Jurnal Elektronika dan Telekomunikasi (JET), Vol. 20, No. 2, December 2020, pp. 36-46
Accredited by RISTEKDIKTI, Decree No: **32a/E/KPT/2017**
doi: 10.14203/jet.v20.36-46

Classification of Privacy Preserving Data Mining Algorithms: A Review

Dedi Gunawan*

Informatics Department
Universitas Muhammadiyah Surakarta
Jl. A. Yani Tromol Pos.1, Pabelan, Kartasura
Surakarta, Indonesia

Abstract

Nowadays, data from various sources are gathered and stored in databases. The collection of the data does not give a significant impact unless the database owner conducts certain data analysis such as using data mining techniques to the databases. Presently, the development of data mining techniques and algorithms provides significant benefits for the information extraction process in terms of the quality, accuracy, and precision results. Realizing the fact that performing data mining tasks using some available data mining algorithms may disclose sensitive information of data subject in the databases, an action to protect privacy should be taken into account by the data owner. Therefore, privacy preserving data mining (PPDM) is becoming an emerging field of study in the data mining research group. The main purpose of PPDM is to investigate the side effects of data mining methods that originate from the penetration into the privacy of individuals and organizations. In addition, it guarantees that the data miners cannot reveal any personal sensitive information contained in a database, while at the same time data utility of a sanitized database does not significantly differ from that of the original one. In this paper, we present a wide view of current PPDM techniques by classifying them based on their taxonomy techniques to differentiate the characteristics of each approach. The review of the PPDM methods is described comprehensively to provide a profound understanding of the methods along with advantages, challenges, and future development for researchers and practitioners.

Keywords: Database, data mining, privacy preserving data mining, sensitive information.

I. INTRODUCTION

In today's era, data can be easily collected from various sources and stored in various types of databases. The collection of data in databases is meaningless until database owners conduct certain data analysis to excavate valuable information from the databases. In general, data analysis is carried out to extract useful information from databases, more specifically when it is used to find hidden knowledge in the databases then it is called data mining. Data mining plays an important role in many applications such as business management, marketing analysis, and science exploration [1]. The true value of data mining techniques does not reside in a set of complex algorithms; instead, it resides in the practical problems that it can help to solve [2]. There are two categories of data mining models such as predictive and descriptive. The predictive model aims to picture some predictions of a certain trend or correlation between one variable with other variables in a database such as regression, classification, and time series analysis. On the other hand, the descriptive model focuses on exploring knowledge from databases. Several data mining tasks that included in these models such as clustering, summarization, association rules, and sequence discovery.

Nowadays, various data mining software has been developed and published in the software market. However, not all people or institutions can utilize the software appropriately due to the lack of resources and limited knowledge in the institutions. A recent trend shows that institutions prefer to hire or use services from a data mining company to mine their data. Handling raw data to other institutions is not encouraged since there might be some sensitive information related to the institutions, their people, or customers.

In reality, several companies do not pay attention to the privacy issues residing in their database and that results in serious privacy violations [3]. Therefore, in this situation, the database owner should have to be careful in handling the database to other companies for the mining process due to some data mining tools that may cause sensitive information breach [4]. In another case, data recipients may also act as adversarial parties who might unfairly use the database to disclose sensitive information of individuals [5].

TABLE 1. PATIENTS DATA TABLE

Name	Birth date	Post code	Occupation	Disease
John	1975/12/10	71794	Engineer	Tuberculosis
Monna	1980/3/15	71780	Accountant	Dengue
Jane	1984/5/10	71794	Teacher	Pneumonia
Matip	1977/2/12	71793	Engineer	HIV
Mark	1978/8/16	71790	Programmer	Pneunomia
Hardy	1981/11/1	71790	IT specialist	Tuberculosis

* Corresponding Author.
Email: dedi.gunawan@ums.ac.id
Received: June 29, 2020 ; Revised: October 02, 2020
Accepted: October 13, 2020 ; Published: December 31, 2020

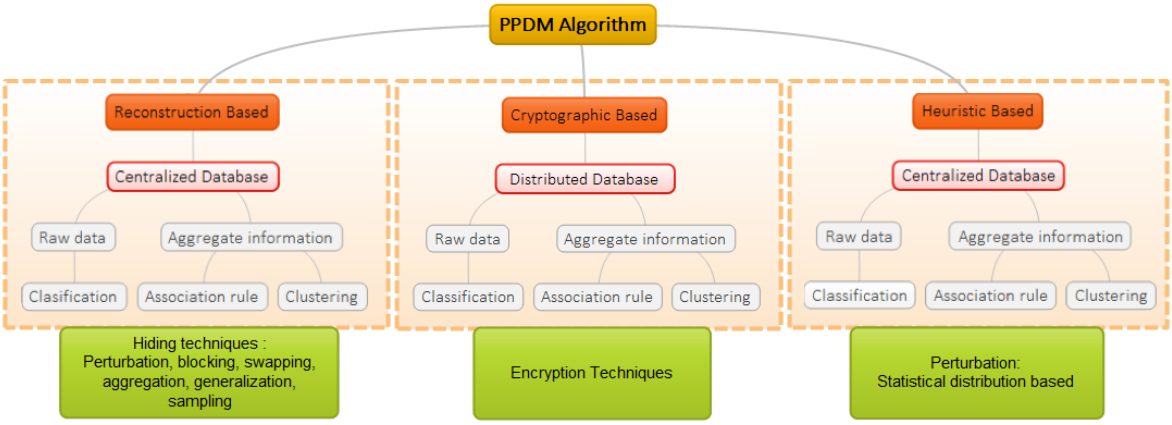

Figure 1. Taxonomy of PPDM techniques

To mitigate the possibility of such breaches a solution called privacy preserving data mining (PPDM) has been developed. Since the last decade, researchers have been developing various privacy protection methods to limit sensitive information leakage by performing data sanitizing into databases. Therefore, prior to sharing or handling a database with other parties, the data owners are encouraged to apply privacy preserving data mining (PPDM) algorithms with respect to balancing the trade-off between privacy and data utility. In the context of PPDM, database owner knows in advance the types of data mining tasks performed by data miner.

There are three requirements that should be satisfied to design PPDM algorithms. The first and important requirement is the entire sensitive values or sensitive itemsets cannot be mined in the sanitized database. While the second requirement is non-sensitive values or itemsets in the original database should also be mined from the sanitized database. The last, the difference between the original database and the sanitized database should be minimized. Obtaining a sanitized database that achieves those three requirements is a very difficult problem, and actually, it has been proved that the problem is NP-Hard [6]. Therefore, various techniques with various settings have been proposed to balance the trade-off and satisfy the requirements of database owners. Since the pioneering work in [7], [8], and [9], several approaches have been proposed in the PPDM area to deal with privacy in data mining.

