Journal of Theoretical and Applied Information Technology
15th November 2022. Vol.100. No 21

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

CLASSIFYING AND EVALUATING PRIVACY-PRESERVING TECHNIQUES BASED ON PROTECTION METHODS: A COMPREHENSIVE STUDY

SHERMINA JEBA [1], MOHAMMED BINJUBIER [2, **], MOHD ARFIAN ISMAIL [2,*], RESHMY KRISHNAN [1], SARACHANDRAN NAIR [1] , and GIRIJA NARASIMHAN [3]

[1] Department Of Computing, Muscat College, Muscat, Oman
[2] Faculty Of Computing, Universiti Malaysia Pahang, Kuantan, Pahang, Malaysia
[3] Information Technology Department, University Of Technology And Applied Sciences, Muscat, Oman

E-mail: *arfian@ump.edu.my, **moh77421143@gmail.com

ABSTRACT

Many data analysis applications encounter the challenge of preserving the privacy of information. Over the past few years, many partially published data have become subjects of various concerns, ranging from unlawful access to private data to privacy breaches and unintended use of personal information. This problem has limited progress in advancing published data, prompting the need for robust privacy-protection techniques, which can minimize the chances of identifying sensitive individual information by unauthorized persons. The simplest solution to preserving sensitive information is to avoid public disclosure of such information. However, this might constitute a problem for data analysis, as there may not be available datasets to analyze and discover interesting patterns. Sometimes, the dataset must be disclosed under government regulations to enable access and subsequent analysis. Sometimes, the data owner may modify the data to ensure privacy and retain sufficient information for a safe release to the public. This process is usually referred to as privacy-preserving data publishing (PPDP). The review in this paper has rigorously evaluated some existing preserving privacy techniques and classified them based on their methods to reduce the risk of disclosing information. Moreover, the review focused on the methods of the current preserving privacy techniques to protect data and preserve the privacy of sensitive information, which is considered a key contribution of this study as it is expected to guide scholars to gain a deeper knowledge of the existing privacy preservation methods. This study also compared and analyzed various privacy-preserving techniques in terms of their advantages and drawbacks.

Keywords: *Big Data Privacy Preservation; Anonymization; Data Publishing.*

1. INTRODUCTION

The volume of big data generation has increased in the past few years because of the emergence of advanced computing techniques, which has given birth to the era of "big data" that significantly transformed modern societies. Many information science scholars have paid significant attention to this aspect, making big data an emerging and exciting field of Information Technology (IT). Big data has been considered the "fourth paradigm" of scientific discoveries in the scientific community [1]. With the increasing popularity of the big data concept, it has become necessary to have a perfect understanding of the big data concept and the challenges associated with its analysis to make substantial effects in this pursuit [1] [2]. As per several researchers, big data is a digital era revolution, which can be described as the era of "the new oil," considering its associated relevance to society [3]. Big data has, therefore, emerged as a fertile ground for gaining a competitive industrial advantage. Still, big data analysis remains a frontier for knowledge and innovation advancements and a better decision-making process.

The idea of big data is yet to be comprehensively understood in the IT and business world; it is rather diverse owing to its rapid evolution. Hence, reaching an acceptable definition of the big data concept has

become difficult. In this regard, several organizations have tried to define big data. For instance, big data was defined by Gartner [4] as "high-volume, high-velocity, and high-variety information assets, which are required to be processed in a low-cost and innovative manner to improve knowledge and decision making." The first aspect of this definition captured the three essential features of big data as introduced by Laney even before the popularity of the big data concept; these characteristics are called the "3Vs" of big data, which are the volume, the velocity, and the variety [4] [5] [2]. The second part of the definition captured the significance of the relationship between costs and outcomes of novel technological capabilities. The last part of the definition portrayed the ultimate goal of value creation through data processing [6].

Recent increases in personal computer usage and public demand for data have caused most businesses regularly publish data in various formats to facilitate access to big data's information content and provide many opportunities with significant benefits in various fields [7] [2]. This information access can aid in improving the performance and efficiency of the organizations and help establish and develop their plans [8] [9] [10] [11]. Data publication is the easiest method of data sharing, which helps various research entities run data analytic operations on published databases to extract knowledge from published data. Such knowledge can represent, interpret, or discover interesting patterns [12] [13]. However, making the gained raw data usable in decision-making and prediction is yet to be realized as scholars face several problems during knowledge extraction from the published data. These challenges are classified into two groups [4]: challenges associated with data and challenges associated with data mining operations.

2.1 Challenges Associated with Data

The increasing volume of collected data can cause storage and management problems. Therefore, it is necessary to process raw data from where heterogeneity of data via data preprocessing (i.e., to remove irrelevant and redundant features), data integration (i.e., to combine/integrate data sourced from different sources), and data conversion into suitable formats for easy processing. Furthermore, secure data collection and propagation must be ensured by denying unauthorized access to unauthorized parties and ensuring non-disclosure of private information [13]. In the past, data were published in statistical databases known as tabular form (macrodata). Nowadays, the need to use and publish specific stored data (microdata) has been rapidly growing. Macrodata consists of the results of precomputed statistics that are typically presented in 2D tables, whereas microdata consists of the actual information. The advantage of microdata over macrodata is the ability to conduct any analysis that may not be possible with microdata. However, the release of microdata may compromise the privacy of companies or individuals whose information is included in the released microdata [14].

There are two fundamental methods for releasing published data in both forms. The first method is a multiple publication model from the same data publisher, which refers to a series of datasets in distinct timestamps that are all extensions in certain aspects (e.g., quarterly released data) [13] [15] [16] [17]. The second method is a single publication model from several data publishers, which refers to anonymized data that are published without considering other published datasets [18] [19] [20] [21]. The issue with this assumption (single publication model) is that, sometimes, the information of an individual may be published by more than one organization [22]. Here, an attacker may launch a composition attack [23] [21] on these published datasets to alter the privacy of the dataset.

2.2 Challenges associated with data mining operations

Data publishing helps many research establishments run big data analytic operations to reveal information embedded therein. In addition, data publishing provides several opportunities with great unprecedented benefits in many fields [7] [2], which can help enhance the efficiency of organizations and support their plans for the future [8] [9] [10] [11]. Consequently, the challenges associated with data mining operations pertain to preventing the disclosure of private information about individuals by utilizing various technologies to extract useful information (knowledge) from data. This information can represent, interpret, or discover exciting patterns. Protecting sensitive data and extracting new knowledge is a significant competitive advantage. Most organizations now rely on data analysis to survive. Adopting big data technology in different fields may not be a luxury but remains the life wire of many organizations in their bid to gain competitiveness [14].

Similarly, a great deal of attention has been paid to potential data privacy violations and data misuse; consequently, the proper protection of released data must be ensured, as failure may cause situations that are detrimental to individuals and

organizations [12] [8]. For this reason, many establishments encounter a challenging dilemma between information sharing and privacy protection to extract valuable information [12]. Several efforts have been devoted to preventing the disclosure of an individual's identity or sensitive information during knowledge discovery [24]. These paradigms, which are known as privacy-preserving data publishing (PPDP) [25] and Privacy-Preserving Data Mining (PPDM) [10], appeared to offer a viable solution to this problem.

Despite the significant overlap between PPDM and PPDP in data privacy protection, they differ in specific ways, particularly regarding their concepts in data mining results. Hence, it may be the intention of the data publisher to only publish a few data due to little or no interest in data mining algorithms and their results. The focus of PPDP is mainly on the data itself and not on data mining results [25]. Besides, the data used and how they are used determine the form of privacy. Therefore, several methods are used to ensure data privacy [26] [12]. These paradigms have recently piqued the interest of academics and designers, and many methods for protecting privacy have been devised, as well as far-reaching policies for sensitive data protection [26] [27] [10]. The type of privacy depends on the data and how it is used. As a result, many methods are employed to provide privacy [26]. Currently, no known generic solutions can address the entire privacy concerns about protecting sensitive information from unintended disclosure while maintaining the data's utility. Studies in this field have focused on finding appropriate treatments for specific issues. When successful data mining is undertaken for privacy protection, data utility and information loss are trade-offs [28] [29] [30] [31] [32] [21].

