rec.*, vol. 29, no. 2, pp. 427–438, Jun. 2000.
[262] P. Das and D. Mandal, *Statistical Outlier Detection in Large Multivariate Datasets*. pp. 1–9. [Online]. Available: http://acsu.buffalo.edu

[263] B. Viswanath, M. A. Bashir, M. Crovella, S. Guha, K. P. Gummadi, B. Krishnamurthy, and A. Mislove, "Towards detecting anomalous user behavior in online social networks," in *Proc. 23rd USENIX Secur. Symp.*, 2014, pp. 223–238.
[264] G. Singh and V. Kumar, "An efficient clustering and distance based approach for outlier detection," *Int. J. Comput. Trends Technol.*, vol. 4, no. 7, pp. 2067–2072, 2013.
[265] H. Wang and R. Liu, "Hiding outliers into crowd: Privacy-preserving data publishing with outliers," *Data Knowl. Eng.*, vol. 100, pp. 94–115, Nov. 2015.
[266] A. Ayadi, O. Ghorbel, A. M. Obeid, and M. Abid, "Outlier detection approaches for wireless sensor networks: A survey," *Comput. Netw.*, vol. 129, pp. 319–333, Dec. 2017.
[267] M. Goldstein and S. Uchida, "A comparative evaluation of unsupervised anomaly detection algorithms for multivariate data," *PLoS ONE*, vol. 11, no. 4, 2016, Art. no. e0152173.
[268] M. Templ, K. Hron, and P. Filzmoser, "Exploratory tools for outlier detection in compositional data with structural zeros," *J. Appl. Statist.*, vol. 44, no. 4, pp. 734–752, Mar. 2017.
[269] A. Pfitzmann and M. Hansen, "A terminology for talking about privacy by data minimization: Anonymity, unlinkability, undetectability, unobservability, pseudonymity, and identity management," Version 0.34, Tech. Rep., Aug. 2010.
[270] B. Gilburd, A. Schuster, and R. Wolff, "k-TTP: A new privacy model for large-scale distributed environments," in *Proc. 10th ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*, New York, NY, USA, 2004, pp. 563–568.
[271] A. Gkoulalas-Divanis, G. Loukides, and J. Sun, "Publishing data from electronic health records while preserving privacy: A survey of algorithms," *J. Biomed. Informat.*, vol. 50, pp. 4–19, Aug. 2014.
[272] L. Taylor, L. Floridi, and B. Van der Sloot, Eds., *Group Privacy: New Challenges of Data Technologies*, vol. 126. Springer, 2016.
[273] A. W.-C. Fu, K. Wang, R. C.-W. Wong, J. Wang, and M. Jiang, "Small sum privacy and large sum utility in data publishing," *J. Biomed. Informat.*, vol. 50, pp. 20–31, Aug. 2014.

ATHANASIOS ZIGOMITROS received the first B.Sc. degree in business planning and information systems from the Technological Educational Institute (T.E.I.) of Patras, and the second B.Sc. degree in informatics, the M.Sc. degree in advanced computer systems, and the Ph.D. degree in data anonymization from the University of Piraeus. He received two independent scholarships for his research and participated in several European and Greek R&D projects. His research interests include multimedia retrieval and metadata, privacy preserving data publishing, anonymity, and watermarking and digital currencies.

FRAN CASINO (Member, IEEE) was born in Tarragona, in 1986. He received the B.Sc. degree in computer science and the M.Sc. degree in computer security and intelligent from Rovira i Virgili University, Tarragona, Spain, in 2010 and 2013, respectively, and the Ph.D. degree *(cum laude)* in computer science from the Rovira i Virgili University, in 2017, with honors (Best Dissertation Award). He was a Visiting Researcher with ISCTE-IUL, Lisbon, in 2016. He is currently a Postdoctoral Researcher with the Department of Informatics, Piraeus University, Piraeus, Greece. He has participated in several European-, Spanish- and Catalan-funded research projects. His research focuses on pattern recognition, and data management applied to different fields such as privacy and security protection, recommender systems, smart health, and blockchain.

AGUSTI SOLANAS (Senior Member, IEEE) received the M.Sc. degree in computer engineering from URV, Spain, in 2004, with honors (Outstanding Graduation Award), and the Ph.D. degree *(cum laude)* in telematics engineering from the Technical University of Catalonia, in 2007. He was a Visiting Researcher from the University of Roma Tre, Italy, in 2012, and the University of Padua, Italy, in 2013. He is currently an Associate Professor with the Department of Computer Engineering and Mathematics, Rovira i Virgili University (URV). He serves as a Scientific Coordinator at APWG.EU. His fields of activity include data privacy, ubiquitous computing, and artificial intelligence. He has participated in several European- and National-funded research projects. He has authored over 150 publications and he has delivered several invited talks worldwide. He serves as an External Expert Reviewer for several National Councils for Scientific Research. He is also the Vice-President of the Computer Society Spain Chapter.

CONSTANTINOS PATSAKIS (Member, IEEE) received the B.Sc. degree in mathematics from the University of Athens, Greece, the M.Sc. degree in information security from the Royal Holloway, University of London, and the Ph.D. degree in cryptography and malware from the Department of Informatics, University of Piraeus. He was a Researcher with the UNESCO Chair in Data Privacy and a Research Fellow with the Trinity College. He is currently an Assistant Professor with the University of Piraeus and an Adjunct Researcher with the Athena Research and Innovation Center. He has authored more than 100 publications in peer-reviewed international conferences and journals. He has participated in several national and European R&D projects. His main areas of research include cryptography, security, privacy, data anonymization, and data mining.