Realizing the fact that various PPDM methods have been proposed, database owners and researchers who newly study this field may find difficulties to determine which method is the most suitable to be applied for protecting privacy in their database and from which point a new research opportunity should be explored.

In this paper, we classify the PPDM methods based on their taxonomy techniques to provide an overall view of current PPDM methods which is distinct from several existing review papers that specifically discuss privacy preserving association rule mining (PPRAM) [10], the trade-off between privacy and data utility [11] and describe broader privacy protection technology for data mining and data publishing [12]. In addition to that, we highlight the strategy of the current PPDM methods as described in Table 2 and present advantages, challenges, and future development of the PPDM methods. An important feature that should not be disregard is the evaluation metrices that assess the quality and the performance PPDM methods.

II. PRIVACY PRESERVING PROBLEM

Prior to describing the classification of the PPDM techniques, it is important to highlight the intuition of why the PPDM techniques need to be developed. In general, databases have several attributes that can be distinguished into three different types such as key attributes, public attributes, and private or sensitive attributes [13]. The key attribute contains information that can be used to identify individuals, for example, user id, customer id, or individuals name. The second attribute holds information accessible to authorized people. Also, this attribute may lead to an individual's privacy breach if not adequately preserved. The last attribute is the attribute that conserves sensitive information and it should be well protected.

Let us consider a tabular database such as in Table 1 which contains several records. The table consists of several attributes that can be categorized into those three. The key attribute of the database is name in which this value directly refers to individuals known as an identifiable attribute (IA), while birth date, zip code, and occupation are the public attributes that refer to quasi-identifier attribute (QA). The disease attribute is a sensitive attribute (SA) of the data table and thus it should be protected.

The sensitive information breach is possible if the one who holds quasi-identifier attributes has a function for constructing logical information to infer sensitive information of an individual through data mining tools. Therefore, PPDM investigates the side effects of data mining methods that originate from the penetration into the privacy of individuals and organizations[14]. Accordingly, to design a PPDM method which can modify databases in such a way data miner could not infer individual sensitive information one should consider various a trade-off between privacy and utility.

TABLE 2. PPDM METHODS AND THEIR STRATEGY

Classification	Method	Strategy
Reconstruction	Additive noise	Modeling noise addition
	Microaggregation	Replacing original values with an aggregate value
	Swapping	Swapping values among records
	Random noise	Generating random value as a noise
Cryptographic	Secure Multiparty Computing	Semi-honest protocol
	Homomorphic encryption	Encryption
Heuristic	Hiding sensitive items	Replacing sensitive items with non-sensitive items
	Item grouping	Generating an identical cluster with the same sub-itemset

The problem which occurs due to the leak of confidential information is referred to as database inference [15]. Additionally, the PPDM should also be able to preserve similar data utility in the sanitized database like that of the original one.

III. CLASSIFICATION OF PPDM ALGORITHM

Currently, various PPDM techniques have been proposed. The proposed algorithms can be categorized into three different groups based on the taxonomy techniques namely reconstruction technique, cryptographic based technique, and heuristic-based technique. The taxonomy techniques are represented in Figure 1 while the strategy of the techniques is described in Table 2.

A. Reconstruction-based Techniques

Reconstruction-based techniques as depicted in Figure 1 rely on perturbing original values so that an adversarial data miner could not find the original values and the perturbed database maintains its statistical properties. The method perturbs databases and reconstructs their data distribution in the aggregate level to estimate the probability distribution of original values as a result the databases' statistical properties do not deviate drastically from that of the original one.

The main idea of data perturbation is delivering a modified database or sanitized database with additional noise that does not result in significant difference from the original data mining results. This method achieves privacy protection by modifying attributes value from a database, such that private value cannot be reconstructed or disclosed. A simple illustration of perturbation is, for example, a database owner considers an attribute says the customers' salary is sensitive, then he can decide how much noise to add to the real value so that the real value cannot be revealed. The amount of noise depends on the data owner view, it can be generated randomly under certain probability distribution.

There are three types of perturbation techniques in PPDM, such as additive noise, microaggregation, and rank swapping.

1) Additive Noise

As it is indicated by the name, additive noise hides sensitive information by adding some values in a data record or adding artificial records in a database. An idea of using additive noise to sanitize a database in privacy preserving sensitive frequent itemset mining for transaction data has been proposed in [16]. The proposed method appends some artificial transactions into the original database. Initially, the method is calculating the number of transactions that should be added in the database, this is called maximum safety bound (MSB). Equation (1) represents the computation of MSB.

$$\max(SB_i) = \left[\frac{|si_i|}{a} - m\right] + 1 \qquad (1)$$

The notation $\max(SB_i)$ refers to the safety bound of each sensitive itemset, while m is the number of records contained in database D and $|si_i|$ represents the count of a sensitive itemset in the database D. Once the $\max(SB_i)$ is determined, the next step is counting the number of items for each additional transaction p_n, based on the standard normal distribution.

Another method for hiding frequent sensitive itemset has been proposed in [17]. The proposed method protects sensitive frequent itemset by inserting noisy items in certain transactions. The noisy items are selected based on queue and random number generator. Moreover, if a transaction has more items then the more noisy items are generated and added in the transaction. Adding some itemsets or artificial transaction records in a database successfully protects frequent sensitive items since it cannot be mined under the same defined support.

The critical side effect of performing item insertion in transactions is it causes significant distortion on item correlation. Consequently, when a certain data mining process such as frequent itemset mining is performed to the sanitized database, the mining result significantly deviates from that of the original one.

Aiming to protect individual privacy in numerical data, [18] proposed *individually adaptable two-phase perturbation method*. In this method, individuals are granted permission to choose their privacy level. The method firstly perturbs original database values using random values generated from an independent identically distributed random variable. Following that, it splits perturbed data into several predetermined intervals. The user then chooses any part of the split and chooses a privacy level e.g. top, high, medium, and low. After selecting the options, it adopts an interval length that corresponds to the selected privacy level, and

finally, generating the perturbed values by sampling the interval uniformly. These values are then dispatched to the data miner.

Since the method relies on users to perturb their original value, the process needs certain secure computation prior to submitting the value. Thus, additional computation resources are needed to support this process.