In this study, the scenario of a single publication model has been considered, whereby the big data are sanitized and independently published by many organizations (data collectors) that share several common individual records. The issue with this assumption is that, sometimes, the information of an individual may be published by more than one organization [22]. In this case, an attacker may launch a composition attack [23] [21] on such published datasets. Therefore, anonymization can only be achieved by altering individual records to conceal the linkage between the individual and specific values to avoid such attacks and preserve the possible utilization of the published data. The commonly known method to sanitize a database while publishing it is preserving privacy techniques [33]. The goal of using preserving privacy techniques is to reduce the risk of disclosing information about individuals and preserve the possible utilization of published data [12] [34] by using the methods of privacy protection, such as perturbing [35] [36] [37] [38], suppressing or generalizing variable values [17], thereby causing uncertainty in the identity inference or sensitive value estimation [39]. Although this method is helpful, information loss using protection methods is inevitable when attempting to attain a high level of privacy [40] [41]. Moreover, privacy-preserving techniques possibly affect the use of data, resulting in the production of imprecise or even impractical extraction of knowledge. Thus, balancing privacy and utility is essential in data applications [12] [41].

The contribution of this review study is: the existing techniques for preserving privacy are outlined and categorized based on the protection methods used to reduce the risk of disclosing information and how preserving privacy techniques can preserve privacy and protect sensitive data. This is regarded as the study's primary contribution, which is expected to help researchers in this field gain a deeper understanding of techniques for preserving privacy. In addition, this study compared and analyzed the benefits and drawbacks of various privacy preservation techniques.

This paper is organized as follows: Section **Error! Reference source not found.** provides an overview of big data and highlights its challenges. Section Error! Reference source not found. describes the issue of the composition attack. Section **Error! Reference source not found.** reviews the privacy-preserving techniques for privacy-preserving of data. Section **Error! Reference source not found.** briefly reviews the anonymization technique. Section **Error! Reference source not found.** discusses the protection methods used with the anonymization techniques. Section **Error! Reference source not found.** concludes this study.

2. COMPOSITION ATTACK

Composition attack results from a combination of different published datasets. Being a combination of different datasets, the attacker relies on the intersection of the datasets to exploit sensitive information, since datasets are rarely isolated. The complexity of this problem increases with the availability of more datasets from several data collectors. A simple example of a composition attack is a scenario where patients may have visited hospitals **A** and **B** for specialized procedures or follow-up. **Error! Reference source not found.A**

and **B** provide the data segments from the two hospitals. The original dataset is first anonymized from its initial value by the data collector (non-sensitive attributes are generalized or replaced with a new one) and made available to the intended recipients [21]. The attacker is assumed to have these knowledge bases to successfully launch a composition attack[13]. Firstly, the attacker is a close relative of the patient (i.e., a friend, neighbor, or colleague) and is assumed to know the hospitals visited by the patient. Secondly, it is assumed that the patient's quasi-identifier (QI) attribute, whose sensitive value is to be inferred as known to the attacker. Quasi-identifier (QI) attributes denote a sequence of individuals' non-explicit attributes (e.g., race, age, date of birth, ZIP code, and gender) wherein no single attribute can provide specific identification of the person, rather, all the attributes must be combined to identify the person. Assume that the following information of the patient is known to the attacker (Age = 25 years old, Sex = Male, and lives in Zipcode = 132000). The attacker is aware of the two hospitals, which were visited by the patient, and that the two hospitals have individually published their data without consulting each other. Consider **Error! Reference source not found.A** and **B** as the anonymized tables of the data made available by the two hospitals. It can be observed that this might increase the chances of breaching the privacy of the patient. The attacker may also find it difficult to exploit the sensitive information of the patient in either dataset as both datasets satisfy k-anonymity and l-diversity.

*Table 1: Patient's Data And Its Generalization At Hospital **A**.*

ID	Equivalence Class	Age	Gender	Zip code	Disease
1	1	<30	*	132000	HIV
2		<30	*	132005	Viral Infection
3		<30	*	132005	Flu
4		<30	*	132000	Viral Infection
5	2	3*	*	132004	Heart Disease
6		3*	*	132005	Flu
7		3*	*	132000	Cancer
8	3	≥40	*	132006	Viral Infection
9		≥40	*	132005	Heart Disease
10		≥40	*	132006	Heart Disease

*Table 2: Patient's Data And Its Generalization At Hospital **B**.*

ID	Equivalence Class	Age	Gender	Zip code	Disease
1	1	20-39	*	132***	HIV
2		20-39	*	132***	Tuberculosis
3		20-39	*	132***	Flu
4		20-39	*	132***	Cancer
5		20-39	*	132***	Heart Disease
6	2	≥40	*	132***	Tuberculosis
7		≥ 40	*	132***	Flu
8		≥ 40	*	132***	Cancer
9		≥ 40	*	132***	Viral Infection
10		≥ 40	*	132***	Heart Disease

However, a composition attack is mainly launched via intersection operation. Records in anonymized datasets are normally arranged in small groups, and each group has an identical set of QI values; hence, each group is referred to as an equivalence class, wherein all individuals are similar and are associated with sensitive values that depend on the applied anonymization method. For instance, **Error! Reference source not found.A** consists of an equivalence class associated with three sensitive values ('HIV', 'Viral Infection', and 'Flu'). Consequently, the attacker may not correctly identify the sensitive value of the victim, even though he/she can rightly identify the victim's equivalence class. Therefore, the privacy of the victim is safeguarded.

However, the intersection of **Error! Reference source not found.A** and **B** includes the two equivalence classes that contain the victim's record; hence, the attack can predict the sensitive value by arranging the QI values for this victim, leading to a breach of privacy [42] [43]. A person's privacy is compromised if the adversary's confidence is significantly larger than a random guess [13]. As a result, the use of privacy-preserving techniques to protect published data has been the most studied method in recent years [33] [12]. Privacy-preserving techniques aim to alter these attributes (QI values) to

conceal the linkage between the individual values and specific values to avoid such attacks and preserve the possible utilization of published data. In detail, the following section will describe these privacy-preserving techniques.

3. PRIVACY-PRESERVING TECHNIQUES

Even though there are many definitions of privacy, it is difficult to derive an accepted, standard definition of privacy because of the context-specific nature of the privacy concept, such as privacy at home, with family, in correspondence, etc. [9] [44] [10] [45]. Primarily, privacy relates to data privacy, communication privacy, body privacy, and territorial privacy. However, data privacy has been described by [44] [10] as the collection and management of personal information. Bodily privacy, on the other hand, is used to protect the body of individuals from certain invasive procedures, such as drug testing and others. Communications privacy refers to the protection of the confidentiality of all communication channels, while territorial privacy refers to protecting the established boundaries of an entity from intrusion.

The scope of this study is limited to data privacy. Data privacy implies not disclosing the sensitive value or inferring the identity of individuals or groups from the information obtained from data collectors. Privacy-preserving techniques ensure that vital information is not disclosed to anyone before being made public. Despite the overlap between privacy and confidentiality in some contexts, they differ in certain ways, especially relating to their concepts and methods of protection. Confidentiality is data related, meaning it is more about data itself and aims to protect data from unauthorized access, alteration, or loss when transferred over a network [12] [46]. By contrast, privacy has an additional data owner-oriented concept - it primarily deals with data owners and aims to protect their private information [47].

Furthermore, the type of privacy changes based on the data and how it is used. Therefore, various techniques are employed to ensure privacy [26] [12]. Three types of techniques were developed to preserve the privacy of information [12] [48]. These techniques are data exchange [49], data cryptographic [50], and data anonymization [12] [51] [52], as shown in **Error! Reference source not found.**.

