VOL. 11, NO. 23, DECEMBER 2016 ISSN 1819-6608
ARPN Journal of Engineering and Applied Sciences

www.arpnjournals.com

AN EXTENSIVE REVIEW ON PRIVACY PRESERVING METHODS IN DATA MINING

Vinoth Kumar Jambulingam and Santhi Vaithiyanathan
School of Computer Science and Engineering, Vellore Institute of Technology University, Vellore, Tamil Nadu, India
E-Mail: jvinoth.kumar2015@vit.ac.in

ABSTRACT

In recent years, Privacy Preservation in data mining has emerged as an essential requirement for exchanging confidential information while publishing and validating data over the internet. On the other hand, the suspicious approaches and conflicts enabled refusal of many information providers towards the protection of data from disclosure results in complete rejection of information sharing or incorrect data sharing. In this paper, an extensive overview of novel perspective and systematic understanding of a list of published literature into various subcategories is presented. In addition, existing privacy-preserving data mining approaches, their advantages, and limitations are also presented. The existing privacy-preserving data mining methods are classified based on variants of k-anonymity, distortion and pattern hiding used along with data mining mechanisms such as association rule mining, classification and the environment such as distributed and outsourced. This extensive study reveals the existing methodologies with their respective limitations, challenges, and emerging trends. Hence, this review would help researchers to carry out further research in privacy preservation data mining.

Keywords: data mining, privacy, K-anonymity, classification, association, clustering, distortion, outsourcing.

1. INTRODUCTION

Online Phishing is an illegitimate way to obtain confidential information such as usernames, passwords, and credit card details by concealing as a reliable entity in an electronic communication. Therefore, enhanced cyberspace protection against such internet phishing becomes a priority. The intimidation exercised by such sophisticated phishing attacks with deceptions creates new challenges in terms of mitigation methods. Recently, online phishing caused important security and economic concerns on the users and enterprises across the world. Internet services such as online banking and trading due to human and software weaknesses have been subjected to remarkable monetary loss. Consequently, improved privacy preserving data mining methods are required for secure and reliable information exchange over the internet. The dramatic increase in storing customers' personal information led to a higher complexity of data mining algorithm with extensive impact on the information sharing. Among several existing algorithm, the Privacy Preserving Data Mining produces exceptional results related to the internal perception of privacy preservation with data mining. Strictly, the privacy must protect all the three main mining algorithms including association rule, classification, and clustering (Sachan *et al*. 2013). [1]The emergence of new cloud computing technology allowed the business collaborators to part the data and store the information for the shared benefits. All of these are linked to the growing capability to collect users' individual data together with the rising complexity of data mining algorithms that disturbs the information interchange. Yet, the concepts, utilization, categorization and various attributes of privacy preservation in terms of its merits and demerits are not regularly reviewed. Currently, several privacy preservation methods for data mining exist including variants of K-anonymity,L-diversity, randomization, cryptography and condensation methods(Sachan *et al*. 2013)[1].

The PPDM methods secure the data by hiding some original information so that confidential information is not disclosed. The purpose is to maintain a trade-off between accuracy and confidentiality. Other approaches that use cryptographic techniques to prevent information leakage are said to be computationally very costly (Ciriani *et al*. 2008) [2]. On the other hand, PPDMs use data distribution with either horizontal or vertical partitioning into several entities. The main objective of implementing such techniques is to maintain person' privacy while developing shared results over the entire data (Aggarwal and Yu 2008) [3]. Despite much research, methods with more satisfactory privacy settings are far from being developed. It is essential to protect the data before it gets distributed to cloud-based data providers. To protect privacy, clients' information needs to be identified before sharing with new users who are not directly allowed to access the relevant data. This can be performed by removing the unique identity attributes such as name and social security number and others. In spite of this, there still exist other types of information such as date of birth, zip code, gender, the number of children and account numbers which can also be used for possible identification of the person. Thus, highly advanced and extensively robust privacy preservation measures coupled with data mining must be implemented to protect the privacy of users. This demonstration highlights that the substantial development of privacy preserving data mining methods is required. The rest of the paper is structured as follows. At the start, a basic differential privacy model has described in the context of internet phishing mitigation along with the framework for all privacy preserving data mining techniques. The subsequent sections discuss different PPDM methods one after the other. Then, principal merits and demerits of the existing methods are discussed in

shortcomings of PPDM methods with the use of the table. Finally, Conclusion concludes the article with the possible future direction of the field.