An idea called *select-a-size* has been proposed in [19] to sanitize a database in privacy preserving association rules mining for transaction data. The proposed method employs uniform randomization to generate random itemset of a transaction. The modified itemset is sent to the server and the server collects the statistical value of the modified transaction. Even though the algorithm is effective in protecting sensitive information, it takes significant computation cost since the method uses a per-transaction strategy which recursively computes the random value.

Performing noise addition by inserting artificial transaction records causes an increase in the database's size. Accordingly, when one performs certain data mining analyses, the data mining resulted from the modified database is significantly different compared to that of the original one. Moreover, when new records are inserted, some item correlations in the modified database are also changing.

Adding noise into the database might be an effective way to guarantee privacy protection in the data mining process. However, we should carefully decide the amount of noise and the strategy to generate the noise since the quality of the sanitized database depends on it.

2) Microaggregation

The underlying concept of microaggretation is releasing a database with continuous values for a data mining task i.e. clustering, where the original values are replaced with values generated from small original values aggregates.

To generate microaggregation from a database we firstly define a number of groups *g*, each group contains at least *k* records. The next phase is calculating the average value for each attribute data for each group and then replace its original averaged values with the average value from each group. The challenge in microagregation is finding optimal *k*-partition. It should maximize homogeneity values within a group to reduce information loss.

In the case when a database contains several attributes, the microaggregation can be performed to aggregate all the data in all attributes or it can also be performed by dividing the attributes into several groups. One of the microaggregation methods that are commonly used is called Maximum Distance to Average Vector (MDAV) [20]–[23]. Described in [24], MDAV consists of six sequential processes to generate a microaggregate database. The steps are described as follows.

Essentially, the MDAV creates a group containing a number of records that contain sensitive values. To form groups of records, MDAV calculates the average value of the records in a database and relies on distance measurements such as Euclidean distance to determine which record should be included in the group. If there are 2 *k* records not belonging to any group, it generates a new group containing those records. At the final step, to hide the sensitive values the method performs an aggregate function of all the values in a group and replaces the original values with the value obtained from the aggregation function.

Microaggregation has been successfully implemented for protecting privacy in query logs which satisfies the *k*-anonymity concept [25]. Since the query logs contain various attributes such as query terms, timestamps, domain name, some distance measurements such as Euclidean distance, Levenshtein distance, and Hausdorff distance are used to calculate aggregation of those attributes values. Several evaluations conducted by [26] indicate that the microaggregation method is suitable for nominal data from a semantic aspect.

Since the classical MDAV method repeatedly computes distance among groups record, it causes significant time computation. Therefore, an improvement method called Fast MDAV (F-MDAV) has been proposed in [27] to speed up the performance of the classical MDAV. The method uses an algebraic approach to compute distance and reuse the result to avoid redundant operations. As a result, the computation time can be reduced.

3) Swapping

The main idea of data swapping is exchanging sensitive values of a record to another record while at the same time maintaining frequency counts. Originally data swapping is developed to protect continuous and categorical values. Data swapping firstly introduced in [28], [29] to protect a database from statistical disclosure. However, since the method does not regard the range value, it increases data utility loss [30].

One variant of data swapping called *rank swapping* hides sensitive values in categorical data and it is successfully implemented for numerical data or ordinal categorical data [31]. Initially, the *rank swapping* determines a value *p* then an attribute value of a database is ranked in ascending order. The next step is randomly selecting the attribute value and swap the value with that of another record while ensuring that they should not differ by more than *p*% of the total number of records in the database.

The rank swapping works well for numerical and categorical data, unfortunately, this method is not suitable for nominal data [32]. Even so, several empirical studies such as in [33] argue that rank swapping results in a balanced trade-off between information loss and disclosure risk.

A recent study in [34] employed swapping techniques namely *partSwap* and *fullSwap* to hide personal tendency of individuals in a set-valued database i.e. transactional database and web click data. The proposed scheme modifies the classical swapping technique by considering values swapping of two records that have a large distance. It uses Euclidean distance measurement to compute the distance between the two records. The intuition of swapping items from two records having a large distance is to avoid an

attacker from inferring and predicting the tendency of individuals in the database.

Experimental results using various types of set-valued databases show that the method works well for hiding personal tendencies. However, due to the nature of data swapping, it causes distortion regarding item correlation in the database.

Overall, perturbation-based techniques can be applied in various database types and provide a strong guarantee in preserving original values from adversarial data miners. However, some side effects occur when such techniques are applied, for example, the data truthfulness is no longer hold. In some critical databases such as health record databases, such a side effect may not be tolerated since it may threaten people's life.

4) Random Noise

Randomization is closely related to perturbation-based technique since most of the perturbation techniques especially for numerical database generates additional noise using some randomization methods. To achieve a sanitized database that protects privacy in the database, the additional noise should be defined carefully to preserve the probability distribution of the database. If we consider *X* is an original database, *Y* is noise and *Z* is a sanitized database, then to generate *Z* we straightforwardly compute $Z = X + Y$.

Referring to [35] the general concept of randomization can be reflected in several points. Initially, we can assume a database D contains a set of records *X*, where $X = \{x_1,...x_N\}$, *N* equals |D|. For each record x_i, a noise y_i is added. The added noise $(y_1, \dots, y_N)$ is generated independently to result in a distorted database. Thus, the database D contains values $Z = x_1+y_1,\dots, x_N + y_n$.

In fact, other than additive strategy randomization there is another variant called multiplicative strategy. Random values for multiplicative strategy also can be generated randomly. Another interesting part is randomization can be applied in the data collection process so that it is not necessary to use a trusted machine for performing data transformations. One important distinct point between additive and multiplicative randomization is that in additive randomization the original aggregate distribution can be reconstructed while in multiplicative randomization not only aggregate distribution that can be reconstructed but also more specific information such as distance between the original value and modified value can be preserved.

Randomization is suitable for data collection scenario, since the collected data can be released for various analysis without concerning further privacy breach [36]. Moreover, [37] states that randomization has efficient computation cost since the noise is randomly obtained from the standard probability distribution. Randomization techniques have been applied in various applications such as Online Analytical Processing (OLAP) and Singular Value Decomposition (SVD)-based collaborative filtering [38].

Performing a randomization strategy needs consideration of balancing the trade-off between privacy and utility of a database. Therefore, an alternative solution to achieve good quality sanitized database by randomization is generating conditional noise that fits to values in the database. The term conditional refers to flexibility in modeling randomization process. Thus, there are still challenges to design a randomization technique that can preserve privacy and retain useful information.

