IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received March 11, 2019, accepted March 26, 2019, date of publication April 1, 2019, date of current version April 16, 2019.

Digital Object Identifier 10.1109/ACCESS.2019.2908452

Privacy Preserving Association Rule Mining: Taxonomy, Techniques, and Metrics

LILI ZHANG[1,2], WENJIE WANG[3], AND YUQING ZHANG[1,3,4]

[1]National Key Laboratory of Integrated Services Networks, Xidian University, Xi'an 710071, China
[2]Henan Joint International Research Laboratory of Cyberspace Security Applications, Information Engineering College, Henan University of Science and Technology, Luoyang 471023, China
[3]National Computer Network Intrusion Protection Center, University of Chinese Academy of Sciences, Beijing 100049, China
[4]State Key Laboratory of Information Security Institute of Information Engineering, Chinese Academy of Sciences, Beijing 100093, China

Corresponding author: Yuqing Zhang (zhangyq@nipc.org.cn)

This work was supported in part by the National Key Research and Development Program of China under Grant 2016YFB0800703, in part by the Chinese National Natural Science Foundation under Grant U1836210, Grant 61572460, Grant 61772174, and Grant 61370220, in part by the National Information Security Special Projects of the National Development and Reform Commission of China under Grant (2012)1424, and in part by the Open Project Program of the State Key Laboratory of Information Security under Grant 2017-ZD-01.

ABSTRACT In the last decades, pervasive computing is generating a growing quantity of data. Data mining (DM) technology has become increasingly popular. However, the excessive collection and analysis of data may violate the privacy of individuals and organizations, which raises privacy concern. Therefore, a new research area known as privacy-preserving DM (PPDM) has emerged and attracted the attention of many researchers who are interested in preventing privacy disclosure during DM. In this paper, we provide a comprehensive review of studies on a specific PPDM, known as privacy-preserving association rule mining (PPARM). We present a detailed taxonomy for the existing PPARM algorithms according to multiple dimensions and then conduct a survey of the most relevant PPARM techniques from the literature. Moreover, we survey and elaborate on each type of metrics used to evaluate PPARM algorithms. Finally, we summarize some conclusions and come up with some future directions and challenges.

INDEX TERMS Data mining, privacy preserving, association rule mining, association rule hiding, frequent itemsets, privacy metrics, data utility metrics, complexity metrics.

I. INTRODUCTION

With the increasing popularity of ubiquitous computing, data mining (DM) technology has become increasingly popular and may be particularly useful in some applications, such as weather forecast, e-healthcare, risk management [1]. DM can be considered as a particular type of knowledge discovery process. It can be defined as the analysis of a volume of data sets to uncover potentially relevant relations and to summarize these relations in a novel and understandable way. In a broad sense, DM can be divided into two categories: predictive DM and descriptive DM. Predictive DM, such as classification, time sequence analysis, focuses on making predictions by historical data. Descriptive DM, such as clustering, association rule mining (ARM), focuses on uncovering the potential rules hidden in the big dataset without having any predefined target.

Recent advances in DM have caused controversies in the fields of science and technology. On the one hand, DM provides a powerful capacity to discover useful and meaningful knowledge from large amounts of data; while on the other hand the excessive collection and analysis of data may violate the privacy of individuals and organizations, which raises privacy concern. To ensure data security and user privacy [2], Privacy Preserving Data Mining (PPDM) has emerged as an increasingly important problem. As a specific type of PPDM, privacy preserving association rule mining (PPARM) has been widely researched in a myriad of areas, such as market basket analysis, e-health, wireless sensor network, etc. The purpose of PPARM is to find interesting relationships among sets of items in the transaction database while protecting the data security and user privacy.

In recent decades, widespread attention to PPARM from the researchers has led to a great many privacy protection techniques. A great many metrics have been proposed to measure the privacy level, data utility and complexity of the proposed algorithms. Consequently, many reviews with regard to PPARM have been presented.

Most of the PPARM reviews concentrate on the classification and summary of the techniques [3]–[9]. In [3], [4],

The associate editor coordinating the review of this manuscript and approving it for publication was Sedat Akleylek.

the authors conducted a comprehensive survey on algorithms, techniques of association rule hiding. In [5]–[7], the authors took a survey of privacy preserving techniques in ARM over distributed data. In [8], [9], the authors presented a review of the state-of-the-art techniques for PPARM. These works mainly classified the existing algorithms based on the privacy protection techniques. In addition, they all lacked metaheuristic techniques, which are higher level heuristic used in association rule hiding in recent years.

Some review articles focus on the metrics to evaluate the different privacy preserving techniques [10], [11]. Bertino and Fovino [10] only summarized the metrics to evaluate the data hiding algorithms. In [11], Fletcher *et al.* only presented some of the existing metrics utilized in clustering and classification algorithms and lacked the metrics utilized in association rule mining algorithms.

Some across-the-board reviews [9], [12], [13] combine techniques and metrics. However, the survey in [12] mainly focused on the techniques and metrics with respect to the association rule hiding. Navale and Mali [9] conducted the survey on the algorithms of PPARM and the metrics to evaluate the algorithms, they only categorized the existing algorithms based on the privacy preserving techniques. They lacked multidimensional topological structure, moreover, they didn't present the comparison between various methods and comparison of the models; Moreover, they only listed a series of the existing metrics in the literature and did not classify and elaborate on these metrics. Although the author in [13] surveyed privacy preserving data mining techniques. They presented a classification based on the data lifecycle phase. This paper concludes ARM, but not focusing on ARM, therefore, the review of ARM was not comprehensive and lacked metaheuristic-based literature.

We bridge a literature gap by providing an up-to-date and comprehensive review of the existing PPARM techniques and metrics. In this paper, we presented a detailed topological structure of the existing PPARM algorithms, then, we survey and review various PPARM techniques from the literature. In addition, we elaborate on each type of metrics dedicated to PPARM algorithms. Finally, we summarize some conclusions and come with up the future directions and challenges.

The remainder of this paper is organized as follows. In Section II, we provide the preliminaries with respect to association rule mining. Classification and a detailed topological structure of PPARM are presented according to multiple dimensions in Section III. In Section IV, we make a comprehensive review of privacy preserving association rule mining techniques. Section V makes a detailed comparison of each PPARM model. In section VI, we survey and elaborate on each type of metrics dedicated to PPARM algorithms. Section VII concludes the paper and Section VIII comes up with some directions and challenges in the future.

II. PRELIMINARIES

Association rule mining Association rule mining algorithms are designed to discover interesting relationships among sets of items in the transaction database. Association rule mining can be stated as follows: Given $I = \{i_1, i_2, \ldots, i_m\}$ is a set of items, $D = \{T_1, T_1, \ldots, T_n\}$ is a set of transactions. Each transaction T is a set of items such that T$\subset$I. An association rule is an implication of the form:X$\Rightarrow$Y, where X$\subset$ Y, Y$\subset$I, X$\cap$Y $= \phi$, both X and Y are itemsets. Support and confidence are two of the most important metrics for evaluating the interestingness of a rule. Support of an association rule X$\Rightarrow$Y is defined as the percentage of records that contain X and Y to the total number of transactions in the database. The confidence of an association rule is defined as the percentage of the number of transactions that contain both X and Y to the total number of transactions that contain X. The support and confidence of a rule can be represented by the following equations.

$$Support(X \Rightarrow Y) = \frac{\sigma(X \cup Y)}{n} \tag{1}$$

$$Confidence(X \Rightarrow Y) = \frac{\sigma(X \cup Y)}{\sigma(X)} \tag{2}$$

where σ(X) denotes the number of transactions that contain the itemset X and n is the total number of transactions. Traditionally, association rules mining can be decomposed into two steps, as shown below:

(1) Identify all the frequent itemsets whose supports are greater than or equal to a user-defined Minimum Support Threshold (MST).

(2) Generate the association rules whose confidences are greater than or equal to a user-defined Minimum Confidence Threshold (MCT).

III. CLASSIFICATION OF THE EXISTING PPARM ALGORITHMS

Many algorithms have been developed to solve the problem of PPARM from different aspects. In this section, we classify PPARM by the following four dimensions: (1) data distribution; (2) the contents that need to be protected; (3) privacy preservation techniques; (4) data modification techniques.

The first dimension refers to the distribution of data. Some approaches are proposed for centralized data. In this case, there is only one data owner who publishes its data to the external site. Other approaches have been designed for distributed data. The distributed data can also be categorized as vertically distributed data and horizontally distributed data. Vertical data distribution means that the number of transactions is the same at each site but the set of attributes is different for all sites. Horizontal data distribution refers to these cases the number of transactions is different at each site but the set of all attributes is the same for all sites.

The second dimension refers to what needs to be protected. Some PPARM algorithms are developed to protect sensitive raw data; some are developed to protect sensitive patterns; however, the others are designed to protect both sensitive raw data and sensitive patterns.

The third dimension refers to the privacy preservation techniques used for protecting the privacy in association

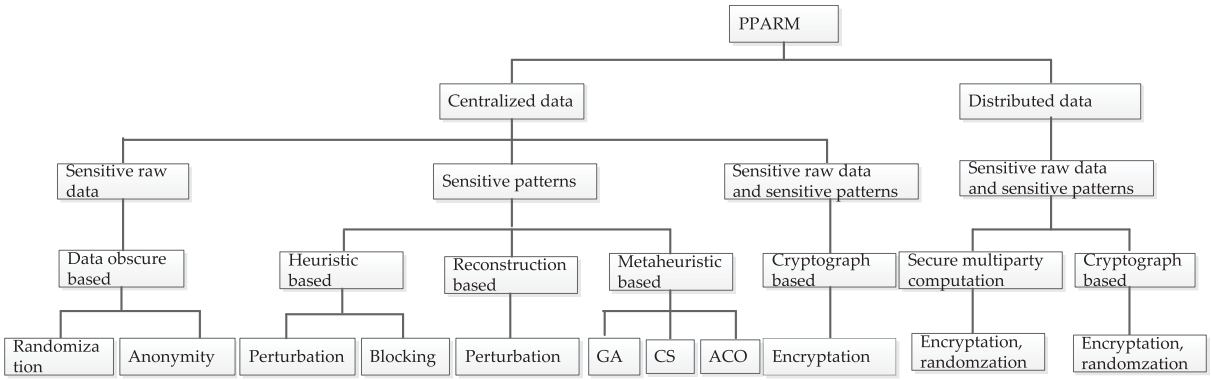

FIGURE 1. A taxonomy of the developed PPARM algorithms.

rule mining. The developed techniques fall into the main six categories.

(1) Data obscure-based techniques. Such techniques can provide privacy protection by adding noise into the original database.

(2) Heuristic-based techniques. Such techniques, such as adaptive modification, modified only selected values rather than all available to minimize the data utility loss.