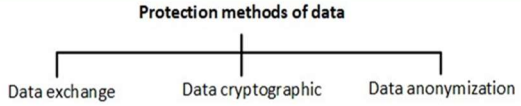

Figure 1: Data protection methods.

The data exchange technique can be used to publish private information from (at least) one data source to another. This technique only works in systems with trustworthy data sources. In other words, none of the data providers intends to endanger publicly available private information. The majority of natural systems, however, do not trust data providers because they could intentionally compromise shared personal data. Therefore, the preservation of private information cannot be guaranteed when using this data exchange technique [53] [54].

In data cryptography, multiple parties (i.e., data providers) typically cooperate to compute results or jointly analyze non-sensitive information, where pairs of public and private keys are available to each data provider. Moreover, the public keys of all data providers should be distributed to everyone, including the data warehouse servers (data collectors). Initially, all data providers are provided with the sum of the public keys as their reference to encrypt their data based on the provided reference for onward transmission to the data warehouse servers. Hence, no involved people know anything beyond their input. Through mathematical manipulations, accurate models can be built by the data warehouse servers based on the received encrypted data; this technique can solve privacy problems among mutual untrusted parties or competitors [55] [56]. However, the technique's complexity may cause high computing costs for data providers and warehouse servers when dealing with massive amounts of data, rendering the technique virtually unusable [57] [58]. Anonymization techniques maintain data usefulness while preventing attempts to identify the identity of the record owner [12] [21]. Data anonymization represents the key aspect of this review, which is thoroughly discussed in the next section.

4. DATA ANONYMIZATION TECHNIQUES

Data anonymization is a technique used to protect private or sensitive information so that an individual cannot be identified or disclose the sensitive value in the collected data. The collected data should be treated as a private table known as microdata table T.

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

The microdata table T comprises a set of tuples t. Each tuple t represents a client i that comprises various specific attributes related to the client (see **Error! Reference source not found.**) [59]. The various categories of these attributes could include identity attribute (IA), which identifies the owners' records like name, phone number, and address, and quasi-identifiers (QI) attribute, which refers to a range of attributes in which a person cannot be identified by any of the single attributes unless all the attributes have been combined, and sensitive attribute (SA), which covers the private information of the individual, such as the type of disease [60] [61].

Table 3: Medical Patient Database.

Identifier (IAs)	Quasi-Identifier (QI)				Sensitive (SA)
Name	Nationality	Age	Gender	Zip code	Disease
Peter	Japanese	30	Male	130350	HIV
Mary	American	23	Male	462351	Cancer
Michel	Canadian	28	Male	150352	Flu
Louis	Chinese	53	Male	160350	Heart Disease
Sofia	American	39	Female	462350	Viral Infection

Organizations are expected to publish only partial data derived from their large datasets in the form of microdata that can add value to their reputation or support plans without divulging the proprietorship of the sensitive data. Even though the Identity Attribute (IAs) that explicitly identify users from the table, such as names and social security numbers, are removed on the assumption of data protection, the remaining data in most of these cases can be used to re-identify the person. Moreover, Sensitive Attribute (SA) may still flow due to linking attacks wherein sensitive data may be revealed by linking the remaining attributes, such as in the published data, with other available data sources. This situation is known as a composition attack or intersection attack [21] [12]. Organizations are expected to publish only partial data derived from their large datasets in the form of microdata that can add value to their reputation or support plans without divulging the proprietorship of the sensitive data. The attributes (IAs), which explicitly identify users from the table, such as names and social security numbers, are removed, assuming that anonymity is maintained. However, the remaining data in most of these cases can be used to re-identify the person. Moreover, SA may still flow due to linking attacks, in which, sensitive data may be revealed by linking the QI attribute in the published data with other available data sources. This situation is known as a composition attack or intersection attack [33]. Thus, anonymization (sanitizing database) can only be achieved by altering these attributes to conceal the linkage between the individual values and specific values to avoid these attacks and preserve the possible utilization of published data, whereby the effective preservation of privacy can be attained by using different protection methods used with anonymization techniques before published data. Protection methods used with anonymization techniques aim to prevent attempts at recognizing the record owner's identity whilst preserving data utility [33]. The following subsections will describe, in detail, each of these protection methods.

5. PROTECTION METHODS USED WITH ANONYMIZATION TECHNIQUES

The anonymization techniques employ various protection methods, which can be used and combined within the same technique to introduce uncertainty into identity inference or sensitive value estimation. Protection methods used with anonymization techniques aim to avoid attempts to identify the record owner's identity by converting a dataset's original values to the anonymized dataset. When performing the extraction of knowledge through data mining operations, the anonymous dataset is used instead of the original dataset. In this study, the protection methods, which are used to sanitize data in anonymization techniques, are classified into three methods, as illustrated in **Error! Reference source not found.**, namely grouping methods, perturbation methods, and measurement correlation (similarity) methods [12] [21]. The following subsections describe these methods in detail.

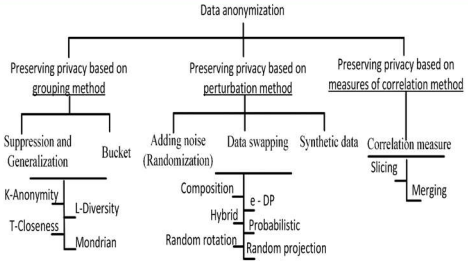

Figure 2: Classification Of The Protection Methods Of Data Anonymization Techniques.

5.1 Preserving Privacy Based on The Grouping Method

This method divides the entire records horizontally into several groups or partitions and only allows each tuple to belong to one group [62] [20]. This operation aims to weaken the linkage between the QI values and SA values, whereby individuals cannot be identified with their sensitive values in the group. Grouping is often implemented depending on suppression and generalization, and both may be implemented using suppression and generalization or bucketisation and/or combined, as described below:

5.2.1 Suppression and generalization

Suppression and generalization are two common ways to anonymize data. The first way is suppression. In the suppression way, the values of the attributes are replaced from the table to ANY, denoted by *. That means some or all attribute values are replaced by "*." The second way is a generalization. In the generalization way, a specific value of the attributes is replaced by a general value that is less specific according to a given taxonomy, thereby making the QI less identifying [17].

There are two major ways of anonymizing information using a generalization method: global recoding and local recoding. For global recording, once an attribute value is generalized, each value occurrence should be replaced by a new generalized value. For example, all values in birth date are in years, or all in nationality are related to continents. In local recoding, values may be generalized to different generalization domains. For example, local recoding may generalize values of the age attribute into [20–39], [40–59], and [60–90]. Local recoding is more similar to the original data and can preserve more information than global recoding; hence, the data mining operations are more accurate. In addition, overlapping intervals are unsuitable for most classification tools as they complicate classification tasks [17] [63].

5.2.1 Bucketisation

Bucketisation was proposed by Xiao and Tao [64] [39]. It divides the original data table into several non-overlapping partitions. Afterward, two tables are formed, namely the QI table and the sensitive table, and each division assigns a GID value. All the tuples in this partition have the same GID value. All the tuples are projected on the QI and confidential attributes to generate a sensitive table. Bucketisation ensures that an individual's attribute values are not distinguishable from others in the same bucket (see **Error! Reference source not found.**).

Table 4: Published by Bucketization.

Quasi Identifier Table				Sensitive Table		
ID	Age	Gender	Zip code	GID	GID	Disease
1	30	Male	130350	1	1	Flu
2	23	Male	130351	1	1	Cancer
3	28	Female	130352	1	1	Flu
4	53	Female	130350	2	2	Heart Disease
5	39	Female	130352	2	2	Flu
6	60	Male	130351	2	2	Heart Disease

Suppression and generalization are effective sensitive data protection techniques from unauthorized access because they hide or replace several details of the attributes. Furthermore, both address different attributes individually, i.e., they only adjust the selected values to minimize utility loss [65]. Moreover, the grouping formed by bucketisation is equivalent to the group formed by generalization, except that bucketisation contains all the original tuple values, whereas generalization contains several generalized tuples values [17]. Generalization has the advantage of providing a representation with consistent attribute values within each group, which makes the analysis of the published data easier [17].