2. RELATED WORKS

In this section, review of related work is carried out and presented elaborately. This literature review clearly helps researchers to find out various existing methodologies and its limitations.

2.1 Differential privacy model

In recent times, differential privacy model is commonly used to provide maximum security to the remote statistical databases by reducing the probabilities of records identification. There are many trusted parties that hold sensitive information in the form of datasets such as patient records, voter registration information, national population register, tourism and email usage. The main goal is to provide global statistical information about the public data while guarding those users confidentiality whose information exists in the dataset. Differential Privacy which is also called as in distinguish ability suggests the privacy in the atmosphere of statistical databases. In general, data privacy is seen as a representative for data safety. Clearly, this view is incorrect because the purposes of the two fields are opposite. On the other hand, security defends the data against unauthorized access when it is transmitted across a network. But, while arriving at an authorized user no additional controls are impressed on the data security to disclose the personal information of individuals. Thus, it is important to find the correlation between data privacy and data security because the latter is a prerequisite of the former.

Data needs to be secure at storage and the transmission also needs to be made through data security protocols. Additionally, if data privacy is a goal, then some more steps should be taken up to protect persons' confidentiality that exists in the data. Also, It is significant to define the process of PPDM in terms of data sharing and data mining operations across a number of users u1,…um with m >= 2. The data is seen as a database of m records, each consisting of n fields, where each record denotes a person p and explains them through its fields. In a simple representation of a table T containing rows which signifies p1…pn and columns which indicate the fields or attributes a1,…al. Considering a static representation, where each person is indicated by a column of components a1,…al. The most important aspect of PPDM is the privacy embedded in the table T, which an attacker wants to attain. The other aspect is the protected data structure, which belongs to one real-world entity and the need to share it with another user (m >= 2). It might be constructed from parts possessed by different real-world entities.

It is essential to present some descriptions to support the basics of PPDM concepts. Particularly, an explicit identifier is an attribute that allows a direct association of an instance (a record or a row in table T) to a user i. For an illustration, by identifying a mobile or a license number of a person it may clearly connect the record or a row in table T, where this explicit identifier to a person i is inserted. Equally, a quasi-identifier which is a set of individuals' non-explicit attributes may also associate a record or a row in table T to a specific person. For instance, in the United States, the quasi-identifier attributes such as date of birth, 5 digit zip code and gender uniquely identify 77 % of the nation's population (Sweeney 2002) [4]. By joining a public medical information dataset with a publicly accessible voters' list and using quasi-identifiers, Sweeney explained that it is easy to find the secret health records of all government employees from a distributed dataset of the Illinois governor, where only explicit attributes or identifiers are removed. Usually, the major PPDM identity protection methods that are based on simple notions are well-known to people as they are easily accessible in the literature. These concepts are described as camouflage or hiding in the crowd. One of the well-known hidings in the crowd approach to data confidentiality is the k-anonymity. Actually, the k-anonymity technique (Sweeney 2002 [4]; Nergiz *et al.* 2009) [5]changes the real data from table T to obtain a table T' such that for any quasi-identifiers q that can be built from attributes of table T there are at least k occurrences in T' so that q matches these instances. Besides, datasets need generalization to satisfy k-anonymity.