(3) Reconstruction-based techniques. In such techniques, the original data are not directly sanitized. The main idea of such techniques is as follows: first sanitize the "knowledge base(KB)" into the sanitized knowledge base, denoted as KB'. The released data is then reconstructed from KB'.

(4) Metaheuristic-based techniques. Such techniques are higher level heuristic techniques.

(5) Cryptography-based techniques. Such techniques are often utilized in the scenarios where the association rule mining was outsourced to the cloud server. These outsourced data are centralized or distributed data.

(6) Secure Multiparty Computation (SMC). SMC is often utilized to achieve the association rule mining in the distributed database. SMC aims to mine the global association rules from the joint database without revealing such input to the other parties.

The fourth dimension refers to the data modification methods. When a database needs to be released to the public, it is necessary to modify the original data to ensure high privacy protection. A data modification method is always dependent on the privacy preserving policy adopted in the algorithm. The following data modification techniques are used to modify the original data in the processing of association rule mining.

(1) Data obscure-based methods used for protecting sensitive raw data in association rule mining, such as data randomization and data anonymity.

(2) Hiding methods used for association rule hiding, which include perturbation, blocking, etc. Perturbation refers to the alteration of an attribute value by a new value (i.e., changing 1 to 0, or adding noise). Blocking is the replacement of an existing attribute value with a question mark (i.e.,?).

(3) Intelligent metaheuristic-based modification methods, such as Genetic Algorithm (GA) method and Cuckoo Search (CS) algorithms, Ant Colony algorithm (ACO).

(4) Encryption methods, such as substitution encryption or more complex public-key encryption.

A similar division of PPDM is provided by Verykios and Bertino [14]. However, some important privacy preserving techniques and data modification techniques weren't considered in [14]. We extend the division in [14]. In addition, we present a detailed topological structure with respect to these four dimensions, shown in Figure 1. Assume that PPARM lies at level 0, then the i-th level ($1 \leq i \leq 4$) corresponds to the i-th dimension.

IV. REVIEW OF THE EXISTING PPARM ALGORITHMS

As mentioned in section III, in the existing PPARM algorithms the protected contents can be classified into three categories: (1) protect sensitive raw data; (2) protect sensitive patterns; (3) protect both sensitive raw data and sensitive patterns.

In this section, we regard these three aspects as three privacy protection models and provide a comprehensive review of the existing PPARM algorithms according to these three models.

A. SENSITIVE DATA PROTECTION MODEL (SDPM)

SDPM aims to protect the sensitive data in the original database. In this model, to protect sensitive raw data, the data owner first sanitizes or removes its sensitive data by various methods before publishing his data to the external site. The external site executes the association rule mining over the sanitized database according to the ARM algorithms. SDPM is shown in Figure 2, where D and D' represent the original database and sanitized database respectively, R denotes association rules mined from the sanitized database.

Two well-known data sanitization methods used in SDPM include data randomization and data anonymity. Traditionally, in data randomization approach, normal distribution or Gaussian distribution is used to add or multiply a random noise for every sensitive dataset item. In general,

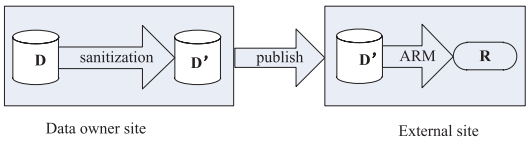

FIGURE 2. Sensitive data protection model.

the data randomization can be classified as additive randomization and multiplicative randomization. In additive randomization,the random noise R is added for every sensitive item in sensitive itemset $S_i = \{S_{i1}, S_{i2}, \ldots S_{in}\}$ as follows.

$$S_{i1} + R_1, S_{i2} + R_2, \ldots, S_{in} + R_n$$

Similarly, in multiplicative randomization,the random noise R is multiplied for every sensitive item in sensitive itemset S_i as follows.

$$S_{i1} \times R_1, S_{i2} \times R_2, \ldots, S_{in} \times R_n$$

In 2002, Rizvi *et al.* expanded the traditional method of randomization for application to the ARM problem and first proposed a randomization-based mask scheme [15] to achieve privacy preserving association rule mining. In addition, they presented an optimization method to decrease high computational overhead in the frequent itemset mining.

Data randomization methods are simple but useful for hiding sensitive information, however, data randomization methods have the risk that public records may be used to discover an identity in the sanitized dataset records. Data anonymity is a raw data protection method which avoids the weakness of randomization methods. Data anonymity aims to prevent the identification disclosure of individual records. K-anonymity, which was introduced by Samarati and Sweeney in 2002. In general, data anonymity functions as follows: (1) remove the record identifier; (2) anonymize the quasi-identifier attributes which can identify the owner of the record. K-anonymity is the most popular anonymity privacy model in which each record is indistinguishable from at least k-1 other records by the quasi-identifier [16]. In the k-anonymity model, k can be utilized to measure the privacy. The larger the value of k is, the more difficult it is to de-anonymize records.

B. SENSITIVE PATTERNS PROTECTION MODEL (SPPM)

The main goal of SPPM is to prevent the disclosure of its sensitive patterns during the process of the association rule mining over the sanitized database at the external site. In this model, as shown in Figure 3, in order to protect the sensitive association rules, the data owner sanitizes the original database D or so-called " knowledge base (KB) "(generated by D) such that certain sensitive association rules can't be discovered through association rule mining techniques. At last, the data owner publishes the sanitized database D' to the external site. At the external size, the server carries the association rule mining over the sanitized data and generate the association rules R', where sensitive rules Rs are

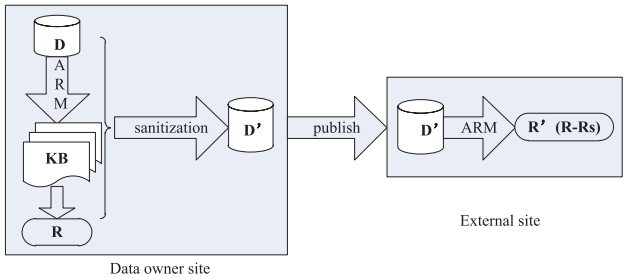

FIGURE 3. Sensitive patterns protection model.

hidden. Sensitive patterns protection model is also known as association rule hiding(ARH) model.

The concept of ARH was proposed firstly in [17]. The problem of ARH is described as follows: let D and D' denotes the original database and sanitized database respectively, R a set of association rules which can be mined from D and Rs a set of sensitive rules in R. The key challenge of ARH is how to transform database D into D' so that all the rules in R can still be mined from D', while the rules in Rs are hidden. In the social network, ARH is often used to protect online social network profiles by hiding sensitive user attributes.

Association rule hiding techniques can be classified into three major categories: heuristic-based techniques, reconstruction-based techniques and metaheuristic-based techniques.

1) HEURISTIC-BASED ARH

A heuristic approach is the supreme used approach for the association rule hiding. "Heuristic" derives from the Greek verb "heuriskein", meaning "to find". In general, heuristics were considered as experience-based algorithms that search the solution space to find a good solution. This approach ignores whether the solution can be proved to be correct, but it usually produces a good solution. According to the data modification strategy, the heuristic-based approaches mainly include two categories: perturbation-based methods and blocking-based methods.

a: PERTURBATION-BASED METHOD

The perturbation-based method is the one widely adopted by the overwhelming majority of researchers in the process of association rule hiding. Perturbation refers to selectively perturbing the values of some attributes, i.e., changing a selected set of 1-values to 0-values so that the support of sensitive rules is lowered in such a way that the utility of the released database is kept to some maximum value. According to hiding strategy, the perturbation-based algorithms can be divided into two categories: itemset-based and sensitive rules-based [4].

Atallah *et al.* [17] first proposed an ARH scheme with the perturbation method. This is an itemset-based scheme. In particular, it selectively hid some frequent itemsets from the original databases by decreasing their supports while having as little as possible effect on other non-selective

frequent itemsets. In [17], Atallah showed that the optimal sanitization was an NP-hard problem. However, this scheme had a disadvantage of hiding only one rule at a time.

Dasseni *et al.* [18] achieved the association rule hiding by sanitizing sensitive rules rather than sanitizing sensitive frequent itemsets. Specifically, they reduced the importance of sensitive rules by making their confidence below a minimum confidence threshold. However, the algorithms in [18] were developed under a strong hypothesis. They assumed that if an item appears in a sensitive rule, then it will not appear in any other sensitive rules. Moreover, their algorithms suffer from two disadvantages: hiding only one rule at a time and generating undesired ghost rules, which reduces the utility of the released database.

On the premise of ensuring security, the heuristic-based algorithms are expected to give attention to the utility issues. A lot of literature considered the balance the privacy and the data utility.

Oliveira and Zaïane [19] proposed a pattern restriction-based algorithms and three item restriction -based algorithms. These algorithms slightly altered the original data while enabling flexibility for someone to tune them. Each algorithm requires two scans of the dataset in order to build the inverted index and alter some sensitive transactions respectively. In this paper, the authors presented some metrics to measure the privacy and utility of the algorithm, such as hiding failure (HF), artifact patterns (AP) and miss cost (MC).

Later, Verykios *et al.* [20] built their schemes on the top of the work [18]. In [20], the authors balanced privacy and disclosure of information by trying to minimize the impact on sanitized transactions. In [21], Modi *et al.* proposed an ARH algorithm named DSRRC (Decrease Support of R.H.S. item of Rule Clusters). This algorithm clustered the sensitive association rules based on R.H.S. of rules and hid as many as possible rules at a time by modifying fewer transactions. DSRRC maintained high data utility because of less modification in the database. Nevertheless, this algorithm generated unwanted side effects. In addition, it couldn't hide the rules with multiple R.H.S. items.

After the literature [21] was published, many subsequent researchers [22], [23] extended and improved the algorithm DSRRC. Komal *et al.* [22] introduced two improved ARH algorithms, named ADSRRC (Advanced DSRRC) and RRLR (Remove and Reinsert L.H.S of Rule). ADSRRC overcame limitation of multiple sorting in the database, as well as it selected transaction to be modified based on different criteria than DSRRC algorithm. The algorithm RRLR could hide association rule with multiple R.H.S items. In addition, the algorithm RRLR is superior to DSRRC algorithm in terms of the number of lost rules and the transactions modified. Nikunj and Ydai [23] modified the algorithm in [22] and proposed a new algorithm, which reduced database modification and side effects by deleting the effective candidate items.