The k-anonymity approach [66], the l-diversity approach [61], the T-closeness approach [18], and

the Mondrian approach [67] are the most favorable approaches for preserving privacy based on data grouping and anonymization. These approaches were proposed to protect privacy in one-time data publication. They destroy the relationships between the attribute values to make personal data anonymous or unattributable to a single source or individual.

Overall, anonymization approaches based on the grouping method are simple and attempt to protect the privacy of individuals. They have an intrinsic drawback. They cannot continually and effectively protect the records' critical values against a composition attack [68]. Furthermore, it has been shown that optimal anonymization is an NP-Hard problem [69] [12]. Furthermore, this method is ineffective because high dimensionality makes it possible to unmask the identities of the primary record-holders via merging data with either background or public (composition attack) information [21] [68]. A full analysis of these approaches may be found in [10] [59] [65] [70] [12].

5.2 Preserving privacy based on the perturbation method

The goal of perturbation is to protect sensitive information in a manner that makes it challenging for an attacker to use attribute linkage attacks to identify a specific person in a published dataset or to infer a specific person's precise, sensitive value. It generally brings about uncertainty in published datasets and negatively affects the chances of inferring the individuals' sensitive information [17] [71]. Among the most favorable methods of anonymity in perturbation is adding noise (randomization) to the data [36] [72] [29] [73], creating synthetic data [37] [74] [75], or swapping attributes [35].

5.2.1 Adding noise (randomization)

One of the most popular perturbation methods is randomization (adding noise) [29] [73] [10]. This method involves a particular perturbation of the original data values either by introducing or multiplying a randomized or stochastic number to conceal the distinguishing values of the records. Accordingly, the opponents cannot deduce the private attributes of a specific person by relating the attributes. Therefore, the perturbed data value of an individual can be considerably different from its original version. An example is a situation where the GPA of a student is fraudulently increased from 3.45 to 3.65.

Work on additive noise was first publicized by Kim [76] with the general expression that $X + \beta$. The general idea is that the data owner publishes the tuples derived from $X + \beta$ instead of X, where X is the original data value, and β is a random value drawn from a certain distribution [77]. Privacy is measured by how closely the original values of a modified attribute can be estimated [78]. Furthermore, the experiments in [79][80][81] suggested that some data can be preserved in the randomized data for data mining operations. Fuller [82] and Kim et al. [83] showed that the addition of random noise would not affect some simple statistical information, such as correlations and means.

Despite the intuitive and simplistic nature of the noise-adding concept, it also has a few drawbacks. Data scholars have noted that the original data, after being distorted by noise addition, will ensure that the personal identifying information is removed, and data is distorted. However, the remaining data can disclose the individual's identity and other sensory attributes [76] [84]. In addition, Kargupta et al. and Binjubeir et al. [85] [12] highlighted that the chance to recover several sensitive values from randomized data is possible when there is a strong correlation between the diverse attributes. Achieving optimal data privacy by adding noise substantially increases the computational cost and results in the loss of several statistical data properties, rendering the dataset almost useless to the user [76]. Therefore, balancing data privacy and data utility is necessary [12].

5.2.1 Data swapping

Data swapping was first presented by T. Dalenius et al. [35] as a way of preserving data privacy, especially in datasets that contain categorical variables. The basis of this way is to transform the original data into a distorted version that will still retain the same frequency count statistics as the original version by altering the data values of selected cells. Data swapping is useful in protecting both numerical [86] and categorical attributes [87].

Swapping allows the masking of information of all individuals and only needs to be performed on sensitive variables to break the relationship between the record and the individual, thereby leaving the QI variables undisturbed.

Swapping works well, but its major drawback is that it may not maintain multivariate relationships. Furthermore, data mining operations may likely be affected by swapping [88]. Swapping may also result in nonsensical combinations. If the microdata table contains gender and type of cancer, the resultant data after a swap may have a record indicating a male with ovarian cancer [41].

Rank swapping is an alternative to the swapping way [89]. The values of an attribute a_i are first ranked in ascending order before swapping each of the ranked values of a_i with another randomly selected ranked values from a specified range. Rank swapping can maintain multivariate relationships more appropriately than ordinary data swapping [88]. The main difference between rank swapping and ordinary data swapping is that the range over which the data can be swapped is restricted. The advantage is that it limits the values that can be swapped with other values, whilst the difficulty is in finding the cells for swapping that will maintain the multivariate relationships of interest [37] [39].

5.2.1 Synthetic data

Privacy in data publishing can be achieved using synthetic data [25]. Synthetic data are used to produce data with distributional characteristics like those of the original data, instead of altering the original dataset or using it as it is. The beauty of synthetic data stems from the fact that it comes from real data and real distributions, making it almost indistinguishable from the original data. Therefore, One of this approach's critical benefits is that an attacker cannot reveal private information by obtaining published data, but the identified data lack sufficient utility [90]. In addition, many statistical disclosure methods are used to generate synthetic data based on patterns found in the original dataset [74]. For example, condensation is used to represent synthetic data [91]. The general idea is to first build a statistical model from the data by condensing the records into multiple groups based on their centers, radii, and sizes. Then, another set of data can be generated based on the statistical information.

In the last decade, the probabilistic approach [92], the e-differential privacy approach $(e-DP)$ [93], the hybrid approach [68], and the composition [94] preserving privacy based on the perturbation method have been proposed for multiple independent data publishing. Composition is the first privacy model to prevent composition attacks in multiple independent data publishing [21]. The proposed approach in [94] combined two new concepts, including (ρ, α)-anonymization via sampling and composition-based generalization to protect independent datasets from composition attacks. In [68], the proposed method merged sampling, generalization, and perturbation by ensuring that, in each equivalence class, Laplacian noise is added to the count of every sensitive value. A new approach called (d, α)-linkable was also proposed by the probabilistic approach, which strives towards reducing the chances of an attacker successfully launching a composition attack by ensuring that association of the d confidential values with a quasi-identifying group with α probability through establishing the relationships between the QI attributes and SAs.

Liu et al. [95] introduced a new protection method called rotating. It changes (rotating) the data in a specific way to protect private information in public data sets from composition attacks. One disadvantage of data rotation is that domain-specific features, such as the inner product and Euclidean distance, are lost. Simultaneously, projection matrices have been used to keep mined data sets anonymous [96]. Where provides a number of random projection matrices that can be used to protect privacy from composition attacks in various data mining applications. However, identifying the actual data's approximation is possible [73].

Mohammed [93] proposed the first noninteractive-based approach, called $e-DP$ based on the generalization method. The proposed solution probabilistically generates a generalized contingency table and then adds noise to the counts. The $e-DP$ provides a strong privacy guarantee for statistical query answering and protection against composition attacks by differential privacy-based data anonymization [23] [68] [97] [98] [21] showed that using $e-DP$ to protect against composition attacks generates a substantial amount of data utility losses during anonymization.

5.3 Preserving privacy based on the measures of correlation (similarity)

In this method, the data sets have several correlated attributes (multiple dimensions) rather than single column distribution to obtain exceptional results for data mining operations in privacy preservation. Previous grouping methods rearrange the data distributions to execute mining for privacy preservation, which involves analyzing each

dimension separately, thereby overlooking the correlations among various attributes (dimensions) [95]. While preserving privacy based on the perturbation method is altering the original values of dataset D to its anonymized version $D1$, which leads to data utility problems depending on the amount and type of noise or the specific properties of that data are not preserved [76]. Using a correlation (similarity) metric to enhance protection and preserve more data utility is a brilliant solution to these issues [21] [38]. The idea behind the correlation measure is to keep data utility by grouping highly correlated attributes together in columns and preserving the correlations between such attributes. The correlation measure protects privacy because it breaks the associations between uncorrelated attributes in another column via protection methods based on anonymization approaches, such as randomly permutated, generalization, and so on [33] [21] [38].