2.2 Privacy preserving data mining framework

Lately, the significance of privacy preserving data mining techniques are exhaustively analyzed and debated by Matwin (2013) [6]. Use of specific approaches shows their ability to stop the excessive use of data mining. Some techniques suggested that any denounced group cannot be directed more on data generalization than the entire population. Vatsalan *et al*. (2013) [7] revised the technique called 'Privacy Preserving Record Linkage' (PPRL), which permitted the linkage of databases to institutions by defending the privacy. Thus, a PPRL methods based categorization is suggested to examine them in 15 dimensions. Qi and Zong (2012) [8] overviewed several existing techniques of data mining for the privacy protection depending on data distortion, distribution, mining algorithms and rule or data hiding. On the topic of data distribution, only a small number of algorithms are presently used for privacy-preserving data mining on distributed or centralized data. Raju *et al*. (2009) [9]recognized the necessity to improve the protocol based homomorphic encryption with the existing method of the digital envelope in attaining collaborative data mining while guarding the sensitive data intact among the multiple parties. The proposed technique displayed significant influence on different applications. Malina and Hajny (2013) [10] and Sachan *et al*. (2013) [1] examined the present privacy preserving solutions for cloud environments, where the solution is drawn based on advanced cryptographic mechanisms. The solution presented the anonymous access, the maintenance of confidentiality the unlinkability of transmitted data. Then, it is implemented and the experimental results are

retrieved and the performance is compared. Mukkamala and Ashok (2011) [11] compared a set of fuzzy based techniques in the context of privacy preserving features and the capability to retain the same connection with other fields. This comparison is subjected to: (1) the initial modification of the fuzzy function definition, (2) the outline of the ways to join different values of a particular data item to a single value, (3) the use of some similarity metrics for the comparison of the original data with that of mapped data, and (4) the assessment of the effect of mapping on the resultant association rule.

3. CATEGORIES OF PPDM METHODS

In order to ensure confidentiality across mining platforms, there can be different privacy preserving methods that can be employed based on mining algorithm, environment being distributed or centralized, use of cryptographic algorithms, simple using variants of anonymization of k-anonymity which ensures better privacy with a lesser computation in standalone scenarios.

3.1Data distortion based PPDM methods

Kamakshi (2012) [12] suggested a novel idea to vigorously classify the sensitive attributes of the dataset. Identification of these attributes is based on the threshold boundary of the sensitivity of each distinguishing feature. It is found that the data owner altered the value under-recognized sensitive attributes using the swapping technique to keep the confidentiality of sensitive information. The data is changed in such a manner that the original characteristics of the data remain unaffected. In spite of the novelty, it is more time expensive. Islam and Brankovic (2011) [13] suggested an architecture involving new techniques that affected all the attributes in the dataset. Experimental findings show that the offered architecture is highly efficient in preserving the original characteristics in a perturbed dataset. Kamakshi and Babu (2010) [14] introduced a model architecture consisting of three elements namely clients, data centers and dataset on every site. The data center is fully inactive so that the site database and clients role appear interchangeable. Li et al. (2009) [15] developed a low-cost and low-risk anonymous perturbation mechanism via anonymous exchange and homomorphism encryption. The presented technique demonstrated robustness for optimized parameters. It is more complex and has a little loss in data utility. Wang and Lee (2008) [16] presented a technique to stop Forward Inference Attacks, in the sanitized data (implies modified data) generated by the sanitization.