Most of the heuristic ARH algorithms are rule-based or itemset-based. In recent years, some hybrid algorithms [24], [25] are proposed. In [24] Ghalehsefdi proposed a hybrid algorithm to achieve association rule hiding. Soon after, Ghalehsefidi and Dehkordi [25] proposed two hybrid algorithms, named ISSDD (Intelligent Selection of Sanitization in Dense Database) and ISSSD (Intelligent Selection of Sanitization in Sparse Database). Their algorithms were based on rules and items with the least amount of side effect on the sparse and dense database through hiding strategy, compared to the previous algorithm [22], [23],

Cheng *et al.* [26] applied the multi-objective optimization mechanism to consider multiple factors for hiding sensitive itemsets. Later, Cheng *et al.* [27] proposed a deletion method to reduce the support or confidence of sensitive rules below specified thresholds for PPDM.

In 2015, Fouladfar and Dehkordi [28] introduced a quick hiding algorithm of association rules, named FHA. In this method, they reduced the overload of ordering transactions by decreasing database scans. Moreover, they reduced the side effects by selecting the appropriate items for performing the modifications. Conducted experiments indicated that FHA exceeded the algorithms proposed in [21]–[23] in term of computation complexity and the number of lost rules.

Telikani and Shahbahrami [29] achieved the optimization of association rule hiding by combining border and heuristic approaches. In this paper. The MaxMin approach is optimal border-based solution. In combination with two heuristics Telikani utilized MaxMin approach to hide association rules. They proposed a new hybrid algorithm, named as Decrease the Confidence of Rule (DCR).With respect to the rule mining time, the number of lost rules and data utility, The experimental results showed that the performance of DCR is superior to ARHIL algorithm proposed by Hai *et al.*

Surendra and Mohan [30] proposed an ARH method without side effect. The novelty of the proposed method is to sanitize the closed itemsets/patterns instead of transactions in the database. In their scheme, hiding failure is 0, moreover, there are lost rules and ghost rules.

b: BLOCKING-BASED METHOD

A blocking-based method refers to replacing the certain value in a selected transaction with a question mark (i.e.,?) in order to hide the sensitive rules. It is sometimes more desirable for specific applications, such as medical applications, to substitute an unknown value for a real value rather than placing a false value. In what follows, we make a review of the existing blocking-based algorithms.

Saygin *et al.* [31] first proposed a blocking-based ARH approach. They proposed a novel technique for hiding sensitive rules by using unknown values. Introducing a question mark into the dataset changes the definition of the support and confidence of an association rule to some extent. As a result, the support and confidence will be altered into a support interval and a confidence interval correspondingly. They provided a framework for association rules hiding when the data set contained unknown values.

Later, Wang and Jafari [32] proposed a blocking-based sanitization algorithm, which achieved further efficiency

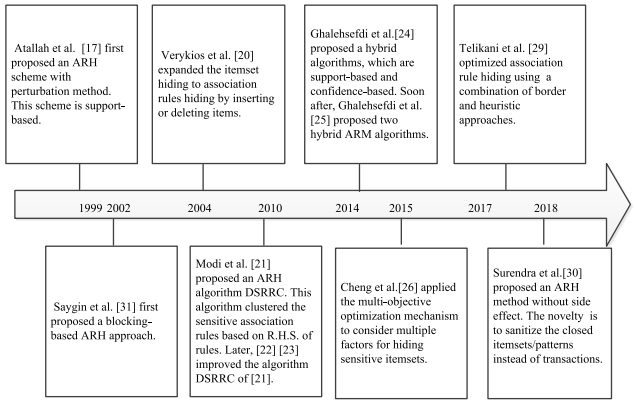

FIGURE 4. A chronological chart about the heuristic-based ARH.

compared to the algorithm in [31]; however, it can only hide all rules containing the hidden items on the left-hand side, while the approach in [31] can hide any specific rule.

Except for the perturbation-based and blocking-based methods, aggregation and disaggregation are also heuristic-based methods often used in ARH. Amiri [33] proposed aggregation and disaggregation sanitization approaches. Aggregation is a similar concept to k-anonimity, where synthetic data is added to the real value instead of generation. In wireless sensor networks(WSN), each sensor is considered a node. Since nodes have low battery capacity, methods to aggregate data from multiple sensors are often used to reduce network traffic. Different from the aggregation approach, the disaggregation method modifies the sensitive transactions in the dataset individually by deleting some items from them until the support of every sensitive itemset is reduced below the minimum support threshold while the number of non-sensitive itemsets whose supports fall below the threshold is minimized.

The classical papers about heuristic-based technique were presented in the chronological chart, as shown in Figure 4.

2) RECONSTRUCTION-BASED ARH

To perform the association rule hiding, many algorithms were proposed to address the problem of privacy preservation by perturbing the data and reconstructing the distributions.

For the first time, Chen *et al.* [34] proposed a reconstruction-based framework, as depicted in Figure 5. The main idea of such techniques is as follows: the data owner firstly find so-called "knowledge base (KB)" (called as frequent itemset in association rule mining) and mined the association rules hidden in the original database, then it sanitizes the KB into KB' such that the sensitive rules Rs can't be mined by the KB'. At last, the data are reconstructed by KB' and released to the external site. In this model, the original data D will not be directly sanitized. In order to hide sensitive frequent itemsets, they proposed a coarse Constraint-based Inverse Itemset Lattice Mining procedure (CIILM). Inspired by the idea in [34], Guo [35] proposed an FP tree based algorithm for inverse frequent itemset mining to

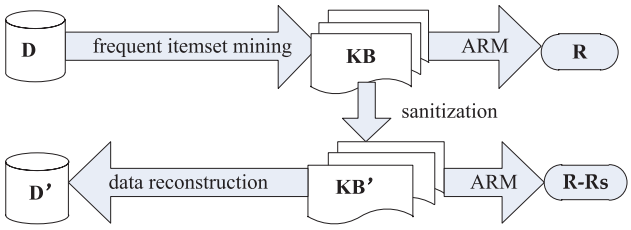

FIGURE 5. Reconstruction-based framework.

reconstruct the original database by using non-characteristic of the database. This algorithm can work more efficiently than CIILM in [34].

3) METAHEURISTIC-BASED ARH

On the premise of ensuring security, the PPARM algorithms are expected to pay attention to the utility issues. It is a challenging problem for traditional PPARH algorithms to minimize side effect in the processing of association rule hiding. Evolutionary algorithms, which are metaheuristic-based solutions, are proposed to find an optimal solution in a short time. Metaheuristics are considered as high-level concepts for exploring search spaces by using different strategies. Evolutionary algorithms, such as Genetic Algorithms (GA), Cuckoo Search (CS) algorithms and Ant Colony (ACO) algorithms are widely used metaheurisics algorithms, which are a type of intelligent algorithms which have a heuristic framework. In recent years, many evolutionary algorithms have been utilized for obtaining the global optimal solution in association rule hiding.

GA is a probabilistic searching algorithm by the Darwinian principle and it transforms an initial population into a new population named as offsprings using crossover and mutation.

Based on the concept of GA and the traditional support and confidence framework, Dehkordi *et al.* [36] introduced a novel algorithm which implemented the distortion in association rule hiding. In [36], they used four fitness strategies for the specification of the fitness function. All of these four strategies are based on the weighted sum function. In this paper, the authors claimed that they minimized the number of lost rules and ghost rules with minimum modification in the original database. However, the number of lost rules and ghost rules are not mentioned.

Khan *et al.* [37] proposed an improved GA architecture with a new fitness function for hiding association rules. In this paper, two databases were used for experimental analysis. The experimental results showed their work has less information loss, lost rules and ghost rules compared to the algorithm in [36].

Lin *et al.* respectively proposed the cpGA2DT [38] and sGA2DT, pGA2DT [39] algorithms for hiding the sensitive itemsets by removing the victim transactions based on GAs. Just recently, based on the NSGAII framework, Lin *et al.* [40] first presented a multi-objective algorithm for hiding the sensitive information with transaction deletion.

CS algorithm is another widely used metaheuristic search algorithm. Be inspired by obligate brood parasitism of some cuckoo species, Yang and Deb [41] first proposed a CS algorithm to effectively solve optimization problems.

Afshari *et al.* [42] presented Cuckoo Optimization Algorithm for Association Rule Hiding (COA4ARH). In this paper, association rule hiding was achieved by the distortion technique. In addition, three fitness functions were defined, which made it possible to achieve a solution with the fewest side effects. A lot of experiments were conducted on the different database and the experimental results showed that COA4ARH hid all association rules and it had fewer side effects, such as lost rules (LR) and ghost rules(GR), compared with the previous algorithms [20], [26], [37].

Doan *et al.* [43] improved the algorithm in [42] to minimize the side effect of the missing non-sensitive rules. Their experimental results indicated that the improved approach had higher performance.

Like GA and CS algorithm, Ant Colony (ACO) algorithms are designed to solve the optimization problems. Dorigo and Gambardella [44] first proposed the Ant Colony System (ACS), which was an extended algorithm of the original ant system.

Narmadha [45] used the ACO solution to hide the sensitive frequent items. Later, some ACO-based approaches [46], [47] were proposed to improve the performance and reduce the side effects.

C. SENSITIVE DATA AND SENSITIVE PATTERNS PROTECTION MODEL (SDSPPM)

SDSPPM aims to protect both the sensitive database and the sensitive patterns, (a.k.a. association rules in ARM). This model is mainly used to solve secure outsourcing of association rule mining over large amounts of data, including centralized data and distributed data, and Secure Multiparty Computation (SMC) problem over distributed data. Researchers have proposed a great many cryptography-based approaches and SMC approaches which are suitable for this model.

1) OUTSOURCING OF SECURE ASSOCIATION RULE MINING

With the development of cloud computing, outsourcing of data mining is acquiring extensive concern. From our survey, outsourcing of association rule mining focus on the outsourcing of association rule mining over the centralized data or distributed data. An overall framework of the outsourcing of association rule mining over centralized data is shown in Figure 6. The data owner encrypts the original database D to transformed database D' and sends D' to the cloud server (a.k.a. miner). The miner computes the encrypted strong association rules AR and sends them the data owner. The data owner then uses a decryption process to convert the encrypted AR to actual association rules AR involving the original items. For example, some hospitals and organizations often outsource association rule mining over e-health records to the cloud for deduced

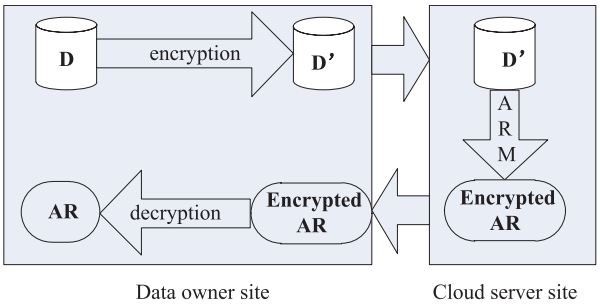

FIGURE 6. Outsourcing of association rule mining over centralized data.

management cost. Privacy preserving problem in outsourcing of frequent itemsets mining and association rule mining has been widely studied in single data owner scenarios [48]–[52].