The degree of association is measured between variables by a correlation coefficient, denoted by r. Where the r plays a significant role in data science techniques that measure the strength of association between variables, choosing a particular similarity measure can significantly cause the success or failure of many algorithms [99].

Pearson correlation coefficient (PCC) and mean square contingency coefficient (MSCC) are the two commonly used measures of association [38]. PCC is used to determine the strength of a linear relationship between two continuous variables. The value of the coefficient r ranges from [−1,+1]. When the value of r is −1 or +1, a perfect linear relationship exists between the considered variables. However, if the value is 0, no linear relationship exists between the pairs of variables. By contrast, MSCC is a chi-square measure of the correlation between two categorical attributes. Unlike PCC, chi-square measures the extent of the significance of the relationship instead of measuring the strength of the relationship. The value of this coefficient r ranges from [0, 1].

The most recent approaches under the measure correlation category are slicing [38], merging [21] and the UL approach. Slicing has received a lot of attention for privacy-preserving data publishing, which is considered a novel data anonymization approach. The authors presented a risk disclosure prevention concept that is devoid of generalization. Therefore, slicing ensures data privacy and preserves data utilities because the attribute values are not generalized. Slicing randomly permutates the values of attributes in the bucket to annul the column-wise relationships. This method protects the privacy of the published records from attribute and membership disclosure risks. In addition, slicing is recommended for high-dimensional data anonymization because it keeps more data utility than a generalization of attribute values. It uses vertical partitioning (attribute grouping) and horizontal partitioning (tuple partition), and its sliced table should be randomly permutated [38] (see **Error! Reference source not found.**).

Table 5: Published Data by Slicing.

(Age, Gender)	(Zip code, Disease)
(30, F)	(130350, ovarian cancer)
(23, M)	(130350, heart disease)
(28, F)	(130352, Flu)
(53, F)	(130350, heart disease)
(39, F)	(130352, Flu)
(60, M)	(130351, heart disease)

Error! Reference source not found. provides the results of measuring the associations (similarity) among the attributes. The attribute group (vertical partitioning) is applied, where the values of the highly correlated attributes are grouped into columns and uncorrelated attributes in other columns. Attribute partition is represented by {Age, Gender}, {Zip code, Disease} while tuple partition is applied by grouping tuples into an equivalence class $\{\{t1, t2, t3\}, \{t4, t5, t6\}\}$. The main part of the tuple partition (horizontal partitioning) is to group all tuples that contain identical values in the same equivalence class or close to each other, thereby making it easier to break down uncorrelated attributes, and check whether an equivalence class satisfies I-diversity [12][67]. For slicing, the values of attributes are randomly permutated between the uncorrelated attributes to break the linkage between different columns. In contrast, the attributes in columns that are highly correlated remain unchanged. Yet, the aspect of it remains an open question: Does randomness always protect the identities of individuals from disclosure? Slicing can result in data utility and privacy-preserving problems, as slicing randomly permutate attribute

values in each bucket with the chances of creating fake tuples which will negatively affect the utility of the published microdata. For instance, in **Error! Reference source not found.**, tuple t_1 has just one matching equivalence class linked with two sensitive values for zip code 130350. Here, any person may be linked with sensitive values with a probability of not more than 1/l via l-diverse slicing because slicing has been shown to satisfy l-diverse slicing by being linked with the sensitive values by 1/2. If the QI attribute, namely, the zip code, is revealed because it has high identical attribute values (sufficient variety) and an adversary relies on background knowledge and has a knowledge of (23, M), then the adversary can determine the SA for the individual. Moreover, if the slicing algorithm switches the sensitive value (randomly) between t_1 and t_2, then incompatibility is created between the SA and QI attribute values, as mentioned in [41]. Additionally, the attacker can rely on the analysis of the fake tuples in the published table to capture the concept of the deployed anonymization mechanism, thereby having the chance to violate the privacy of published data [41] [12] [10].

The merging approach was designed by Hasan et al. [21] to protect the personal identity from disclosure, and it is considered an extension of slicing. The basic aim of merging is privacy preservation in multiple independent data publications via cell generalization and random attribute value permutation to break the linkage between different columns. To compute data utility and privacy risks, the merging approach that preserves data utility has small privacy risks because it increases the false matches in the published datasets. However, the major drawback of merging is the random permutation procedure for attribute values to break the association between different columns. Besides increasing the false matches for unique attributes in the published datasets, these procedures may generate a small fraction of fake tuples but result in a large number of matching buckets (more than the original tuples), which will lead to loss of data utility and can produce erroneous or infeasible extraction of knowledge through data mining operations [100] [101].

The UL approach for data anonymization was proposed by Mohammed et al. [33]. The UL approach strikes a better balance between utility, information loss, and privacy. The UL approach divides the data into horizontal and vertical partitions in particular. The values of the unique and identical attributes are then determined using the lower and upper protection levels (LPL and UPL). To protect the published data from disclosure risks and to ensure that the published table is l-diverse, the unique and identical attributes are rank swapped. The rest of the cells are protected from attribute disclosure and membership disclosure due to their presence in more than one equivalence class.

Therefore, the primary reason for revealing people's identity is the existence of unique attributes in the table, or it involves allowing several attributes in the row to match the attributes in other rows, leading to the possibility of accurately extracting the attributes of a person [38] [21] [12].

Previous studies [13] [38] confirmed the importance of allowing a tuple to match multiple buckets to ensure protection against attribute and membership disclosure. This finding implies that mapping the records of an individual to more than one equivalence class results in the formation of a super equivalence class from the set of equivalence classes.

Table 6 summarizes the three protection methods that classify all privacy-preserving techniques. The criterion for evaluating the efficiency of the anonymization technique is the capability of data privacy-preserving by decreasing the likelihood of revealing people's data and keeping the possibility of the published data being used. The privacy preservation level refers to difficulty disclosing information to individuals [66]. Data utility refers to what extent we can use the sterile database for intensive analyses.

Table 6: A summary of protection methods

Protection methods	Data utility	Privacy Preservation Level
grouping methods	high	low
perturbation methods	low	high
measurement correlation methods	high	high

6. CONCLUSION

Many data analysis applications face the challenge of maintaining information privacy. In recent years, many partially published data sets have been the subject of various concerns, ranging from unauthorized access to private data to privacy violations and unintended use of personal information. This issue has slowed progress in publishing data, necessitating robust privacy-protection techniques that can reduce the chances of

unauthorized individuals identifying sensitive information. The existing privacy-preserving techniques are intensively reviewed and classified in this paper based on their protection methods to minimize the risk of information disclosure. Furthermore, this review investigated and analyzed the benefits and drawbacks of various protection methods used with anonymization techniques. They were classified based on the method used to reduce the risk of information disclosure, representing the study's main contribution to providing researchers in this field with a comprehensive understanding of privacy-preserving techniques. The study's findings show that privacy-preserving techniques continue to face potential challenges and open issues, paving the way for additional research by scholars in the data privacy and protection field. More technical research is needed to propose an effective solution to the highlighted challenges of these techniques raised in this research.

ACKNOWLEDGEMENT

This paper is part of a funded Project 'Developing an effective Privacy Protection Framework for Oman's current Healthcare Sector in Big data environment' from the Ministry of Higher Education, Research and Innovation (MoHERI/BFP/MC/01/2020), Sultanate of Oman in call 2020; and from ump with vote no. RDU220304.

REFERENCES:

[1] D. Zhang, "Granularities and Inconsistencies in Big Data Analysis," *International Journal of Software Engineering and Knowledge Engineering*, vol. 23, no. 06, pp. 887–893, Aug. 2013, doi: 10.1142/S0218194013500241.