3.2Association rule mining based PPDM methods

Aggarwal and Yu (2008) [3] underscored two major factors involving the association rule mining such as support and confidence. For an association rule X =>Y, the support is the number of transactions in the entire dataset which must have X U Y. The confidence (also called strength) of an association rule X => Y is the ratio of the transactions that actually exhibits such a rule in the entire dataset. Additionally, Belwal *et al*. (2013) [17] slightly changed the basis of support and confidence of sensitive association rules without changing directly the given database. On the other hand, the modification can indirectly be implemented by newly incorporating parameters connected to database transactions. New changes involve modified support called M support, modified confidence called M confidence and Hiding counter. The algorithm used the definition of support and confidence. Thus, it concealed the all sensitive association rule without any side-effect. Though, it can hide only the rules for a single sensitive item on the Left Hand Side of the association rule. Naeem *et al*. (2010) [18] presented an architecture which separated the classified association rules with complete removal of the well-known side effects such as the non-genuine association rules, generation of unwanted rules while yielding no 'hiding' failure. In this architecture, ordinary statistical measures are employed instead of the conventional concept of support and confidence to construct association rules, a particularly weighing system based on central tendency. Li and Liu (2009) [19] presented an association rule mining algorithm for privacy preserving known as DDIL. The proposed algorithm is built on data disturbance and inquiry limitation. The original data can be disturbed or hidden by using DDIL algorithm to improve the confidentiality more efficiently. This is an effective technique for generating frequent itemsets from transformed data. Experimental results show that the proposed technique is more efficient to generate acceptable values of privacy stability with suitable choice of random parameters.

3.3Association rule hiding basedPPDM methods

Fast Hiding Sensitive Association Rules (FHSAR) algorithm is developed by Weng *et al.* (2008) [20]. This protected the SAR with lower side-effects, where an approach is made to avoid hidden failures. Also, two heuristic techniques are created to increase the efficiency of the system to solve the limitations. The heuristic function is used vigorously to find the earlier weight for each particular record so that the order of changed records can be decided more effectively. Consequently, the association between the sensitive association rules and each record in the original dataset are examined by successfully selecting the suitable item for modification. This leads to effective sanitization of confidential information in the updated dataset. Dehkordi *et al*. (2009) [21] proposed a novel multi-objective technique to conceal the sensitive association rules while improving the security of dataset. In fact, this retained the utility of extracted rules at a more efficient level. The proposed method is based on a genetic algorithm where the accuracy and privacy of database are improved considerably. Li et al. (2009b) [16]offered a new algorithm to sanitize transactional datasets. This is itemset based, where the support of larger item sets are significantly lowered below the threshold limit set by the client. Thus, no rules can be obtained from some restricted item sets. A new technique is also proposed to choose the items that requested removal from the database to avoid the detection of a set of rules. The main demerits are

related to the choice of victim items without moving the non-sensitive patterns when the sanitization of the third and the fourth sensitive records are defined. Kasthuri and Meyyappan (2013) [22] proposed a novel scheme to determine the sensitive items by concealing the sensitive association rules. The developed technique found the frequent itemsets and generated the association rules. Representative association rules concept is chosen to find the sensitive items. Hiding the sensitive association rules using chosen sensitive items is found useful. Quoc *et al*. (2013) [23] have proposed a heuristic algorithm based on the intersection lattice of frequent item sets to protect the set of confidential association rules employing distortion method. To lower the side effects, the heuristic for support and confidence minimization oriented intersection lattice (HCSRIL) algorithm are used. This identified the victim item and reduced the number of records by affecting least impact on item-sets variations in Generate (FI) function of the algorithm.

3.4Classification based PPDM methods

Xiong *et al*. (2006) [24] presented a closet-neighbour classification technique based on Secure Multi-party Computation techniques to solve the privacy tests in few phases including the pf selection of the privacy preserving closet-neighbour and the classification of privacy preserving. The newly created algorithm is balanced in terms of performance, accuracy and privacy protection. Also, it is flexible by a number of settings to fulfill different optimization conditions. Singh et al. (2010) [25] developed a simple and efficient privacy preserving classification for cloud-based environments. Jaccard similarity measure is employed to calculate the nearest neighbors for KNN classification and the equality test is presented to calculate it between two encrypted transactions. This approach simplified a protected local neighbor calculation at each node in the cloud and classified the hidden records via weighted KNN classification scheme. It is noteworthy to focus on allowing the robustness of the proposed approach so that generalization to many data mining tasks can be made simultaneously, where privacy and security are needed. Baotou (2010) [26] proposed an effective algorithm based on random perturbation environment to safeguard privacy classification mining. It is employed on discrete data of Boolean type, character type, number type and classification type. The experimental analysis discovered the considerably improved features of proposed algorithm in terms of accuracy of mining computation and privacy protection, where the computation process is simplified but at a greater cost. Vaidya *et al.* (2008) [27] proposed amethod for vertically partitioned dataset mining. This technique could transform and cover a variety of data mining applications such as decision trees. More effective solutions are required to find abetter upper bound on the difficulty. Kantarcioglu and Vaidya (2003) [28] emphasized the use of secure summation and logarithm, where the distributed naive Bayes classifier algorithms are protected. The experimental results show the concept of few useful protocols that enabled the secure distribution of different types of distributed data mining algorithms.