B. Cryptographic based techniques

Different from the previous one cryptographic based technique takes part in securing sensitive information from a data mining task under a distributed computing system. In general, cryptographic techniques such as homomorphic encryption and secure multiparty computation are used in this technique.

Cryptography is a very dynamic research field in computer science and mathematics. Presently, a lot of cryptographic techniques have been developed and successfully implemented in various areas of computer science including PPDM. Therefore, it is not surprising that this field attracts many researchers to utilize cryptographic based techniques to design PPDM methods. Referring to Figure 1, the scenario of cryptographic based techniques always relates to distributed system to ensure that the computation of various data analysis can be protected from adversarial nodes.

1) Secure Multiparty Computation

One of the most used techniques in PPDM which considers distributed system scenario is *Secure Multiparty Computation* (SMC). In this scenario, a data owner wants other partners to perform computation over a database without revealing any private data in the database. Generally, SMC-based approaches consider a semi-honest model where all involved parties permanently obey the protocol.

Pioneering work that employed cryptographic technique in a set-valued database or transactional database for association rule mining which horizontally partitioned proposed in [39]. The proposed technique consists of five steps. First, all the involved parties should encrypt their itemsets using commutative encryption schemes. Second, each party exchanges its encrypted itemsets to another party. The party who received the encrypted itemsets should re-encrypt it. Third, one party sends a token to another party. The token contains item frequency count and a random variable to its neighbor. Forth, the neighbor then adds its item frequency count and sends back the token to its party. The last is comparing between the initiating party and the final one to know whether the final result is higher than the defined threshold and its random value.

This type of scheme has been applied in health care database [40] in a vertical format. However, in this research, the number of collaborators is very limited. Therefore, further investigation by adding more collaborators is still needed to evaluate the efficiency of the proposed method.

Another research in the same task has also been proposed in [41]. The technique is implemented to preserve privacy for *k*-means clustering task over a vertically partitioned database. In every procedure of the

clustering process, each data point is successfully and securely computed to finds the smallest distance to its cluster center and mean value.

As the SMC involves a lot of parties to perform computation it results in high computational cost [6]. Thus, a method in [42] tries to reduce the drawback by introducing an outsourcing strategy. The method aims to enforce privacy preserving in clustering scenario by proposing weighted average protocol as a result the computational cost can be minimized. Computation performance is very crucial in SMC since it always involves a number of users. Therefore, designing SMC-based PPDM methods that can reduce the computation cost is still an open issue.

2) Homomorphic Encryption

Another research for protecting private information when a linear regression model is performed in a database has been proposed in [43]. The method used fully homomorphic encryption schemes and assumed that all involved partners are semi-honest. Each independent attribute in the database is held by different individuals.

To encrypt plaintext *x* from attribute values, the method needs to compute a function $f : f(x) = (x + rp)\, c^r \% n$, where c and r are the number of blocks that need to be checked, $c \leq n$ and r refers to the pseudo-random key of the function f, respectively [44]. While to decrypt the ciphertext *y*, it needs to perform $x = f^{-1}(y) = (b^{tr})^{-1}\, y \,\%\, p$, where $n = p$ x q, both p and q are prime, while b is a primitive root mod $n * c$. In addition, r is any positive integer less than n, while t is a discrete logarithm, $b^t \equiv c \,(mod\ n)$.

The following steps are the ways of the proposed algorithm achieve privacy preserving linear regression model:

1) A key generator (KG) sends public encryption key and different private encryption keys to each partner.
2) Each partner performs encryption to their original data and sent the ciphertext result to the data miner. Since data miner does not have private decryption key, the plaintext cannot be obtained.
3) Data miner performs a calculation to generate encrypted coefficient correlation value E_β.
4) KG decrypts the coefficient correlation value to obtain the regression coefficient correlation value β.

Homomorphic encryption can also be implemented in a set-valued database to find frequent association rules [45], [46]. The scenario in that research assumes that two parties own horizontally partitioned database D_A and D_B, they want to determine the interesting association rules from the combination of their database $D = \{D_A \cup D_B\}$ without compromising individual sensitive condition. All the parties have their secret number, later it will be used to encrypt the number of their frequent itemset. The first stage is each party A and B determine their global frequent itemset L_g based on the given minimum support s. To determine whether itemset from A and that from B are frequent, each party have to send the item counts to another party in the following way:

1) A sends its itemset count c_A and $|D_A|$ to B.
2) B also sends its itemset count c_B and $|D_B|$ to A.
3) Both A and B privately compute whether the itemsets are frequent using the following equation (2).

$$\frac{c_A + c_B}{|D_A + D_B|} \geq \frac{s}{100} \qquad (2)$$

Once each party determined its frequent itemset, then both parties should generate their global frequent itemset L_g. The last step is generating association rule from L_g using the given minimum confidence threshold c. Both parties should split each of their frequent itemset into two parts to generate all possible combinations of association rules from each itemset.

In the area of data classification, a method based on homomorphic encryption has been proposed in [47]. The proposed method is designed to solve multi-label classification and employing Paillier cryptosystem to encrypt and decrypt class label. According to the experimental results both theoretical and simulation show that the computation cost is low since all the processes of encrypt and decrypt are conducted in cloud servers that normally have high specification and cost. We should also note that the Paillier cryptosystem works only for non-negative integers, while in real application many data labels are in real number format which may limit the performance of the method.

Even though the cryptographic schemes are quite promising for guaranteeing privacy protection in PPDM, it is still challenging to be implemented in a real situation since databases for data mining processes usually have very large size which may result in high computation costs and time-consuming processes. In addition, we should also consider the selection of cryptosystem schemes since different cryptosystem may have different limitations in real-world applications. Moreover, the trend of cryptosystem technology no longer relies on traditional techniques, instead it is now moving to quantum cryptosystem which is predicted to be the future technology in computer security. Therefore, designing PPDM schemes that utilize quantum cryptosystem techniques is promising to secure the mining process.

C. Heuristic Based Techniques

Achieving a sanitized database with respect to preserve maximum data utility and maximum privacy protection is a hard problem [47]. Therefore, a lot of heuristic approaches have been proposed to generate sanitized databases with acceptable privacy protection as well as retain enough data utility.

Aiming to hide sensitive frequent itemset in transactional database several methods proposed heuristic solution. An initial work called *Sanit* which follows a heuristic algorithm has been proposed in [46].