[2] M. Bin Jubeir, M. A. Ismail, S. Kasim, H. Amnur, and others, "Big Healthcare Data: Survey of Challenges and Privacy," *JOIV: International Journal on Informatics Visualization*, vol. 4, no. 4, pp. 184–190, 2020.

[3] H. Hu, Y. Wen, T. S. Chua, and X. Li, "Toward scalable systems for big data analytics: A technology tutorial," *IEEE Access*, vol. 2, pp. 652–687, 2014, doi: 10.1109/ACCESS.2014.2332453.

[4] A. Taouli, amar bensaber Djamel, N. Keskes, K. Bencherif, and B. Hassan, "Semantic for Big Data Analysis: A survey," in *Proceedings of the 7th Innovation and New Trends in Information Systems conference*, 2018.

[5] D. Laney, "3D data management: Controlling data volume, velocity and variety," *META group research note*, vol. 6, no. 70, p. 1, 2001.

[6] V. Rubin and T. Lukoianova, "Veracity roadmap: Is big data objective, truthful and credible?," *Advances in Classification Research Online*, vol. 24, no. 1, p. 4, 2013.

[7] C. L. Philip Chen and C. Y. Zhang, "Data-intensive applications, challenges, techniques and technologies: A survey on Big Data," *Information Sciences*, vol. 275, pp. 314–347, 2014, doi: 10.1016/j.ins.2014.01.015.

[8] Lei Xu, Chunxiao Jiang, Jian Wang, Jian Yuan, and Yong Ren, "Information Security in Big Data: Privacy and Data Mining," *IEEE Access*, vol. 2, pp. 1149–1176, 2014, doi: 10.1109/access.2014.2362522.

[9] L. Cranor, T. Rabin, V. Shmatikov, S. Vadhan, and D. Weitzner, "Towards a Privacy Research Roadmap for the Computing Community," *arXiv preprint arXiv:1604.03160*, 2016.

[10] R. Mendes and J. P. Vilela, "Privacy-Preserving Data Mining: Methods, Metrics, and Applications," *IEEE Access*, vol. 5, pp. 10562–10582, 2017, doi: 10.1109/ACCESS.2017.2706947.

[11] G. Jagannathan and R. N. Wright, "Privacy-preserving imputation of missing data," *Data and Knowledge Engineering*, vol. 65, no. 1, pp. 40–56, 2008, doi: 10.1016/j.datak.2007.06.013.

[12] M. Binjubeir, A. A. Ahmed, M. A. Bin Ismail, A. S. Sadiq, and M. Khurram Khan, "Comprehensive Survey on Big Data Privacy Protection," *IEEE Access*, vol. 8, pp. 20067–20079, 2020, doi: 10.1109/ACCESS.2019.2962368.

[13] A. Gkoulalas-Divanis and G. Loukides, *Medical Data Privacy Handbook*. Cham: Springer International Publishing, 2015.

[14] J. M. Cavanillas, E. Curry, and W. Wahlster, *New Horizons for a Data-Driven Economy*. Cham: Springer International Publishing, 2016.

[15] B. C. M. Fung, K. Wang, A. W.-C. Fu, and J. Pei, "Anonymity for continuous data publishing," in *Proceedings of the 11th international conference on Extending database technology: Advances in database technology*, 2008, pp. 264–275.

[16] R. C.-W. Wong, A. W.-C. Fu, J. Liu, K. Wang, and Y. Xu, "Global privacy guarantee in serial data publishing," in *2010 IEEE 26th International Conference on Data Engineering (ICDE 2010)*, 2010, pp. 956–959.

[17] R. C.-W. Wong and A. W.-C. Fu, "Privacy-preserving data publishing: An overview," *Synthesis Lectures on Data Management*, vol. 2, no. 1, pp. 1–138, 2010, doi: https://doi.org/10.2200/S00237ED1V01Y201003DTM002.

[18] N. Li, T. Li, and S. Venkatasubramanian, "t-Closeness: Privacy Beyond k-Anonymity and l-Diversity," in *2007 IEEE 23rd International Conference on Data Engineering*, 2007, pp. 106–115, doi: 10.1109/ICDE.2007.367856.

[19] A. Machanavajjhala, D. Kifer, J. Gehrke, and M. Venkitasubramaniam, "L -diversity," *ACM Transactions on Knowledge Discovery from Data*, vol. 1, no. 1, p. 3, Mar. 2007, doi: 10.1145/1217299.1217302.

[20] L. Sweeney, "k-anonymity: A model for protecting privacy," *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems*, vol. 10, no. 05, pp. 557–570, 2002.

[21] A. Hasan, Q. Jiang, H. Chen, and S. Wang, "A New Approach to Privacy-Preserving Multiple Independent Data Publishing," *Applied Sciences*, vol. 8, no. 5, p. 783, May 2018, doi: 10.3390/app8050783.

[22] B. Malin and L. Sweeney, "How (not) to protect genomic data privacy in a distributed network: using trail re-identification to evaluate and design anonymity protection systems," *Journal of Biomedical Informatics*, vol. 37, no. 3, pp. 179–192, Jun. 2004, doi: 10.1016/j.jbi.2004.04.005.

[23] S. R. Ganta, S. P. Kasiviswanathan, and A. Smith, "Composition attacks and auxiliary information in data privacy," in *Proceeding of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining - KDD 08*, 2008, p. 265, doi: 10.1145/1401890.1401926.

[24] M. I. Jordan and T. M. Mitchell, "Machine learning: Trends, perspectives, and prospects," *Science*, vol. 349, no. 6245, pp. 255–260, 2015.

[25] B.-C. Chen, D. Kifer, K. LeFevre, and A. Machanavajjhala, "Privacy-Preserving Data Publishing," *Foundations and Trends® in Databases*, vol. 2, no. 1–2, pp. 1–167, Jun. 2009, doi: 10.1561/1900000008.

[26] A. Shah and R. Gulati, "Privacy Preserving Data Mining: Techniques, Classification and Implications - A Survey," *International Journal of Computer Applications*, vol. 137, no. 12, pp. 40–46, 2016, doi: 10.5120/ijca2016909006.

[27] P. Bhaladhare and D. Jinwala, "A sensitive attribute based clustering method for k-anonymization," in *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 2012, vol. 7135 LNCS, pp. 163–170, doi: 10.1007/978-3-642-29280-4_18.

[28] Y. Ding and K. Klein, "Model-driven application-level encryption for the privacy of e-health data," in *ARES 2010 - 5th International Conference on Availability, Reliability, and Security*, 2010, pp. 341–346, doi: 10.1109/ARES.2010.91.

[29] A. Shah and R. Gulati, "Privacy Preserving Data Mining: Techniques, Classification and Implications-A Survey," *International Journal of Computer Applications*, vol. 137, no. 12, 2016.

[30] P. R. Bhaladhare and D. C. Jinwala, "Novel approaches for privacy preserving data mining in k-anonymity model," *Journal of Information Science and Engineering*, vol. 32, no. 1, pp. 63–78, 2016.

[31] J. Yu, Z. Kuang, B. Zhang, W. Zhang, D. Lin, and J. Fan, "Leveraging content sensitiveness and user trustworthiness to recommend fine-grained privacy settings for social image sharing," *IEEE transactions on information forensics and security*, vol. 13, no. 5, pp. 1317–1332, 2018.

[32] J. Yu, B. Zhang, Z. Kuang, D. Lin, and J. Fan, "iPrivacy: image privacy protection by identifying sensitive objects via deep multi-task learning," *IEEE Transactions on Information Forensics and Security*, vol. 12, no. 5, pp. 1005–1016, 2016.

[33] M. BinJubier, M. Arfian Ismail, A. Ali Ahmed, and A. Safaa Sadiq, "Slicing-Based Enhanced Method for Privacy-Preserving in Publishing Big Data," *Computers, Materials & Continua*, vol. 72, no. 2, pp. 3665–3686, 2022, doi: 10.32604/cmc.2022.024663.