3.5Clustering based PPDM methods

Yi and Zhang (2013) [29] reviewed several existing solutions to reserve privacy of k means clustering and provided a correct definition for similarly contributed multiparty protocol. A similarly contributed multiparty k means clustering is used on the vertically partitioned dataset, wherein each data site contributed k means clustering equally. According to the mechanism, data sites collaborated to encrypt k values with a shared public key in each phase of clustering. Then, it securely compared k values and returned the index of the minimum without showing the intermediate values. In certain situation, this is more practical and efficient than Vaidya–Clifton protocol (Vaidya *et al*. 2008) [27].

3.6 Privacy preserving outsourcing based PPDM methods

Giannotti *et al*. (2013) [30] discussed the issues concerning the outsourcing of association rule mining activity for a commercial privacy preserving network. An attack model is created based on the contextual knowledge for privacy preserving outsourced mining. An encryption technique called as Rob Frugal is developed. This is based on one-to-one substitution ciphers of items, which involved the fake records in sharing each cipher item with the same occurrence as of k-1 to the others. A simple summary of the fake records is utilized for true support of mined rules from which the server can be restored effectively. It is described that the proposed technique is effective against an adversarial attack which is based on the actual itemsets and their exact support count. This strategy is based on that the attacker is not aware of such information. Besides, any relaxation may damage our encryption technique and bring privacy weaknesses. They examined encryption techniques that could repel such privacy weaknesses. The strategies for the enhancement of the Rob-Frugal algorithm to reduce the number of spurious rules are also studied. Worku *et al*. (2014) [31] improved the efficiency of the above technique by minimizing the computationally oriented operations such as bilinear mapping. The technique exposed secure results after a detailed examination on the performance front. Though, the data block insertion made the presented technique static. Therefore, the need for a completely secure and dynamic public auditing scheme remains as a challenge for a cloud environment. Arunadevi and Anuradha (2014) [32] examined the issues associated with theoutsourcing of frequent itemsets for a commercial privacy preserving networking model. An attack model is presented by assuming that the attackers are aware of the items and support count of the item. In such an eventuality, the attackers are fully aware of the details of the encryption scheme and some pairs of items with the related cipher values. These new assumptions significantly enhanced the security of the system and prevented the itemset based attack as well as decreased the processing time. Kerschbaum and Julien (2008) [33] proposed a

searchable encryption technique for outsourced data mining. In this technique, the client had to encrypt the data once and transmit that encrypted information via the network to the data analyst. The data analyst then initiated a number of queries for obtaining necessary permission from the client to translate the contents of the data in the queries. The novel encryption technique allowed the search of range and keyword queries. Also, the technique allowed queries to reuse the output of earlier queries as tokens to make dependent queries without boundary. The presented scheme is found to be highly secure. There are several open issues that exist in the area of searchable encryption. In the case of outsourced data mining, it is most remarkable to associate the efficiency enhancements possible for range queries with the required security requirements via pairing based cryptography.