Wong *et al.* [48] proposed a secure encryption solution based on a 1-to-n item mapping. In their approach, a set of real items are mapped into a set of pseudo items, and then each transaction was transformed in a non-deterministic way. Their scheme was claimed to be resistant to frequency analysis attack; however, it lacked a formal theoretical analysis and proof of security guarantees.

However, Molloy *et al.* [49] indicated that the solution in [48] couldn't counter known frequency analysis attacks. A successful attack on the algorithm in [48] mainly depended on the existence of fake items. In [49], they demonstrated that the random fake items can be removed by detecting the low correlations between items and that the top frequent items can be recognized by attackers.

Tai *et al.* [50] concentrated on outsourcing of frequent itemset mining. In this paper, they introduced k-support anonymity to protect each item with k-1 other items with the same support against the adversary with exact support information. Specifically, items were introduced in the encrypted database to achieve k-support anonymity, and a pseudo taxonomy tree was constructed to ensure the protection of the original items. Giannotti *et al.* [51], [52] also proposed k anonymity-based schemes to achieve the outsourcing of privacy preserving association rule mining. Their schemes ensure that each transformed item is indistinguishable from at least k-1 other transformed items. To counter frequency analysis attack, some fictitious transactions are inserted in the encrypted database to conceal frequencies of the items. After inserting the fictitious transactions, any item in the encrypted database has the same frequency with at least k-1 other items. The data owner sends the encrypted database, including both the real and fictitious transactions, to the cloud server. The cloud server runs a classic frequent itemset mining algorithm and returns the result (frequent itemsets and their supports) to the data owner. The data owner recovers the real supports of these itemsets by subtracting them with occurrence counts of these itemsets in the fictitious transactions respectively. Finally, the data owner decrypts the received itemsets with the revised supports higher than the frequency threshold and generates association rules based on the found frequent itemsets.

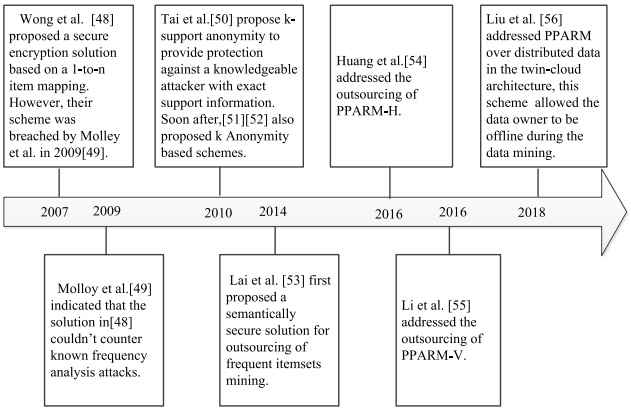

FIGURE 7. A chronological chart about the outsourcing of ARM.

However, none of these works [48]–[52] provide semantic security, which means that the adversary with knowledge on chosen plaintext-ciphertext pairs is able to infer partial information about the raw data and mining results.

Based on predication encryption and dual system encryption, Lai *et al.* [53] first proposed a semantically secure solution for outsourcing of frequent itemsets mining. This solution could resist chosen-plaintext attacks on encrypted items; however, it had no capacity to solve frequency analysis attacks.

The above schemes [49]–[53] only supported the outsourcing of association rule mining over the centralized data owned by a single data owner. In recent years, some researchers focus on the problem of association rule mining over encrypted distributed data [54]–[56].

Huang and Lu [54] proposed an efficient and flexible privacy-preserving mining of association rule algorithm, named as EFPA. This scheme can support the outsourcing of privacy-preserving association rule mining over horizontally distributed data (PPARM-H for short). Li and Lu [55] addressed outsourcing of privacy-preserving association rule mining on vertically partitioned databases (PPARM-V for short). However, these schemes needed the data owner to be online during the data mining. Liu *et al.* [56] addressed PPARM over distributed data in the twin-cloud architecture, this scheme allowed the data owner to be offline during the data mining.

We depict a chronological chart about the development history of the outsourcing of association rule mining, as shown in Figure 7.

2) SECURE ASSOCIATION RULE MINING OVER DISTRIBUTED DATA

There are scenarios where multiple parties wish to collaborate for exacting the interesting global association rules over the conjunction of all partitioned data, without revealing their respective data to each other. Undoubtedly, traditional centralized approaches aren't fundamentally adaptable to these situations. To solve this problem, privacy preserving distributed association rule mining (PPDARM) has emerged as an important research area.

For the first time, Clifton *et al.* [57] proposed the problem of privacy preserving in the distributed data. In this paper, they presented numerous secure multiparty computation (SMC) techniques, such as secure sum, secure set union, secure size of set intersection and scalar product. The proposed techniques are often used as tools in distributed data mining.

In the last decades, many schemes with respect to PPARM over distributed data have been proposed and can be divided into two classes, SMC-based schemes and cryptography-based schemes. SMC aims to mine the global association from the unified database without revealing each party's private data to the other parties. Different from SMC, cryptography-based approaches are suitable for the other cases, where there are multiple data owners and one or multiple third-party servers which are responsible for most of the mining work. In distributed scenarios, there are two main types of data distribution: (1) horizontal distribution; (2) vertical distribution.

a: PRIVACY PRESERVING ASSOCIATION RULE MINING OVER VERTICALLY DISTRIBUTED DATA(PPARM-V)

For privacy preserving association rule mining over vertically distributed data, the main task is to secure computation of the support count of an itemset. If the support count of such an itemset is larger than or equal to the given minimum support threshold, then the miner considers it to be a frequent itemset.

Vaidya and Clifton [58] first addressed the problem of secure association rule mining over vertically partitioned databases. In this paper,to computer the support of itemsets, they proposed a secure scalar product protocol, which was a secure two-party computation protocol. However, the computation overhead of the solution [58] is $\boldsymbol{O}(n^2)$, where n is the number of the transactions. In addition, it can disclose numerous linear combinations of each party's private data to the other parties.

For vertically distributed data, Zhong [59] proposed two algorithms with two levels of privacy, respectively. The algorithm with weak privacy was based on probabilistic public-key encryption, and the one with strong privacy was based on homomorphic encryption. Both of them achieved the privacy guarantee superior to the existing solution. Moreover, they were more efficient than the one in [58] in that their computational complexity is $\boldsymbol{O}(\mathbf{n})$, n is the number of the transactions.

However, Vaidya and Clifton [60] *et al.* adopted a secure set intersection cardinality protocol to enable secure association rule mining over vertically partitioned data. The basic idea of this protocol is as follows: Each party encrypted all their items with communicative public key encryption and passed them to the next party until all parties had encrypted all items. As a result of commutative encryption, the encrypted values from different sets were equal if and only if the original values were the same. Therefore it was needed to only count the number of values that were present in all of the encrypted

itemsets. The set intersection cardinality protocol enabled the anonymity of the ownership of an item set. In addtion, this technique didn't require decryption.

In addition, Ge *et al.* [61] utilized a secret sharing technique to achieve secure association rule mining in multiple parties. In this paper, a collusion-resistant algorithm was proposed, which effectively prevented the collusive behaviors and performed secure multiparty computation.

Similar to [58], Kharat *et al.* [62] utilized the secure scalar product in multiple parties. In respect with the communication cost, the method in [62] exceeded the algorithm in [59], [61]. The communication cost of [62] is N(n-1) + 1, that of [59] is N(n-1) + N + N-1 and that of [61] is 2Nn + n-1. Here N and n represent the number of the transactions and parties respectively.

In both algorithms in [58] and [59], the secure scalar product was achieved by introducing randomness to an input vector, and the final output was retrieved by canceling out the randomness.

Unlike all of the above schemes, both [63] and [55] achieved association rule mining over the vertically distributed database with the assistance of the third-party server. In [63], Rozenberg presented a two-party algorithm and a new version of the multi-party algorithm for finding all frequent itemsets in vertically distributed databases, without revealing their individual transaction values. In what follows, we summarize the main idea of their algorithms: a data owner was designated as the master, which was responsible for the major work of the mining, and the other data owners were considered as slaves. Each slave introduced some fictitious transactions into his own datasets and passed the resulting datasets to the master. Each data owner also sent his set of real transactions' IDs to the third-party server. The master generated all candidate frequent itemsets from the unified database. For each candidate rule $X \Rightarrow Y$, the master sent the ID lists of the transactions containing $X \cup Y$ and the transactions containing X to the third-party server. The server verified if the rule is qualified one. However, this solution suffers from the following disadvantages: (1) The master does the majority of the computation. (2) Though fictitious data are added in datasets, the master is able to learn significant information about other data ownersąŕ raw data from the received datasets.

Li and Lu [55] also provided a cloud-aided frequent itemset mining solution, in which data owners collaboratively mined global association rules from their joint database with the aid of the cloud. To protect the privacy of data, they designed an efficient homomorphic encryption scheme and a secure comparison scheme. Compared with the scheme with strong privacy in [60], which had the same privacy as theirs, their scheme is 3 to 5 orders of magnitude higher. Soon after [55], Hammami *et al.* [64] also used homomorphic encryption to achieve privacy preserving data mining in a cloud computing environment. The processing time consumption of their algorithm is superior to the approach proposed in [65] fitting in the same trend and in the same data characteristics.

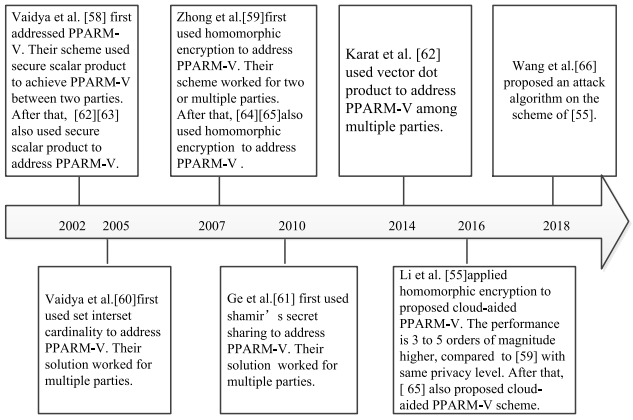

FIGURE 8. A chronological chart about ARM-V.

Wang *et al.* [66] proposed an attack algorithm on the symmetric homomorphic encryption scheme in [55]. In [66], they illustrated li *et al.* overvalued the privacy of their scheme. They showed that they could retrieve the secret key from several known plaintext/ciphertext pairs through the continued fraction algorithm and the Euclidean algorithm.

We depict a chronological chart about the development history of the association rule mining over vertically distributed data, as shown in Figure 8.

b: PRIVACY PRESERVING ASSOCIATION RULE MINING OVER HORIZONTALLY DISTRIBUTED DATA(PPARM-H)

Mining private association rules from horizontally distributed data aims to find all strong associations rules that are hidden in the joint database while minimizing the disclosure of private data at each data owner site.