TABLE 3. THE PROS AND CONS OF THE PPDM METHODS

Method	Pros	Cons
Additive noise	+ Provide a strong privacy guarantee + Statistical properties can be preserved + Can be applied to various types of databases	– Data truthfulness may decrease – May not be suitable for health database – Artificial items/records reduce data utility – Artificial items/records distort database properties such as data size and the number of items/records in the sanitized database
Microaggregation	+ Sensitive attribute values can be preserved + The data values semantically consistent + Due to no artificial value/records inserted so that the data properties such as data size and the number of records in the database remain the same	– Suitable only for databases with continuous values – Difficult to find an appropriate amount of noise due to it relies on certain aggregation function of the attribute values
Swapping	+ Individual privacy can be preserved + No additional or artificial item/records + Statistical properties can be preserved	– Item correlation is distorted – Random swapping may cause false sensitive attribute correlation – Due to correlation distortion, data mining results such as frequent itemset mining and association rule mining of the sanitized database may differ significantly from that of the original one
Random noise	+ Suitable for data collection phase + Provides efficient computation cost + Can be applied into various types of databases + Widely used to protect privacy in OLAP + Flexible to determine the amount of added noise	– Difficult to generate conditional noise that fits to values in the database – The flexibility leads to significant data distortion due to different amount of added noise in the data values – Needs additional difficult work to prepare a random number generator which tempers to entropy flow attack
Secure Multiparty Computing	+ Provide stronger privacy protection due to the encryption process + Can be applied in various data types	– Computation extensive due to the encrypt and decrypt process especially for big data size – Only suitable for distributed scenario – Needs additional difficult work to prepare a random number generator which tempers to entropy flow attack
Homomorphic encryption	+ Provide stronger privacy protection due to the encryption process + Can be applied in various data types	– Computation extensive due to the encrypt and decrypt process especially for big data size – Needs additional difficult work to prepare a random number generator which tempers to entropy flow attack
Heuristics	+ Can be applied in various database + Usually applicable in many real cases + Can be combined with other methods to achieve better privacy protection	– The results may not the optimal one – Some heuristics approach needs significant computational cost

The proposed method is specifically designed to protect sensitive items in a transactional database. Initially, *Sanit* generates a sorted graph of a frequent itemset in descending order based on the items' support value. Some sensitive items are then omitted from several records while minimizing a side effect such as frequent itemset lost.

Different from the previous one, [48] proposed a heuristic method called Item Grouping Algorithm (IGA). IGA groups itemsets into several identical clusters where each itemset cluster shares the same sub-itemset. By performing such grouping IGA could assign a victim item in each group. If there are overlapping items among the groups, all the itemsets in overlappingareas will be removed, as a result, each group only holds their distinct itemsets. Experimental results show that IGA successfully reduces misses cost which means sensitive itemset cannot be mined from a sanitized database while non-sensitive itemset can still be mined in a sanitized database.

Another heuristic technique for hiding sensitive frequent itemset also has been proposed in [49] namely Maximum Item Conflict First (MICF). The method can achieve a sanitized database by removing sensitive items so that it reduces the support value of the sensitive items. There are several main steps in MICF such as identifying sensitive transactions or records (transaction containing sensitive itemset) from a database and determine a part of the transactions to be sanitized. For each sensitive transaction, it decides an item to be removed, called victim item and perform data modification. The data modification result is re-written in memory as a sanitized database.

The proposed method in [50] assumes that data owner has ability to determine sensitive items in a database and define a support threshold for frequent

itemset mining tasks. The hiding strategy firstly scans all records in the database then records that contain sensitive items are subjected to be modified while other records without any sensitive items are kept as is. To determine which sensitive items in a record that should be removed, they propose *degree sensitivity*, *si* as a boundary. Thus, any sensitive items which occur more than the *si* value will be omitted from the transaction.

A method based on heuristic solution [45] has also been proposed in [51] to solve hiding sensitive itemset problem. The method uses evolutionary multi-objective optimization to selectively find records and insert certain items into the selected records. This strategy allows the support of sensitive frequent itemsets decreasing and thus when the mining process is conducted those itemsets cannot be mined from the database.

In a situation where achieving an exact result is NP-hard, heuristic approach is an alternative solution to generate a database that protects privacy and maintain data utility. Although the results might not be optimal, and may not be correct, it usually applicable in a real situation [48].

Each proposed model may have some advantages and disadvantages to deal with privacy protection and database utility preservation. To sum up the pros and cons of the methods we describe it briefly in Table 3.

IV. MEASUREMENT

Generating a sanitized database which achieves maximum privacy protection and maintains data utility for knowledge discovery is an NP-hard problem. Therefore, various techniques have been proposed since the last decades in which it uses various measurement strategies to evaluate the result of those methods.

A. Privacy protection measurement

To measure the privacy protection over a sanitized database, [52] proposed a quantification metric for perturbation based technique. that is if a perturbed value can be estimated under a confidence level c % which belongs to an interval $[x_{1,2}]$, then we can estimate the privacy by subtracting x_1 to x_2 with the confidence c %.

Measuring privacy in multiplicative random noise has also been described in [53]. In this measurement, it assumes that if x_1 is an original attribute value and x_2 is the distorted value of the x_1, we can estimate the original value using the following equation (3).

$$\frac{\mathrm{Var}(x_1 - x_2)}{\mathrm{Var}(x_1)} \quad (3)$$

Another important privacy measurement is called *hiding failure* (*HF*) which is firstly introduced in [54]. This measurement plays an important role to quantify the balance between privacy and knowledge discovery in a database. The hiding failure calculates ratio between the number of sensitive frequent patterns in a sanitized database that still can be mined $\#P(D')$ and the number of that in the original database $\#P(D)$, the formula of calculating *HF* is described in the following equation. A good data sanitization method would result in a minimum percentage of *HF*. Therefore, since there is a trade-off between privacy and data utility designing a data sanitization method which can minimize *HF* or even zero *HF* is still a challenge. To compute *HF*, one can use the formula in (4).

$$HF = \frac{\#P(D')}{\#P(D)} \quad (4)$$

B. Utility measurement

Measuring data utility in PPDM should also be taken into account since it represents the quality of a sanitized database. It is further acknowledged in [54] that there are no generic measurements to evaluate utility in a sanitized database. Therefore, various data utility measurements have been proposed.