[34] A. Majeed and S. Lee, "Anonymization Techniques for Privacy Preserving Data Publishing: A Comprehensive Survey," *IEEE Access*, vol. 9, pp. 8512–8545, 2021, doi: 10.1109/ACCESS.2020.3045700.

[35] S. E. Fienberg and J. McIntyre, "Data swapping: Variations on a theme by dalenius and reiss," in *International Workshop on Privacy in Statistical Databases*, 2004, pp. 14–29.

[36] R. Brand, "Microdata Protection through

Noise Addition," in *Inference control in statistical databases*, Springer, 2002, pp. 97–116.

[37] C. K. Liew, U. J. Choi, and C. J. Liew, "A data distortion by probability distribution," *ACM Transactions on Database Systems (TODS)*, vol. 10, no. 3, pp. 395–411, 1985.

[38] T. Li, N. Li, J. Zhang, and I. Molloy, "Slicing: A New Approach for Privacy Preserving Data Publishing," *IEEE Transactions on Knowledge and Data Engineering*, vol. 24, no. 3, pp. 561–574, Mar. 2012, doi: 10.1109/TKDE.2010.236.

[39] T. A. Lasko and S. A. Vinterbo, "Spectral Anonymization of Data," *IEEE Transactions on Knowledge and Data Engineering*, vol. 22, no. 3, pp. 437–446, Mar. 2010, doi: 10.1109/TKDE.2009.88.

[40] A. Mehmood, I. Natgunanathan, Y. Xiang, G. Hua, and S. Guo, "Protection of big data privacy," *IEEE Access*, vol. 4, no. January, pp. 1821–1834, 2016, doi: 10.1109/ACCESS.2016.2558446.

[41] A. S. M. T. Hasan, Q. Jiang, J. Luo, C. Li, and L. Chen, "An effective value swapping method for privacy preserving data publishing," *Security and Communication Networks*, vol. 9, no. 16, pp. 3219–3228, Nov. 2016, doi: 10.1002/sec.1527.

[42] S. Gambs, M.-O. Killijian, and M. N. del Prado Cortez, "Show me how you move and I will tell you who you are," in *Proceedings of the 3rd ACM SIGSPATIAL International Workshop on Security and Privacy in GIS and LBS - SPRINGL '10*, 2010, p. 34, doi: 10.1145/1868470.1868479.

[43] A. S. M. Hasan, Q. Jiang, and C. Li, "An effective grouping method for privacy-preserving bike sharing data publishing," *Future Internet*, vol. 9, no. 4, p. 65, 2017.

[44] D. Banisar and S. Davies, "Global trends in privacy protection : An international survey of privacy, data protection, and surveillance laws and developments," *The John Marshall Journal of Computer and Information Law*, vol. 18, no. 1, pp. 1–111, 1999.

[45] W. Gan, J. Chun-Wei, H.-C. Chao, S.-L. Wang, and S. Y. Philip, "Privacy preserving utility mining: A survey," in *2018 IEEE International Conference on Big Data (Big Data)*, 2018, pp. 2617–2626.

[46] A. Senosi and G. Sibiya, "Classification and evaluation of privacy preserving data mining: a review," in *2017 IEEE AFRICON*, 2017, pp. 849–855.

[47] T. Wang, Z. Zheng, M. H. Rehmani, S. Yao, and Z. Huo, "Privacy Preservation in Big Data From the Communication Perspective—A Survey," *IEEE Communications Surveys & Tutorials*, vol. 21, no. 1, pp. 753–778, 2019, doi: 10.1109/COMST.2018.2865107.

[48] E. Bertino, D. Lin, and W. Jiang, "A Survey of Quantification of Privacy Preserving Data Mining Algorithms," in *Privacy-preserving data mining: Models and Algorithms*, Springer, Boston, MA, 2008, pp. 183–205.

[49] R. Conway and D. Strip, "Selective partial access to a database," in *Proceedings of the 1976 Annual conference, ACM 1976*, 1976, pp. 85–89, doi: 10.1145/800191.805537.

[50] Z. Yang, S. Zhong, and R. N. Wright, "Privacy-Preserving Classification of Customer Data without Loss of Accuracy," in *Proceedings of the 2005 SIAM International Conference on Data Mining*, 2005, pp. 92–102, doi: 10.1137/1.9781611972757.9.

[51] R. Agrawal and R. Srikant, "Privacy-preserving data mining," in *ACM Sigmod Record*, 2000, vol. 29, no. 2, pp. 439–450.

[52] S. Agrawal and J. R. Haritsa, "A framework for high-accuracy privacy-preserving mining," in *Data Engineering, 2005. ICDE 2005. Proceedings. 21st International Conference on*, 2005, pp. 193–204.

[53] C. Clifton *et al.*, "Privacy-preserving data integration and sharing," in *Proceedings of the ACM SIGMOD International Conference on Management of Data*, 2004, pp. 19–26, doi: 10.1145/1008694.1008698.

[54] N. Zhang, "Privacy-preserving data mining," Texas A&M University, 2006.

[55] H. Vaghashia and A. Ganatra, "A Survey: Privacy Preservation Techniques in Data Mining," *International Journal of Computer Applications*, vol. 119, no. 4, pp. 20–26, 2015, doi: 10.5120/21056-3704.

[56] Y. Lindell, "Secure Multiparty Computation for Privacy Preserving Data Mining," in *Encyclopedia of Data Warehousing and Mining*, vol. 1, no. 1, IGI Global, 2011, p. 5.

[57] T. ElGamal, "A Public Key Cryptosystem and a Signature Scheme Based on Discrete Logarithms," in *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 1985, vol. 196 LNCS, no. 4,

pp. 10–18, doi: 10.1007/3-540-39568-7_2.

[58] Z. Luo and C. Wen, "A chaos-based multiplicative perturbation scheme for privacy preserving data mining," in *2014 IEEE 5th International Conference on Software Engineering and Service Science*, 2014, pp. 941–944, doi: 10.1109/ICSESS.2014.6933720.

[59] Y. A. A. S. Aldeen, M. Salleh, and M. A. Razzaque, "A comprehensive review on privacy preserving data mining," *SpringerPlus*, vol. 4, no. 1, p. 694, Dec. 2015, doi: 10.1186/s40064-015-1481-x.

[60] A. Sharma, "Literature Survey of Privacy Preserving Data Publishing (PPDP) Techniques," *International Journal Of Engineering And Computer Science*, vol. 6, no. 5, pp. 1–12, 2017, doi: 10.18535/ijecs/v6i4.12.

[61] A. Machanavajjhala, J. Gehrke, D. Kifer, and M. Venkitasubramaniam, "L-diversity: privacy beyond k-anonymity," in *22nd International Conference on Data Engineering (ICDE'06)*, 2006, pp. 24–24, doi: 10.1109/ICDE.2006.1.

[62] X. Xiao and Y. Tao, "Anatomy: Simple and effective privacy preservation," in *Proceedings of the 32nd international conference on Very large data bases*, 2006, pp. 139–150.

[63] S. Wen, W. Wu, and A. Castiglione, *Cyberspace Safety and Security*, vol. 10581. Cham: Springer International Publishing, 2017.

[64] X. He, Y. Xiao, Y. Li, Q. Wang, W. Wang, and B. Shi, "Permutation Anonymization: Improving Anatomy for Privacy Preservation in Data Publication," in *Pacific-Asia Conference on Knowledge Discovery and Data Mining*, 2012, pp. 111–123.

[65] 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," in *SIGMOD Record*, 2004, vol. 33, no. 1, pp. 50–57, doi: 10.1145/974121.974131.

[66] R. J. Bayardo and R. Agrawal, "Data Privacy through Optimal k-Anonymization," in *21st International Conference on Data Engineering (ICDE'05)*, 2005, pp. 217–228, doi: 10.1109/ICDE.2005.42.