Kantarcioglu and Clifton [67] first proposed the method for exacting association rule in horizontally partitioned databases. In this paper, Yao's generic secure computation protocol was utilized to securely compute the union of locally large itemsets without revealing the data owners of particular itemsets. Moreover, a novel approach was presented to securely test if the support count exceeded the given threshold.

Schuster *et al.* [68] stated the approaches of [67] are unsuitable for the large-scale systems for the following reasons: (1) A message must traverses through all the parties twice, one-by-one. (2) Slight changes of the data in one party could cause the whole algorithm to have to be executed over again. They proposed a scalable distributed association rule mining algorithm, which was shown to be scalable to millions of resources. Based on additively homomorphic encryption, they presented a cryptographic privacy preserving association rule mining algorithm.

Tassa [69] improved the protocol in [67] in terms of simplicity and efficiency as well as privacy. In [69], they proposed two novel secure protocols. One was utilized to compute the union or of private subsets held by each player and the other was utilized to test the presence of an element held by the client in a subset possessed by another.

TABLE 1. Distributed privacy preserving techniques and the number of data owners included in the system. Note : superscript i represents the ith algorithm of this reference.

Vertically partitioned data	Horizontally partitioned data
Secure scalar product, two parties [58], [63], $[64]^1$	Secure set union, multiple parties($\geq$3)[67-69]
Secure scalar product, multiple parties($\geq$3)[62], $[64]^2$	Homomorphic encryption multiple parties($\geq$3)[68], [73]
Set intersection cardinality, multiple parties($\geq$3) [60]	Elliptic-curve cryptography, multipleparties($\geq$3) $[71]^1$
Shamir's secret sharing, multiple parties($\geq$3) [61]	Shamir's secret sharing, multiple parties($\geq$3) $[71]^2$
Cryptography techniques, two parties $[62]^1$	ECDH+ ECDSA, multiple parties($\geq$3) [72]
Homomorphic encryption, two parties $[62]^2$, [55], [65], [66]	Oblivious transfer protocol multiple parties($\geq$3) [70]

Based on distributed oblivious transfer protocol (OTP), Xie and Zhu [70] presented an algorithm for secure mining of association rules in horizontally distributed databases. This algorithm could not resist the sites collusion attack but failed to prevent the attack from the external adversary.

Chahar *et al.* [71] proposed two protocols to generate global association rule in a horizontally partitioned database. The first protocol utilized Elliptic-curve cryptography (ECC) to address the privacy problem of individual site's information; however, this protocol could not resist the collusion of two sites. The second protocol overcame the weakness by using Shamir's secret sharing, but it required higher communication cost, compared with the first protocol.

All of the schemes assume that the communication channel for data exchange is secure. In practice, this assumption is strong. In [72], Chirag N *et al.* proposed a privacy preserving association rule mining algorithm in horizontally partitioned databases when the communication channel is insecure between involving sites. In this paper, Elliptic Curve-based Diffie-Hellman (ECDH) was utilized to protect the privacy of data at each site, and Elliptic Curve-based Digital Signature Algorithm (ECDSA) was applied to improve the security of the data exchanged among multiple sites.

Different from the above literature, some works addressed the outsourcing problem of association rule mining in the distributed environment, satisfying that the mining process is confidential and the mining results are private.

In such a system structure, there is a third-party server, called the cloud server or master miner. The overall process of such schemes is as follows: to protect the privacy of data, all data owners encrypt their data based on a special encryption algorithm and upload them to the cloud server. When receiving the query of mining global association rules from the data user, the cloud server performs the association rule mining over the encrypted database and obtains the encrypted support and the confidence of each rule.

In [73], the proposed solution was based on homomorphic encryption, specifically, all data owners encrypted their data with the public key of the data user; the cloud server performed the homomorphic computations over the encrypted database. However, in [54], the authors achieved the encryption at the data owner sites by the encryption algorithm proposed in [74]. Each item is one-time masked with a random number.

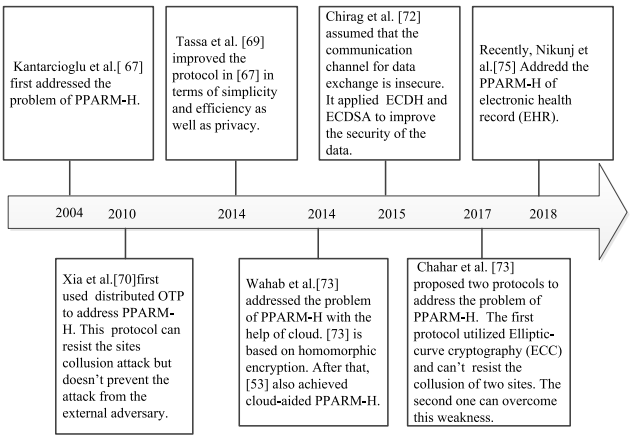

FIGURE 9. A chronological chart about ARM-H.

We depict a chronological chart about the development history of the association rule mining over horizontally distributed data, as shown in Figure 9.

In privacy preserving association rule ming over distributed database, many techniques are utilized. In what follows, we present a brief summary of these techniques and the number of data owners included in the system, as shown in Table 1.

V. COMPARISON OF EACH MODEL

In section IV, we reviewed a variety of PPARM algorithms corresponding to three models. Each model has specific security requirements and application fields. Appropriate selection of the model should depend on the practical requirements.

According to the survey, we make a comparison of these three models from the following six aspects: data distribution, protected content, entities in models, whether ARM is required at the data owner site, overall processing, privacy preserving techniques and data modification methods, as shown in Table 2.

Moreover, we summarize the main advantages as well as the weaknesses of the PPARM techniques, as shown in Table 3.

VI. METRICS FOR QUANTIFYING PPARM ALGORITHMS

The main objective of PPARM is to preserve a certain level of privacy while minimizing at the time the loss of data utility. An important aspect in the design of PPARM algorithms

TABLE 2. Comparison of these three models from six aspects.

	SDPM	SPPM	SDSPPM: Outsourcing of ARM	SDSPPM: SMC
Data distribution	Centralized data	Centralized data	Centralized or distributed data	Distributed data
Protected content	Sensitive data	Sensitive patterns	-Sensitive data -Sensitive patterns	-Sensitive data -Sensitive patterns
Entities in models	-A data owner, -An external miner	-A data owner, -An external miner	-A or multiple data owners, -A cloud server	-Multiple data owners
ARM at the data owner site	Not required	Required	Not required	Not required
Overall process	-Sanitize -Publish -Mine	-Mine -Sanitize -Publish -Mine	-Encrypt -Publish -Request -Mine	-Data sanitization -SMC
Privacy preserving techniques	Data perturbation-based	–Heuristic- based -Reconstruction-based -Metaheuristic-based	Cryptography-based	Secure multiparty computation
Data modification methods	-Data randomization -Data anonymity	-Perturbation -Blocking -Data reconstruction -CS-based -GA-based -ACO-based	-Substitution encryption -Public-key encryption	-Communicative encryption -Homomorphic encryption -Data randomization

TABLE 3. Summary of main advantages and weaknesses of PPARM techniques.

Techniques	Advantages	Weaknesses
Randomization-based	Useful and simple for hiding information	Pubic records may be used to discovered an identifier in sanitized database
Anonymity-based	Hide the identifier of records(transactions) owners	Lose considerable information in the original database
Heuristic:pertuebation-based	More effective, Scalable	Suffer from side effect (ghost rules, lost rules)
Heuristic:blocking-based	Maintain data quality since it just blocks the original value instead of insert false value	Suffer from side effect (ghost rules, lost rules)
Reconstruction-based	Have lesser side effects than heuristic-based approach	The open problem is to restrict the number of transactions in the new database.
SMC-based	Achieve secure mining over distributed data among multiple sites without the third party	High computation time, High communication time
Cryptography-based	Robust, More secure than other types of techniques, Lost rules and ghost rules aren't produced	High execution time for encrypting data, Overall cryptography is a long process, Can't resist frequency analysis attack

is to identify suitable evaluation criteria. In the existing PPARM algorithms, some metrics have been proposed. Based on what aspect of the PPARM algorithms is measured, the existing metrics may be divided into three main classes:

(1) Privacy level metrics, which measure how secure the data is from a disclosure point of view.

(2) Data utility metrics, which quantify the loss of data quality.

(3) Complexity metrics, which are utilized to measure the efficiency and scalability of a proposed algorithm.

In what follows, we survey and elaborate on each type of metrics utilized in PPARM algorithms.

A. PRIVACY METRICS

Generally, the privacy level is measured by the degree of uncertainty. The higher the degree of uncertainty attained by a PPARM algorithm, the better the data privacy is protected by this PPARM algorithm. In the existing PPARM algorithms, the degree of uncertainty is estimated in a variety of ways.

As mentioned in section III, according to the privacy preserving techniques, PPARM algorithms can be classified into data perturbation-based approaches, association rule hiding approaches, and cryptography-based approaches. In the following section, we make a review of the metrics used in each category of privacy preserving techniques.

1) PRIVACY METRICS IN DATA PERTURBATION-BASED TECHNIQUES

In the SDPM model, data randomization is one of the two well-known data perturbation-based techniques. For the first time, Rizvi and Haritsa [15] proposed a randomization-based PPARM scheme. In this paper, they defined user privacy as the following percentage:

$$P(p) = (1 - R(p)) \times 100\% \quad (3)$$

where p refers to the probability the random number r take a value 1 in the processing of randomization, R(p) is the reconstruction probability of the original data from the distorted data.

Data anonymity is another data perturbation technique. K-anonymity was firstly proposed by Sweeney in [16]. If a database is k-anonymous with respect to quasi-identifier attributes, then this database is indistinguishable from at least other transactions by the quasi-identifier attributes. At this time the degree of uncertainty of the quasi-identifier attributes is at least 1/k. Therefore, k-anonymity model has a certain control over the privacy level.

However, both metrics are specific to some privacy preserving techniques.

2) PRIVACY METRICS IN ASSOCIATION RULE HIDING TECHNIQUES

In the association rule hiding model, most of the existing algorithms are designed with the goal of hiding all sensitive patters. However, it is obvious that the more sensitive patterns are hidden, the more non-sensitive information is lost. Thus, some ARH algorithms have been developed, which considers the balance between privacy and data utility. In order to measure the balance between privacy and data utility, Oliveira and Zaiane [19] presented a new and important result privacy metric, named hiding failure (HF). HF is represented as the percentage of sensitive patterns that are discovered from the sanitized database D', to the sensitive patterns found in the original database D.

$$HF = \frac{\#R_P(D')}{\#R_P(D)} \tag{4}$$

where # R_p(D) and # R_p(D') denote the number of sensitive patterns discovered from the original database D and the sanitized database D' respectively. Ideally, HF is 0, meaning all sensitive patterns are successfully hidden. Later, many association rule hiding algorithms [19]–[23], [25], [28], [31], [33], [35], [37], [42]–[43] utilized HF as the result privacy metrics of their schemes.