There are two important measurements to quantify data utility in PPDM, the first is called *Misses Cost* (MC) and the second is *Artificial Pattern* (AP). MC refers to the number of non-sensitive patterns that are accidentally hidden due to performing PPDM algorithm. The formula to compute MC is stated in (5), where notations $\#P_{ns}(D)$ and $\#P_{ns}(D')$ denote the number of non-sensitive patterns in an original database and that in a sanitized database, respectively.

$$MC = \frac{\#P_{ns}(D) - \#P_{ns}(D')}{\#P_{ns}(D)} \quad (5)$$

Meanwhile, AP represents the number of artificial pattern that is generated in the sanitized database as seen in (6). Artificial pattern refers to an occurrence of patterns that previously does not exist in the original database but it becomes exist in the sanitized database.

$$AP = \frac{|P(D)| - |P(D) \cap P(D')|}{|P(D)|} \quad (6)$$

A closely related metric with MC and AP namely NTH (not to hidden) has been proposed in [54]. The metric computers the number of non-sensitive itemsets that are accidentally hidden in a sanitized database due to the data sanitization. The NTH formula is denoted in (7).

$$NTH = |FI_s - SI_s - FI_s'| \quad (7)$$

where, SI_s refers to the number of sensitive items that should be hidden from an original database, while FI_s and FI_s' denote the number of sensitive frequent itemsets in an original database and the number of that in the sanitized database.

Achieving the lowest value or even zero of MC and AP is desirable in designing the PPDM algorithm. However, we should note that there is always a trade-off between data utility and privacy.

A measurement called misclassification error (M_E) has also been proposed in [55] to evaluate the quality of sanitized databases for clustering task. M_E is measuring the number of information loss resulted from clustering algorithms. Misclassification error can be computed using the following equation in (8).

$$M_E = \frac{1}{N}\sum_{i=1}^{k}(|C_i(D)| - |C_i(D')|) \quad (8)$$

The notation N refers to the number of points in the original database while k is the number of clusters. $|C_i(D)|$ and $|C_i(D')|$ represent the number of data points in cluster i th from an original database and that in a sanitized database, respectively. Since data sanitation somehow changes the values inside the database, it is important to maintain the consistency of the clustering results.

C. Similarity measurement

Measuring similarity of a sanitized database should also be taken into account since it represents the closeness between an original database and a sanitized database. It is further believed that by knowing the similarity between those to databases, the data owner can avoid disbelief from the database recipients [55].

To measure similarity in the transactional database, [55] proposed dissimilarity measurement ($Diss$). The underlying idea of such measurement is comparing the histogram frequency of items in an original database with that of the sanitized one.

$$Diss = \frac{\sum_{i=1}^{n} |fD(i) - fD^{i}(i)|}{\sum_{i=1}^{n} fD(i)} \quad (9)$$

As described in (9), $fD(i)$ represents the frequency of item i in the original database, whereas $fD^{i}(i)$ refers to the frequency of item i in the sanitized database. It is obvious that dissimilarity between original and sanitized databases should be minimized to provide acceptable data similarity in knowledge discovery process.

V. Research Opportunities and Challenges

In today's era where most people are connected to the Internet and doing their activities via on-line systems in many different ways, individual privacy protection is an essential issue that should be addressed. It is generally known that each individual may have a different concern about privacy, for example, one may think that his political view is sensitive information while some others do not think so. Thus, developing some ideas to guarantee personalized individual privacy while not changing the general data pattern is an interesting issue.

Looking from the fact that generated data in this era, e.g. mobile technology and IoT technology, result in various data format and an abundant amount of data size, designing distributed PPDM algorithms that is resilient to handle very large databases with ensuring its communication security and data integrity will be very prominent in the future.

In the area of cryptographic based PPDM, developing a method which can reduce the computation complexity might be a priority since the main problem of such techniques is the computation performance. In addition to that, the development of quantum cryptosystems as the future backbone of computer security will also lead to enrich PPDM development.

Another crucial part of PPDM development is measurement or metric to evaluate the performance of PPDM methods. It has been stated in [55] that generic measurements are limited to basic statistics such as mean and covariance, while in the real case, the evaluation metrics must be suitable for a specific application. Thus, designing assessment frameworks that can be used to assess various PPDM methods will foster their development with respect to improving privacy guarantee and data utility.

The application of PPDM is also important for practitioners who aim to implement some of the described methods. Practitioners or database owners should consider various aspects prior to implementing the methods. Firstly, they should know what type of database they have such as transactional or relational databases since different databases need different techniques. Secondly, the purpose of data analysis should be determined in advance since different analysis needs a different approach, for example, if they want to protect customers transaction pattern then heuristic approach might be the solution, however, if the aim is protecting certain sensitive attribute of the customers then additive noise methods are suitable. The last is considering the computation resource especially if they want to utilize cryptographic based method to preserve database privacy.

VI. Conclusion

PPDM is one of the field studies in data mining area which aims to protect private information in a database that might leak during the knowledge discovery process. Various PPDM algorithms have been proposed not only to ensure privacy protection but also to maintain data usefulness from a modified database. However, there are still many areas to be explored.

Since each algorithm has its design purpose, none of the proposed algorithms can fit to protect privacy from different mining tasks. The implementation of PPDM algorithms should also consider the type of databases that are used whether it is a statistical database, a categorical database, or a transactional database since different types of databases need different treatments.

Even though some PPDM algorithms seem very promising in protecting privacy and data utility based on its empirical studies, we still need to ensure their applicability and effectiveness with respect to the performance and computation costs because data mining tasks usually involve very large databases.

Ensuring PPDM algorithms results is also another important thing. Thus, various measurement tools have also been suggested to evaluate performance of the PPDM algorithms. However, utilizing one metric is not adequate since there might be multiple parameters in a database that should be evaluated. Moreover, the proposed measurement metrics are application-specific. As a result, it is usually difficult to compare the existing PPDM techniques between one and another.

References

[1] U. M. Fayyad, G. Piatetsky-Shapiro, and P. Smyth, "Advances in Knowledge Discovery and Data Mining," U. M. Fayyad, G. Piatetsky-Shapiro, P. Smyth, and R. Uthurusamy, Eds. Menlo Park, CA, USA: American Association for Artificial Intelligence, 1996, pp. 1–34.

[2] H. Mark, M. Erik, and V. Sunil, *Java Data Mining: Strategy, Standard, and Practice*, 1st Editio. Morgan Kaufmann, 2006.

[3] Merdeka.com, "Mafia Jual Beli Data Pribadi." 2020.

[4] S. Xingzhi and P. S. Yu, "A border-based approach for hiding sensitive frequent itemsets," 2005, doi: 10.1109/ICDM.2005.2.