[67] K. LeFevre, D. J. DeWitt, and R. Ramakrishnan, "Mondrian Multidimensional K-Anonymity," in *22nd International Conference on Data Engineering (ICDE'06)*, 2006, pp. 25–25, doi: 10.1109/ICDE.2006.101.

[68] J. Li, M. M. Baig, A. H. M. Sarowar Sattar, X. Ding, J. Liu, and M. W. Vincent, "A hybrid approach to prevent composition attacks for independent data releases," *Information Sciences*, vol. 367–368, pp. 324–336, Nov. 2016, doi: 10.1016/j.ins.2016.05.009.

[69] E. W. Chambers, A. De Mesmay, and T. Ophelders, "On the complexity of optimal homotopies," in *Proceedings of the Annual ACM-SIAM Symposium on Discrete Algorithms*, 2018, pp. 1121–1134, doi: 10.1137/1.9781611975031.73.

[70] N. Zhang and W. Zhao, "Privacy-Preserving Data Mining Systems," *ieee*, vol. 40, no. 4, pp. 52–58, Apr. 2007, doi: 10.1109/MC.2007.142.

[71] E. Zorarpacı and S. A. Özel, "Privacy preserving classification over differentially private data," *WIREs Data Mining and Knowledge Discovery*, vol. 11, no. 3, p. e1399, May 2021, doi: 10.1002/widm.1399.

[72] S. Chawla, C. Dwork, F. McSherry, A. Smith, and H. Wee, "Toward privacy in public databases," in *Theory of Cryptography Conference*, 2005, pp. 363–385.

[73] X. Li, Z. Yan, and P. Zhang, "A Review on Privacy-Preserving Data Mining," in *2014 IEEE International Conference on Computer and Information Technology*, 2014, pp. 769–774, doi: 10.1109/CIT.2014.135.

[74] D. B. Rubin, "Statistical disclosure limitation," *Journal of official Statistics*, vol. 9, no. 2, pp. 461–468, 1993.

[75] J. Domingo-Ferrer, *Inference Control in Statistical Databases*, vol. 2316. Berlin, Heidelberg: Springer Berlin Heidelberg, 2002.

[76] K. Mivule, "Utilizing noise addition for data privacy, an overview," *Proceedings of The International Conference on Information and Knowledge Engineering (IKE 2012)*, pp. 65–71, 2013.

[77] J. J. Kim, "A method for limiting disclosure in microdata based on random noise and transformation," in *Proceedings of the section on survey research methods*, 1986, pp. 303–308.

[78] D. Agrawal and C. C. Aggarwal, "On the design and quantification of privacy preserving data mining algorithms," in *Proceedings of the twentieth ACM SIGMOD-SIGACT-SIGART symposium on Principles of database systems*, 2001, pp. 247–255.

[79] A. Charu and S. Y. Philip, *Privacy-Preserving Data Mining*, vol. 34. Boston, MA: Springer US, 2008.

[80] W. Du and Z. Zhan, "Using randomized response techniques for privacy-preserving data mining," in *Proceedings of the ninth ACM SIGKDD international conference on Knowledge discovery and data mining - KDD '03*, 2003, p. 505, doi: 10.1145/956750.956810.

[81] A. Evfimievski, R. Srikant, R. Agrawal, and J. Gehrke, "Privacy preserving mining of association rules," in *Proceedings of the eighth ACM SIGKDD international conference on Knowledge discovery and data mining - KDD '02*, 2002, vol. 29, no. 4, p. 217, doi: 10.1145/775047.775080.

[82] W. Fuller, "Masking procedures for microdata disclosure," *Journal of Official Statistics*, vol. 9, no. 2, pp. 383–406, 1993.

[83] J. J. Kim, W. E. Winkler, and others, "Masking microdata files," in *Proceedings of the Survey Research Methods Section, American Statistical Association*, 1995.

[84] V. Ciriani, S. De Capitani di Vimercati, S. Foresti, and P. Samarati, "κ-Anonymity," in *Secure data management in decentralized systems*, Springer, 2007, pp. 323–353.

[85] H. Kargupta, S. Datta, Q. Wang, and Krishnamoorthy Sivakumar, "On the privacy preserving properties of random data perturbation techniques," in *Third IEEE International Conference on Data Mining*, 2003, pp. 99–106, doi: 10.1109/ICDM.2003.1250908.

[86] S. P. Reiss, M. J. Post, and T. Dalenius, "Non-reversible privacy transformations," in *Proceedings of the 1st ACM SIGACT-SIGMOD symposium on Principles of database systems - PODS '82*, 1982, p. 139, doi: 10.1145/588111.588134.

[87] S. P. Reiss, "Practical data-swapping: the first steps," *ACM Transactions on Database Systems*, vol. 9, no. 1, pp. 20–37, Mar. 1984, doi: 10.1145/348.349.

[88] G. J. Matthews and O. Harel, "Data confidentiality: A review of methods for statistical disclosure limitation and methods for assessing privacy," *Statistics Surveys*, vol. 5, pp. 1–29, 2011, doi: 10.1214/11-SS074.

[89] B. C. M. Fung, K. Wang, A. W.-C. Fu, and P. S. Yu, *Introduction to Privacy-Preserving Data Publishing*. Chapman and Hall/CRC, 2010.

[90] J. Heldal and D.-C. Iancu, "Synthetic data generation for anonymization purposes. Application on the Norwegian Survey on living conditions/EHIS," 2019.

[91] C.-A. Saita and F. Llirbat, "Clustering Multidimensional Extended Objects to Speed Up Execution of Spatial Queries," in *International Conference on Extending Database Technology*, 2004, pp. 403–421.

[92] A. H. M. S. Sattar, J. Li, J. Liu, R. Heatherly, and B. Malin, "A probabilistic approach to mitigate composition attacks on privacy in non-coordinated environments," *Knowledge-Based Systems*, vol. 67, pp. 361–372, Sep. 2014, doi: 10.1016/j.knosys.2014.04.019.

[93] N. Mohammed, R. Chen, B. C. M. Fung, and P. S. Yu, "Differentially private data release for data mining," in *Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining - KDD '11*, 2011, p. 493, doi: 10.1145/2020408.2020487.

[94] M. M. Baig, J. Li, J. Liu, X. Ding, and H. Wang, "Data Privacy against Composition Attack," in *International Conference on Database Systems for Advanced Applications*, 2012, pp. 320–334.

[95] K. Chen and L. Liu, "Privacy preserving data classification with rotation perturbation," in *Proceedings - IEEE International Conference on Data Mining, ICDM*, 2005, pp. 589–592, doi: 10.1109/ICDM.2005.121.

[96] K. Liu, H. Kargupta, and J. Ryan, "Random projection-based multiplicative data perturbation for privacy preserving distributed data mining," *IEEE Transactions on Knowledge and Data Engineering*, vol. 18, no. 1, pp. 92–106, 2006, doi: 10.1109/TKDE.2006.14.

[97] G. Cormode, C. M. Procopiuc, Entong Shen, D. Srivastava, and Ting Yu, "Empirical privacy and empirical utility of anonymized data," in *2013 IEEE 29th International Conference on Data Engineering Workshops (ICDEW)*, 2013, pp. 77–82, doi: 10.1109/ICDEW.2013.6547431.

[98] R. Sarathy and K. Muralidhar, "Evaluating Laplace noise addition to satisfy differential privacy for numeric data.," *Trans. Data Priv.*, vol. 4, no. 1, pp. 1–17, 2011.

[99] J. Han, M. Kamber, and J. Pei, "Introduction," in *Data Mining*, 3rd ed., Elsevier, 2012, pp. 1–

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

38.

[100] 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*, Springer Nature Switzerland, 2020, pp. 57–65.

[101] S. Rohilla, "Privacy Preserving Data Publishing through Slicing," *American Journal of Networks and Communications*, vol. 4, no. 3, p. 45, 2015, doi: 10.11648/j.ajnc.s.2015040301.18.