In addition, Wu *et al.* [47] presented a similar definition to HF, named as F-T-H (failure to hide some sensitive information). F-T-H is defined as the number of sensitive itemsets appearing in the sanitized database D'.

$$F - T - H = \left|SI_S \cap FI_s'\right| \tag{5}$$

where SI_S denotes the set of sensitive itemsets that the data owner wants to hide. FI_S is the set of frequent itemsets in the sanitized database D'.

3) PRIVACY METRICS IN CRYPTOGRAPHY-BASED TECHNIQUES

Generally, the cryptography-based techniques are used in the association rule mining outsourcing and secure multiparty computation (SMC). In association rule mining outsourcing, the data owner applies encryption algorithms to encrypt the original transactions, which are sent to the service provider, i.e., miner. In the SMC model, two or more parties want to conduct a joint computation based on their private data, but neither party is willing to disclose its own output to other parties, therefore, encryption is sometimes required in the SMC model.

In general, cryptography-based methods can achieve a high level of privacy. However, they may be inferior with respect to other metrics like computation complexity that will be discussed below.

B. DATA UTILITY METRICS

In the algorithm based on data obscure-based (randomization-based and k-anonymity-based)algorithms and ARH algorithms, data utility was degraded because original database is sanitized to protect sensitive raw data or association rule. Therefore, data utility was considered as an important aspect in this type of algorithms. Data utility is measured by functionality loss, which quantifies the account of lost information after the application of privacy preserving processing. Various data utility metrics have been proposed, however, currently, no metrics are widely accepted by the research community. Data utility metrics can be categorized into data utility metrics and result utility metrics.

1) DATA UTILITY METRICS

From a data utility point of view, Agrawal and Aggarwal [77] defined the information loss as a metric to measure the accuracy of any reconstruction algorithm. Denoting the original density function of attribute X and the reconstructed density function by f_X(x) and $\hat{f}_X$(x) respectively, the information loss is computed by the following formula.

$$I(f_X(x), \hat{f}_X(x)) = \frac{1}{2}E[\int_{\Omega X} \left|f_X(x) - \hat{f}_X(x)\right| dx] \tag{6}$$

The information loss ranges from 0 and 1. In the best case, I(f_X(x), $\hat{f}_X$(x)) is 0, implying that there is no information loss. I(f_X(x), $\hat{f}_X$(x))is 1 means that there is no overlap between f_X(x) and its estimate $\hat{f}_X$(x).

A data sanitization approach is measured mainly by the capacity to hide sensitive rules or itemsets with minimal effect on data utility of non-sensitive information. In [33], data utility and data accuracy were utilized to measure the performance of the data sanitization approaches. Data utility was defined as the percentage of the non-sensitive itemsets that are not concealed. Amiri defined data accuracy with two levels, named transaction accuracy and item accuracy. Transaction accuracy represents the percentage of transactions that are accurate in the sanitized database. Item accuracy, denoted by IA, can be computed as the percentage of frequencies of the items that are not deleted from the sanitized database.

In some literature [21]–[23], data utility was evaluated by dissimilarity (Diss), which measured the difference between the original and the modified dataset. Assuming f_X(i) represents the frequency of the item i in the dataset X, D represents the original database, D' represents the sanitized database and n represents the number of the items in the original dataset D.

Diss is computed as follows.

$$Diss(D, D') = \frac{1}{\sum_{i=1}^{n} f_D(i)} \times \sum_{i=1}^{n} [f_D(i) - f_{D'}(i)] \quad (7)$$

2) RESULT UTILITY METRICS

Result utility metrics are often based on the comparison between the data mining results over the perturbed data and the original data.

In [15], Rizvi *et al.* utilized support error (ρ) and identify error (σ) as their result utility metrics. Support error reflected the average relative error in the reconstructed support values for all the itemsets which are correctly identified to be frequent. Identify error (σ), which has two components: σ^- and σ^+, referred to the percentage error in identifying frequent itemsets. Support error (ρ) and identify error (σ) can be computed by the following formulas.

$$\rho = \frac{1}{|f|} \sum_f \frac{|rec\_\sup_f - act\_\sup_f|}{|act\_\sup_f|} \times 100 \quad (8)$$

$$\sigma^+ = \frac{|R - F|}{|F|} \times 100, \quad \sigma^- = \frac{|F - R|}{|F|} \times 100 \quad (9)$$

where rec_sup$_f$ and act_sup$_f$ are reconstructed support and actual support respectively and R denotes the reconstructed set of frequent itemsets, and F denotes the correct set of frequent itemsets.

To measure result utility loss, Oliveira and Zaiane [19] presented two interesting metrics, called the Miss Cost (MC) and the Artifact Patterns (AP). MC measures the percentage of non-sensitive association rules that are missed from D'. Legitimate rules were lost when the support of some non-sensitive patterns is decreased below the threshold due to the sanitization. MC is defined as follows:

$$MC = \frac{\#\sim R_P(D) - \#\sim R_P(D')}{\#\sim R_P(D)} \times 100\% \quad (10)$$

where D and D' are the original database the sanitized database respectively; P is a set of all frequent patterns that can be mined from D., #$\sim$ R_P(D) and #$\sim$ R_P(D') denote the number of non-restrictive discovered from the original database and the sanitized database respectively. Ideally, MC should be equal to 0%. In the design of the association rule hiding, compromise often is made between MC and HF. Similar to MC, N-T-H (not to be hidden) was defined in [47] and it measured the number of non-sensitive itemsets hidden in the sanitized database D'.

$$N - T - H = \left|FI_s - SI_s - FI'_s\right| \quad (11)$$

where FI_s and FI'_s are the set of frequent itemsets in original dataset D and the sanitized database D' respectively, and SI_s denotes the set of sensitive itemsets that needs to be hidden.

In [19], AP was defined as the percentage of the discovered patterns that were artifact patterns. Representing the set of all patterns in D and D' by P and P' respectively, AP is computed by the following formula:

$$AP = \frac{|P'| - |P \cap P'|}{|P'|} \quad (12)$$

where the symbol "| |" denotes the cardinality. Ideally, AP is 0, implying that no artificial patterns incurred during the sanitization processing.

A similar metric to AP, i.e., N-T-G (not to be hidden), was presented in [47]. N-T-G is defined as the number of frequent itemsets in the sanitized database D' that are infrequent in the original database D.

$$N - T - G = \left|FI_s - FI'_s\right| \quad (13)$$

where FI_s and FI'_s represent the set of frequent itemsets in original dataset D and the sanitized database D' respectively.

Later a great many association rule hiding works [19]–[23], [28], [25], [35], [37], [42], [43] also took AP and MC as the result utility metrics of the proposed algorithms. In [21]–[23], apart AP and MC metrics, they evaluated the result utility of their proposed algorithms by considering SEF (side effect factor) metrics. Similar to MC, SEF is used to evaluate the amount of non-sensitive association rules that are lost because of the effect of the sanitization process.

There are some works [31], [32], [35] which used the combination of some metrics as their result utility. In [37], apart from Ghost Rule (GR) and Lost Rule (LR), Khan *et al.* used Loss of Information (LI) as their metrics of result utility. LI is defined as:

$$IL = NumberofDataItemModified \times LostRule(LR) \quad (14)$$

Table 4 summarizes the privacy level metrics and data utility metrics applied in the existing privacy preserving association rule mining algorithms. As shown in Table 4, HF is the most commonly used metric to measure privacy, MC and AP are the most commonly used metrics to evaluate data utility.

C. COMPLEXITY METRICS

The complexity of most PPARM algorithms concerns efficiency and scalability. The efficiency is generally measured in term of time and space required to implement the given algorithm. Space is evaluated according to the amount of memory required in the process of executing the given algorithm. Time is generally assessed by the following three aspects: (1) the CPU time; (2) the computational time; (3) the communication time.

In [19], Oliveira *et al.* tested the relationship between the CPU time and the database size, while keeping the sensitive patterns constant. In addition, they measured the relationship between the CPU time and the number of the sensitive patterns, while keeping the number of the transactions fixed. In their test, they fixed the disclosure threshold and support threshold.

In many cryptography-based PPARM outsourcing approaches, the authors tested the performance of the algorithms in term of the computation time, such as encryption

TABLE 4. The privacy level metrics and data utility metrics applied in the existing privacy preserving association rule mining algorithms.

Privacy Level Metrics	Data Utility Metrics: Data Metrics	Data Utility Metrics: Result Metrics
-1-R(p), where R(p) is the reconstructed probability of original database. [15] -K acts as privacy metric in k-anonymity privacy model [16] -Hidden Failure (HF) [19-23],[28-29] ,[25][31], [33], [35], [37],[42-43] -F-T-H (failure to hide some sensitive information)[47]	-Reconstruction accuracy [33] -Data utility [33] -Data accuracy [33] -Number of transaction modified [36] -Information loss incurred in the process of reconstruction [77] -Dissimilarity(Diss) [21-23], [29]	-Miss cost(MC) [19-23],[28], [25], [33], [35], [37], [42-43] -Artifact patterns (AP) [19-23][28], [28-30], [35], [37], [42-43] -N-T-H (not to be hidden) [47] -Support error (ρ) [15] -Identify error (σ) [15] -Loss of information(LI)= Number of Data Items Modified $\times$ Lost Rules (LR) [37] -Combination of AP and MC [31] -Side effect factor (SEF) [21-23]

time [48]–[55], decryption time [53], mining time [48], [53] and storage cost [56]. It is well known that an algorithm with polynomial complexity exceeds another one having linear complexity or exponential complexity. SMC is often used to address the problem of association rule mining over distributed data, therefore, in this type of algorithms, the authors often analyze the computation complexity based on the protocol utilized in the algorithms. Xie and Zhu [70] analyzed the computation complexity based on the time of executing DOT and Lagrange's interpolation. Tassa [69] analyzed the computation cost based on the encryption time and decryption in steps of the protocol.

In the PPARM algorithms over distributed data, communication time [58], [59], [61], [62], [69], [70] is often measured based on either on the time, or the number of exchanged messages, and the bandwidth consumption.

VII. CONCLUSION

In the last decades, as a new and rapidly emerging research area, PPARM has been widely researched in a myriad of fields. There are a variety of approaches which have been employed for PPARM.

In this survey, the existing PPARM algorithms are divided according to four dimensions and a multidimensional topological structure of PPARM are presented. An inclusive overview of the existing PPARM algorithms is provided according to the content that needed to be protected from disclosure. The obvious advantages and notable disadvantage of the existing algorithms are analyzed and emphasized. At last, we present various metrics to measure PPARM algorithms. The presented survey indicates the ever increasing interest of researchers in the area of protecting sensitive data and mined patterns from malicious users.