[5] X. Sun and P. S. Yu, "A Border-Based Approach for Hiding Sensitive Frequent Itemsets," in *Proceedings of the Fifth IEEE International Conference on Data Mining*, 2005, pp. 426–433, doi: 10.1109/ICDM.2005.2.

[6] M. Atallah, A. Elmagarmid, M. Ibrahim, E. Bertino, and V. Verykios, "Disclosure Limitation of Sensitive Rules," in *Proceedings of the 1999 Workshop on Knowledge and Data Engineering Exchange*, 1999, pp. 45--, [Online]. Available: http://dl.acm.org/citation.cfm?id=519168.788219.

[7] R. Agrawal and R. Srikant, "Privacy-preserving Data Mining," in *Proceedings of the 2000 ACM SIGMOD International Conference on Management of Data*, 2000, pp. 439–450, doi: 10.1145/342009.335438.

[8] B. Pinkas, "Cryptographic techniques for privacy-preserving data mining," *ACM SIGKDD Explorations Newsletter*, 2002, doi: 10.1145/772862.772865.

[9] Y. Lindell and B. Pinkas, "Privacy preserving data mining," *Journal of Cryptology*, 2003, doi: 10.1007/s00145-001-0019-2.

[10] V. S. Verykios, A. K. Elmagarmid, E. Bertino, Y. Saygin, and E. Dasseni, "Association Rule Hiding," *IEEE Transactions on Knowledge and Data Engineering*, 2004, doi: 10.1109/TKDE.2004.1269668.

[11] M. B. Malik, M. A. Ghazi, and R. Ali, "Privacy preserving data mining techniques: Current scenario and future prospects," 2012, doi: 10.1109/ICCCT.2012.15.

[12] V. S. Verykios *et al.*, "State-of-the-art in Privacy Preserving Data Mining Classification of Privacy Pre," *ACM SIGMOD Record*, vol. 33, no. 1, pp. 50–57, 2004, doi: 10.1145/974121.974131.

[13] L. Chun-Wei, H. Tzung-Pei, C. Chia-Ching, and W. Shyue-Liang, "A Greedy-based Approach for Hiding Sensitive Itemsets by Transaction Insertion," *Journal of Information Hiding and Multimedia Signal Processing.*, vol. 4, no. 4, pp. 201–2014, 2013.

[14] J.-L. Lin and Y.-W. Cheng, "Privacy Preserving Itemset Mining Through Noisy Items," *Expert Syst. Appl.*, vol. 36, no. 3, pp. 5711–5717, Apr. 2009, doi: 10.1016/j.eswa.2008.06.052.

[15] L. Liu, M. Kantarcioglu, and B. Thuraisingham, "The Applicability of the Perturbation Based Privacy Preserving Data Mining for Real-world Data," *Data Knowl. Eng.*, vol. 65, no. 1, pp. 5–21, Apr. 2008, doi: 10.1016/j.datak.2007.06.011.

[16] 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*, 2002, pp. 217–228, doi: 10.1145/775047.775080.

[17] J. Domingo-Ferrer and V. Torra, "Ordinal, Continuous and Heterogeneous k-Anonymity Through Microaggregation," *Data Min. Knowl. Discov.*, vol. 11, no. 2, pp. 195–212, Sep. 2005, doi: 10.1007/s10618-005-0007-5.

[18] C. C. Aggarwal and P. S. Yu, *Privacy-Preserving Data Mining: Models and Algorithms*, 1st ed. Springer Publishing Company, Incorporated, 2008.

[19] G. Navarro-Arribas, V. Torra, A. Erola, and J. Castellà-Roca, "User k-anonymity for privacy preserving data mining of query logs," *Information Processing & Management*, vol. 48, no. 3, pp. 476–487, 2012, doi: https://doi.org/10.1016/j.ipm.2011.01.004.

[20] S. Martínez, D. Sánchez, and A. Valls, "Semantic adaptive microaggregation of categorical microdata," *Computers and Security*, vol. 31, no. 5, pp. 653–672, 2012, doi: 10.1016/j.cose.2012.04.003.

[21] M. Rodriguez-Garcia, M. Batet, and D. Sánchez, "A semantic framework for noise addition with nominal data," *Knowledge-Based Systems*, 2017, doi: 10.1016/j.knosys.2017.01.032.

[22] M. Batet, A. Erola, D. Sánchez, and J. Castellà-Roca, "Semantic anonymisation of set-valued data," 2014, doi: 10.5220/0004811901020112.

[23] M. Rodriguez-Garcia, M. Batet, and D. Sánchez, "Semantic noise: Privacy-protection of nominal microdata through uncorrelated noise addition," 2016, doi: 10.1109/ICTAI.2015.157.

[24] A. Rodríguez-Hoyos, J. Estrada-Jiménez, D. Rebollo-Monedero, A. M. Mezher, J. Parra-Arnau, and J. Forné, "The Fast Maximum Distance to Average Vector (F-MDAV): An algorithm for k-anonymous microaggregation in big data," *Engineering Applications of Artificial Intelligence*, vol. 90, no. January, p. 103531, 2020, doi: 10.1016/j.engappai.2020.103531.

[25] T. Dalenius and S. P. Reiss, "Data-swapping: A technique for disclosure control," *Journal of Statistical Planning and Inference*, vol. 6, no. 1, pp. 73–85, 1982, doi: https://doi.org/10.1016/0378-3758(82)90058-1.

[26] M. Rodriguez-Garcia, M. Batet, and D. Sánchez, "Utility-preserving privacy protection of nominal data sets via semantic rank swapping," *Information Fusion*, vol. 45, no. February 2018, pp. 282–295, 2019, doi: 10.1016/j.inffus.2018.02.008.

[27] W. E. Winkler, "Masking and Re-identification Methods for Public-Use Microdata: Overview and Research Problems," in *Privacy in Statistical Databases*, 2004, pp. 231–246.

[28] A. Hundepool *et al.*, *Statistical Disclosure Control*. 2012.

[29] J. Domingo-Ferrer, D. Sánchez, and J. Soria-Comas, "Database Anonymization: Privacy Models, Data Utility, and Microaggregation-based Inter-model Connections," *Synthesis Lectures on Information Security, Privacy, and Trust*, 2016, doi: 10.2200/s00690ed1v01y201512spt015.

[30] D. P, lane J.i, T. J.J.M, and Z. L.V, *Confidentiality, disclosure, and data acces: theory and practical applications for statistical agencies*. Amsterdam: Elsevier Science, 2001.

[31] D. Gunawan and M. Mambo, "Data anonymization for hiding personal tendency in set-valued database publication," *Future Internet*, vol. 11, no. 6, 2019, doi: 10.3390/FI11060138.

[32] C. C. Aggarwal and P. S. Yu, "Chapter 2 A General Survey of Privacy-Preserving Data Mining Models and Algorithms," *Privacypreserving data mining*, pp. 11–52, 2008, doi: 10.1007/978-0-387-48533.

[33] Y. Wang, X. Wu, and D. Hu, "Using randomized response for differential privacy preserving data collection," *CEUR Workshop Proceedings*, vol. 1558, 2016.

[34] C. Bettini and D. Riboni, "Privacy protection in pervasive systems: State of the art and technical challenges," *Pervasive and Mobile Computing*, vol. 17, no. PB, pp. 159–174, 2015, doi: 10.1016/j.pmcj.2014.09.010.

[35] Z. Xian, Q. Li, X. Huang, and L. Li, "New SVD-based collaborative filtering algorithms with differential privacy," *Journal of Intelligent and Fuzzy Systems*, vol. 33, no. 4, pp. 2133–2144, 2017, doi: 10.3233/JIFS-162053.

[36] J. Vaidya and C. Clifton, "Privacy Preserving Association Rule Mining in Vertically Partitioned Data," in *Proceedings of the Eighth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2002, pp. 639–644, doi: 10.1145/775047.775142.

[37] N. Domadiya and U. P. Rao, "Privacy Preserving Distributed Association Rule Mining Approach on Vertically Partitioned Healthcare Data," *Procedia Computer Science*, vol. 148, no. Icds 2018, pp. 303–312, 2019, doi: 10.1016/j.procs.2019.01.023.

[38] W. Fang, C. Zhou, and B. Yang, "Privacy preserving linear regression modeling of distributed databases," *Optimization Letters*, vol. 7, no. 4, pp. 807–818, Apr. 2013, doi: 10.1007/s11590-012-0482-8.

[39] B. Denham, R. Pears, and M. A. Naeem, "Enhancing random projection with independent and cumulative additive noise for privacy-preserving data stream mining," *Expert Systems with Applications*, vol. 152, 2020, doi: 10.1016/j.eswa.2020.113380.

[40] Y. Li, Z. L. Jiang, L. Yao, X. Wang, S. M. Yiu, and Z. Huang, "Outsourced privacy-preserving C4.5 decision tree algorithm over horizontally and vertically partitioned dataset among multiple parties," *Cluster Computing*, vol. 22, pp. 1581–1593, 2019, doi: 10.1007/s10586-017-1019-9.

[41] M. G. Kaosar, R. Paulet, and X. Yi, "Fully Homomorphic Encryption Based Two-party Association Rule Mining," *Data Knowl. Eng.*, vol. 76–78, pp. 1–15, Jun. 2012, doi: 10.1016/j.datak.2012.03.003.

[42] Y. Liu, Y. Luo, Y. Zhu, Y. Liu, and X. Li, "Secure multi-label data classification in cloud by additionally homomorphic

encryption," *Information Sciences*, vol. 468, pp. 89–102, 2018, doi: 10.1016/j.ins.2018.07.054.

[43] F. K. Dankar, "Privacy Preserving Linear Regression on Distributed Databases," *Trans. Data Privacy*, vol. 8, no. 1, pp. 3–28, Dec. 2015, [Online]. Available: http://dl.acm.org/citation.cfm?id=2870564.2870566.

[44] J. J. Yang, J. Q. Li, and Y. Niu, "A hybrid solution for privacy preserving medical data sharing in the cloud environment," *Future Generation Computer Systems*, vol. 43–44, pp. 74–86, 2015, doi: 10.1016/j.future.2014.06.004.

[45] S. R. M. Oliveira and O. R. Zaane, "Privacy Preserving Frequent Itemset Mining," *Proceedings of the IEEE international conference on Privacy, security and data mining*, 2002.

[46] D. Gunawan and G. Lee, "Heuristic Approach on Protecting Sensitive Frequent Itemsets in Parallel Computing Environment," in *The 1ST UMM International Conference on Pure and Applied Research (UMM-ICOPAR 2015)*, 2015, pp. 41–49.

[47] Y.-C. Li, J.-S. Yeh, and C.-C. Chang, "MICF: An effective sanitization algorithm for hiding sensitive patterns on data mining," *Advanced Engineering Informatics*, vol. 21, no. 3, pp. 269–280, 2007, doi: 10.1016/j.aei.2006.12.003.

[48] S. R. M. Oliveira and O. R. Zaïane, "Privacy Preserving Clustering By Data Transformation," *Proc. of the 18th Brazilian Symposium on Databases*, pp. 304–318, 2003, doi: 10.1.1.2.42.

[49] P. Cheng, C. W. Lin, and J. S. Pan, "Use HypE to hide association rules by adding items," *PLoS ONE*, 2015, doi: 10.1371/journal.pone.0127834.

[50] L. Zhang, W. Wang, and Y. Zhang, "Privacy Preserving Association Rule Mining: Taxonomy, Techniques, and Metrics," *IEEE Access*, vol. 7, pp. 45032–45047, 2019, doi: 10.1109/ACCESS.2019.2908452.

[51] J. Domingo-Ferrer and V. Torra, "Disclosure risk assessment in statistical data protection," *Journal of Computational and Applied Mathematics*, 2004, doi: 10.1016/S0377-0427(03)00643-5.

[52] D. Gunawan and M. Mambo, "Set-valued data anonymization maintaining data utility and data property," Jan. 2018, doi: 10.1145/3164541.3164583.

[53] J. C. W. Lin, T. P. Hong, P. Fournier-Viger, Q. Liu, J. W. Wong, and J. Zhan, "Efficient hiding of confidential high-utility itemsets with minimal side effects," *Journal of Experimental and Theoretical Artificial Intelligence*, vol. 29, no. 6, pp. 1225–1245, 2017, doi: 10.1080/0952813X.2017.1328462.

[54] S. R. M. Oliveira and O. R. Zaiane, "Privacy preserving frequent itemset mining," *Proceedings of the IEEE international conference on Privacy, security and data mining-Volume 14*, vol. 14, pp. 43–54, 2002, [Online]. Available: http://portal.acm.org/citation.cfm?id=850782.850789.

[55] J. Salas and J. Domingo-Ferrer, "Some Basics on Privacy Techniques, Anonymization and their Big Data Challenges," *Mathematics in Computer Science*, vol. 12, no. 3, pp. 263–274, 2018, doi: 10.1007/s11786-018-0344-6.