VIII. SOME DIRECTIONS AND CHALLENGES IN THE FUTURE

The evolution of PPARM is spurred by a variety of privacy requirements in numerous application domains. Different applications have different requirements for privacy and data utility. This discrepancy leads to a great many PPARM algorithms and techniques. However, none of the existing PPARM algorithms can surpass all the other algorithms with regard to all the metrics, such as privacy, data utility and complexity. It is often a reasonable choice to consider a trade-off among many aspects, such as the desired privacy level, the data utility, the computation complexity and even the practical feasibility and scalability of the schemes. Apart from some classical requirements, there are still some challenges from other aspects in designing the PPARM algorithms.

(1)In most of the existing ARM algorithms, interestingness of the association rules was measured by the support and confidence. Depending on the feature of specific applications, different metrics can be utilized to measure the interestingness of association rules.

(2)Each user has different concern and requirements over privacy. Therefore, it is required to design the schemes that can achieve personalized privacy. Personalized privacy enables the users to have a level of control over the specificity of their data; however, it is a challenging problem to implement personalized privacy.

(3)Many existing metrics are specific to some applications, therefore, this leads to a difficult comparison among the advantages and disadvantages of the existing PPARM schemes. Therefore, more universal applicable metrics, such as the information entropy-based methods, are required for an effective comparison of different PPARM schemes.

(4)Traditional PPARM algorithms are developed to discover useful and meaning association rules in structured data. However, heterogeneity is the inherent factor of distributed data. Therefore, it is a challenging problem for PPARM to uncover the qualified association rule hidden in unstructured and/or semi-structured data.

(5)So far, traditional DM schemes are aimed at static data. Data mining should be an online and continuous process. So data mining over data streams is a new challenging problem for data mining researcher.

(6)Few researches have touched on the problem of discovering graphs and structured patterns from large data. Therefore it is a challenging problem to discover graph and structured patterns in the future.

(7)Homomorphic encryption and the oblivious transfer protocol are two widely used techniques for preserving privacy. These two techniques can achieve full privacy without incurring in a loss of utility. However, these techniques are often not efficient for real-time applications. So it is required to develop a more efficient secure protocol with better utility.

REFERENCES

[1] H. A. Madni, Z. Anwar, and M. A. Shah, "Data mining techniques and applications—A decade review," in *Proc. Int. Conf. Automat. Comput. (ICAC)*, Huddersfield, U.K., 2017, pp. 1–7.
[2] L. Zhang, Y. Zhang, and H. Ma, "Privacy-preserving and dynamic multi-attribute conjunctive keyword search over encrypted cloud data," *IEEE Access*, vol. 6, pp. 34214–34225, 2018.
[3] I. S. Alwatban and A. Emam, "Comprehensive survey on privacy preserving association rule mining: Models, approaches, techniques and algorithms," *Int. J. Artif. Intell. Tool*, vol. 23, no. 5, pp. 10563–10582, May 2014.
[4] P. Gayathiri and B. Poorna, "Association rule hiding for privacy preserving data mining: A survey on algorithmic classifications," *Int. J. Appl. Eng. Res.*, vol. 12, no. 23, pp. 14917–14926, 2017.
[5] M. N. Kumbhar and R. Kharat, "Privacy preserving mining of association rules on horizontally and vertically partitioned data: A review paper," in *Proc. Int. Conf. Hybrid Intell. Syst. (ICHIS)*, Gammarth, Tunisia, 2013, pp. 231–235.
[6] J. Vaidya, "A survey of privacy-preserving methods across vertically partitioned data," in *Privacy-Preserving Data Mining* Boston, MA, USA: Springer, 2008, pp. 337–358.
[7] W. Gan, J. C.-W. Lin, J. Zhan, and H.-C. Chao, "Data mining in distributed environment: A survey," *WIREs Data Mining Knowl Discovery*, vol. 7, no. 6, p. e1216, 2017. doi: 10.1002/widm.1216.
[8] K. Sathiyapriya and G. S. Sadasivam, "A survey on privacy preserving association rule mining," *Int. J. Data Mining Knowl. Manage. Process*, vol. 3, no. 2, pp. 119–131, Mar. 2013.
[9] G. S. Navale and S. N. Mali, "Survey on privacy preserving association rule data mining," *Int. J. Rough Sets Data Anal.*, vol. 4, pp. 63–80, Apr. 2017.
[10] E. Bertino and I. N. Fovino, "Information driven evaluation of data hiding algorithms," in *Proc. Int. Conf. Data Warehousing Knowl. Discovery*, Copenhagen, Denmark, 2005, pp. 418–427.
[11] S. Fletcher and M. Z. Islam, "Measuring information quality for privacy preserving data mining," *Int. J. Comput. Theor. Eng.*, vol. 7, no. 1, pp. 21–28, May 2015.
[12] V. S. Verykios and A. Gkoulalas-Divanis, "A survey of association rule hiding methods for privacy," in *Privacy-Preserving Data Mining*. Boston, MA, USA: Springer, 2008, pp. 267–289.
[13] 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.
[14] V. S. Verykios, I. N. Fovino, I. N. Fovino, Y. Saygin, Y. Saygin, and E. Bertino, "State-of-the-art in privacy preserving data mining," *ACM SIGMOD Rec.*, vol. 33, no. 1, pp. 50–57, May 2004.
[15] S. J. Rizvi and J. R. Haritsa, "Maintaining data privacy in association rule mining," in *Proc. 28th Int. Conf. Very Large Data Bases*, Hong Kong, 2002, pp. 682–693.
[16] L. Sweeney, "K-anonymity: A model for protecting privacy," *Int. J. Uncertainty, Fuzziness Knowl.-Based Syst.*, vol. 10, no. 5, pp. 557–570, 2002.
[17] M. Atallah, E. Bertino, A. Elmagarmid, M. Ibrahim, and V. Verykios, "Disclosure limitation of sensitive rules," in *Proc. Knowl. Data Eng. Exchange*, Chicago, IL, USA, 1999, pp. 45–52.
[18] E. Dasseni, V. S. Verykios, A. K. Elmagarmid, and E. Bertino, "Hiding association rules by using confidence and support," in *Proc. 4th Inf. Hiding Workshop*, Prague, Czech Republic, 2000, pp. 369–383.
[19] S. R. M. Oliveira and O. R. Zaïane, "Privacy preserving frequent itemset mining," in *Proc. Int. Conf. Privacy Secur. Data Mining*, Maebashi, Japan, 2002, pp. 43–54.
[20] V. S. Verykios, A. K. Elmagarmid, E. Bertino, Y. Saygin, and E. Dasseni, "Association rule hiding," *IEEE Trans. Knowl. Data Eng.*, vol. 16, no. 4, pp. 434–447, Apr. 2004.
[21] C. N. Modi, U. P. Rao, and D. R. Patel, "Maintaining privacy and data quality in privacy preserving association rule mining," in *Proc. Int. Conf. Comput. Commun. Netw. Technol.*, Karur, India, 2010, pp. 259–367.
[22] K. Shah, A. Thakkar, and A. Ganatra, "Association rule hiding by heuristic approach to reduce side effects & hide multiple R.H.S. items," *Int. J. Comput. Appl.*, vol. 45, no. 1, pp. 1–7, May 2012.
[23] N. H. Domadiya and U. P. Rao, "Hiding sensitive association rules to maintain privacy and data quality in database," in *Proc. Adv. Comput. Conf. (IACC)*, Ghaziabad, India, 2013, pp. 1306–1310.
[24] N. J. Ghalehsefidi and M. N. Dehkordi, "A hybrid approach to privacy preserving in association rules mining," *Adv. Comput. Sci.*, vol. 3, no. 12, pp. 69–72, Nov. 2014.
[25] N. J. Ghalehsefidi and M. N. Dehkordi, "A hybrid algorithm based on heuristic method to preserve privacy in association rule mining," *Indian J. Sci. Technol.*, vol. 9, no. 27, pp. 1–10, Jul. 2016.
[26] P. Cheng, C.-W. Lin, and J.-S. Pan, "Use hype to hide association rules by adding items," *PLoS ONE*, vol. 10, no. 6, Jul. 2015, Art. no. e0127834.
[27] P. Cheng, J. F. Roddick, S.-C. Chu, and C.-W. Lin, "Privacy preservation through a greedy, distortion-based rule-hiding method," *Appl. Intell.*, vol. 44, no. 2, pp. 295–306, Feb. 2016.
[28] M. Fouladfar and M. N. Dehkordi, "A heuristic algorithm for quick hiding of association rules," *Adv. Comput. Sci.*, vol. 4, no. 13, pp. 16–21, Jan. 2015.
[29] A. Telikani and A. Shahbahrami, "Optimizing association rule hiding using combination of border and heuristic approaches," *Appl. Intell.*, vol. 47, no. 2, pp. 544–557, Sep. 2017.
[30] H. Surendra and H. S. Mohan, "Hiding sensitive itemsets without side effects," *Appl. Intell.*, vol. 49, no. 4, pp. 1213–1227, 2018. doi: 10.1007/s10489-018-1329-5.
[31] Y. Saygin, V. S. Verykios, and A. K. Elmagarmid, "Privacy preserving association rule mining," in *Proc. Int. Workshop Res. Issues Data Eng. (RIDE)*, 2002, pp. 151–158.
[32] S. L. Wang and A. Jafari, "Using unknowns for hiding sensitive predictive association rules," in *Proc. Int. Conf. Inf. Reuse Integr.*, Las Vegas, NV, USA, 2005, pp. 223–228.
[33] A. Amiri, "Dare to share: Protecting sensitive knowledge with data sanitization," *Decis. Support Sys.*, vol. 43, no. 1, pp. 181–191, 2007.
[34] X. Chen, M. Orlowska, and X. Li, "A new framework of privacy preserving data sharing," in *Proc. 4th Int. Conf. Data Mining*, Brighton, MI, USA, 2004, pp. 47–56.
[35] Y. H. Guo, "Reconstruction-based association rule hiding," in *Proc. SIGMOD Workshop IDAR*, Beijing, China, 2007, pp. 51–56.
[36] M. N. Dehkordi, K. Badie, and A. K. Zadeh, "A novel method for privacy preserving in association rule mining based on genetic algorithms," *J. Softw.*, vol. 4, no. 6, pp. 555–562, Aug. 2009.
[37] A. Khan, M. S. Qureshi, and A. Hussain, "Improved genetic algorithm approach for sensitive association rules hiding," *World Appl. Sci. J.*, vol. 31, no. 12, pp. 2087–2092 2014.
[38] C.-W. Lin, B. Zhang, K.-T. Yang, and T. P. Hong, "Efficiently hiding sensitive itemsets with transaction deletion based on genetic algorithms," *Sci. World J.*, vol. 2014, Sep. 2014, Art. no. 398269.
[39] C.-W. Lin, T.-P. Hong, K.-T. Yang, and S.-L. Wang, "The GA-based algorithms for optimizing hiding sensitive itemsets through transaction deletion," *Appl. Intell.*, vol. 42, no. 2, pp. 210–230, 2015.
[40] J. C.-W. Lin, Y. Zhang, P. Fournier-Viger, Y. Djenouri, and J. Zhang, "A metaheuristic algorithm for hiding sensitive itemsets," in *Proc. Int. Conf. Database Expert Syst. Appl.*, Regensburg, Germany, 2018, pp. 492–498.
[41] X.-S. Yang and S. Deb, "Cuckoo Search via Lévy flights," in *Proc. World Congr. Nat. Biol. Insp. Comput. (NaBIC)*, Coimbatore, India, 2009, pp. 210–214.
[42] M. H. Afshari, M. N. Dehkordi, and M. Akbari, "Association rule hiding using cuckoo optimization algorithm," *Expert Syst. Appl.*, vol. 64, pp. 340–351, Dec. 2016.
[43] K. Doan, M. N. Quang, and B. Le, "Applied cuckoo algorithm for association rule hiding problem," in *Proc. 8th Int. Symp. Inf. Commu. Technol.*, 2017, pp. 26–33.
[44] M. Dorigo and L. M. Gambardella, "Ant colony system: A cooperative learning approach to the traveling salesman problem," *IEEE Trans. Evol. Comput.*, vol. 1, no. 1, pp. 53–66, Apr. 1997.
[45] S. Narmadha, "Privacy preserving data mining based on ant colony optimization," *Int. J. Comput. Appl.*, vol. 83, no. 8, pp. 21–25, Dec. 2013.

[46] P. T. Selvan and S. Veni, "Social ant based sensitive item hiding with optimal side effects for data publishing," *Indian J. Sci. Technol.*, vol. 9, no. 2, pp. 1–9, Jan. 2016.

[47] J. M.-T. Wu, J. Zhan, and J. C.-W. Lin, "Ant colony system sanitization approach to hiding sensitive itemsets," *IEEE Access*, vol. 14, pp. 10024–10039, 2017.

[48] W. K. Wong, D. W. Cheung, B. Kao, N. Mamoulis, and E. Hung, "Security in outsourcing of association rule mining," in *Proc. 33rd Int. Conf. Very Large Data Bases (VLDB)*, Vienna, Austria, 2007, pp. 111–122.

[49] I. Molloy, N. Li, and T. Li, "On the (IN)security and (IM)practicality of outsourcing precise association rule mining," in *Proc. IEEE Int. Conf. Data Mining*, Washington, DC, USA, 2009, pp. 872–877.

[50] C.-H. Tai, P. S. Yu, and M.-S. Chen, "k-Support anonymity based on pseudo taxonomy for outsourcing of frequent itemset mining," in *Proc. Int. Conf. Knowl. Discov. Data Mining*, Washington, DC, USA, 2010, pp. 473–482.

[51] F. Giannotti, L. V. S. Lakshmanan, D. Pedreschi, H. Wang, and A. Monreale, "Privacy-preserving data mining from outsourced databases," in *Computers, Privacy and Data Protection: An Element of Choice*, S. Gutwirth, Y. Poullet, P. De Hert, and R. Leenes, Eds. Dordrecht, The Netherlands: Springer, 2011, pp. 411–426.

[52] F. Giannotti, L. V. S. Lakshmanan, D. Pedreschi, H. Wang, and A. Monreale, "Privacy-preserving mining of association rules from outsourced transaction databases," *IEEE Sys. J.*, vol. 7, no. 3, pp. 385–395, Sep. 2013.

[53] J. Lai, Y. Li, J. Weng, C. Guan, Q. Yan, and R. H. Deng, "Towards semantically secure outsourcing of association rule mining on categorical data," *Inf. Sci*, vol. 267, pp. 267–286, May 2014.

[54] C. Huang and R. X. Lu, "EFPA: Efficient and flexible privacy-preserving mining of association rule in cloud," in *Proc. Int. Conf. Commun. China*, Shenzhen, China, 2016, pp. 1–6.

[55] L. Li, K.-K. R. Choo, A. Datta, Jun Shao, and R. Lu, "Privacy-preserving-outsourced association rule mining on vertically partitioned databases," *IEEE Trans. Inf. Forensics Security*, vol. 11, no. 8, pp. 1847–1861, Aug. 2016.

[56] L. Liu *et al.*, "Privacy-preserving mining of association rule on outsourced cloud data from multiple parties," in *Proc. ACISP*, W. Susilo and G. Yang Eds. Cham, Switzerland: Springer, 2018, pp. 431–451.

[57] C. Clifton, M. Kantarcioglu, J. Vaidya, X. Lin, and M. Y. Zhu, "Tools for privacy preserving distributed data mining," *ACM SIGKDD Explor. Newslett.*, vol. 4, no. 2, pp. 28–34, Dec. 2002.

[58] J. Vaidya and C. Clifton, "Privacy preserving association rule mining in vertically partitioned data," in *Proc. Int. Conf. Knowl. Discovery Data Mining*, Edmonton, AB, Canada, 2002, pp. 639–644.

[59] S. Zhong, "Privacy-preserving algorithms for distributed mining of frequent itemsets," *Inf. Sci.*, vol. 177, no. 2, pp. 490–503, Jan. 2007.

[60] J. Vaidya and C. Clifton, "Secure set intersection cardinality with application to association rule mining," *J. Comput. Secur.*, vol. 13, no. 4, pp. 593–622, 2005.

[61] X. Ge, L. Yan, W. Shi, and J. Zhu, "Privacy-preserving distributed association rule mining based on the secret sharing technique," in *Proc. Int. Conf. Softw. Eng. Data Mining*, Chengdu, China, 2010, pp. 345–350.

[62] R. Kharat, M. Kumbhar, and P. Bhamre, "Efficient privacy preserving distributed association rule mining protocol based on random number," in *Intelligent Computing, Networking, and Informatics*, D. P. Mohapatra and S. Patnaik, Eds. New Delhi, India: Springer, 2014, pp. 827–836.

[63] B. Rozenberg and E. Gudes, "Association rules mining in vertically partitioned databases," *Data Knowl. Eng.*, vol. 59, no. 2, pp. 378–396, Nov. 2006.

[64] H. Hammami, H. Brahmi, S. Ben Yahia, and I. Brahmi, "Using homomorphic encryption to compute privacy preserving data mining in a cloud computing environment," in *Proc. Eur. Medit., Middle Eastern Conf. Inf. Syst. (EMCIS)*, Coimbra, Portugal, 2017, pp. 397–413.

[65] M. Waddey, P. Poncelet, and S. Ben Yahia, "A novel approach for privacy mining of generic basic association rules," in *Proc. ACM 1st Int. Workshop Privacy Anonymity Very Large Databases (PAVLAD)*, Hong Kong, 2009, pp. 45–52.

[66] B. Wang, Y. Zhan, and Z. Zhang, "Cryptanalysis of a symmetric fully homomorphic encryption scheme," *IEEE Trans. Inf. Forensics Security*, vol. 13, no. 6, pp. 1460–1467, Jun. 2018.

[67] M. Kantarcioglu and C. Clifton, "Privacy-preserving distributed mining of association rules on horizontally partitioned data," *IEEE Trans. Knowl. Data Eng.*, vol. 16, no. 9, pp. 1026–1037, Sep. 2004.

[68] A. Schuster, R. Wolff, and B. Gilburd, "Privacy-preserving association rule mining in large-scale distributed systems," in *Proc. Int. Symp. Cluster Comput. Grid*, Chicago, IL, USA, 2004, pp. 411–418.

[69] T. Tassa, "Secure mining of association rules in horizontally distributed databases," *IEEE Trans. Knowl. Data Eng.*, vol. 26, no. 4, pp. 970–983, Apr. 2014.

[70] X. Juan and Z. Yanqin, "Application of distributed oblivious transfer protocol in association rule mining," in *Proc. 2nd IEEE Int. Conf. Comput. Eng. Appl.*, Washington, DC, USA, Mar. 2010, pp. 204–207.

[71] H. Chahar, B. N. Keshavamurthy, and C. Modi, "Privacy-preserving distributed mining of association rules using Elliptic-curve cryptosystem and Shamir's secret sharing scheme," *Sādhanā*, vol. 42, no. 12, pp. 1997–2007, Dec. 2017.

[72] C. N. Modi and A. R. Patil, "Privacy preserving association rule mining in horizontally partitioned databases without involving trusted third party (TTP)," in *Proc. 3rd Int. Conf. Adv. Comput. Netw. Inf. (ICACNI)*. New Delhi, India: Springer, 2015, pp. 549–555.

[73] O. A. Wahab, M. O. Hachami, M. Vivas, G. G. Dagher, and A. Zaffari, "*DARM*: A privacy-preserving approach for distributed association rules mining on horizontally-partitioned data," in *Proc. 18th Int. Database Eng. Appl. Symp.*, Porto, Portugal, 2014, pp. 1–8.

[74] R. X. Lu, H. Zhu, J. K. Liu, J. Shao, and X. Liu, "Toward efficient and privacy-preserving computing in big data era," *IEEE Netw.*, vol. 28, no. 4, pp. 46–50, Jul./Aug. 2014.

[75] N. Domadiya and U. P. Rao, "Privacy-preserving association rule mining for horizontally partitioned healthcare data: A case study on the heart diseases," *Sādhanā*, vol. 43, no. 8, pp. 127–136, 2018.

[76] N. R. Nanavati, P. Lalwani, and D. C. Jinwala, "Analysis and evaluation of schemes for secure sum in collaborative frequent itemset mining across horizontally partitioned data," *J. Eng.*, vol. 2014, Nov. 2014, Art. no. 470416. doi: 10.1155/2014/470416.

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

LILI ZHANG received the M.S. degree in cryptography from Xidian University, Xi'an, China, in 2008, where she is currently pursuing the Ph.D. degree with the National Key Laboratory of Integrated Services Networks, Xidian University. Her current research interests include security protocol and cloud computing.

WENJIE WANG was born in 1964. He is currently an Associate Professor with the University of Chinese Academy of Sciences. His research interests include network security and artificial intelligence.

YUQING ZHANG received the Ph.D. degree in cryptography from Xidian University, Xi'an, China, in 2000. He is currently a Professor of computer sciences and the Director of the National Computer Network Intrusion Protection Center, University of Chinese Academy of Sciences. His research interests include network and system security, cryptography, and networking.