3.7 Distributed environment based PPDM methods

Li (2013) [34] discussed the merits and demerits of privacy preserving methods by creating and analyzing a private key based privacy preserving technique to support mining counts. An incentive-based approach is offered to analysis the secure computation by developing a reputation system in the networks. The proposed technique provided an incentive for misbehaving nodes to act correctly. Experiments show the system is efficient in identifying the misbehaving nodes and improving the average throughput in the entire network. Moreover, Dev *et al.* (2012) [35] recognized the privacy risks associated with data mining on a cloud system and developed a distributed framework to eliminate such risks. The proposed method involved disintegration, classification, and distribution. This eliminated the data mining by protecting the privacy levels, splitting the data into portions and keeping them onto suitable cloud providers. However, the proposed system presented a proper way to secure privacy from mining based attacks, but it resulted in a performance overhead as client retrieved the data more often. For an example, a client had to perform a global data analysis for a whole dataset, where the analysis needed accessing the data over different locations with a reduced performance. Tassa (2014) [36] proposed a protocol for secure mining of association rules in horizontally distributed dataset. The proposed scheme exhibited benefits over other leading schemes in terms of performance and safety. It consists of two set of rules including (1) a multiparty scheme to compute the union or intersection of sensitive subsets held by each client and (2) a scheme to check the presence of an item possessed by the client in a subset held by another. Methods based on Field and Row Level distribution of transactional dataset are proposed by Chan and Keng (2013). They proposed a distributed framework to preserve association rule mining and analyzed the probability of its deployment. Based on the characteristics of the information in the database, they are distinguished by their distribution to multiple servers. Its privacy concepts are studied from two separate perspectives such as K-anonymity and distribution of support values. The proposed mechanisms for assigning transactions to outsourced servers depend on the significance of the types of privacy conception to a user. Dong and Kresman (2009) [37] described the relation between prevention of accidental leak of private data and distributed data mining in privacy preserving schemes, where two mechanisms are developed to eliminate such disclosures. The first one was a simplified protocol used for different application, whereas the second one delivered the correctness of fewer broadcasts and collusion resistance. The easiness of the proposed schemes allowed negligible requirements for data structures, data storage, and computation. Inan and Saygin (2010) [38] proposed a method to collect dissimilarity matrix for horizontal data mining in distributed environment. The evaluation needed all the operations on records in the form of the pair for sensitive personal datasets which are distributed in different sites horizontally. This method took the data either in the form of numerical or character. For these two data sets of different types, a number of comparison functions are made accessible. Yet, as estimated, guaranteeing privacy has its costs, taking the comparison against the standard protocol where secretive data is shared with third parties. They used the secure comparison schemes for collecting horizontally partitioned datasets. There are several other application areas of these methods such as outlier detection and record linkage problems. Nanavati and Jinwala (2012) [39] expanded different methods used to find global and partial cycles keeping the privacy of the particular parties secured in a distributed setup. The enclosed algorithm is modified to find global cycles in cyclic association rules confidentially. The privacy preservation techniques are suggested based on secret sharing and homomorphic approach. It is established that the methods based on Shamir's secret sharing can be used to determine the partial global cycles. Still, few open research tests including the application of these privacy preserving models to other temporal rule mining approaches like temporal predicate association rules and calendric association rules need to be lectured. Another research challenge also includes deciphering the most accurate and efficient technique in this scenario by practically comparing the cost for each method. Agrawal and Srikant (2000) [40] established a uniform randomization technique based association rule for the categorical datasets. In this method, the client replaces each data item by a new data item which is absent in the dataset before sending the same data to the server. The substitution process of specific values from datasets with other values is termed as uniform randomization. This isa simplification of Warner's randomized response method. Some other types of data reconstruction techniques involve the original data being taken for sanitizing called as a knowledge base. As a result, recently acquired data is then collected based on the sanitized knowledge. The usefulness of randomization based on reconstruction for categorical attributes is demonstrated. Wang et al. (2010) [41] suggested a modified algorithm called privacy preserving frequent data mining and related computation technique based on the Frequent Data Mining (FDM) to maintain privacy. The method included the computation required for privacy preserving technique along with total

support count while confirming the local large item set and local support count source is protected. Accordingly, the time required for the communication is avoided and protected the distributed data privacy at each site. The experimental results confirmed the efficiency and correctness of the scheme for real-world application, particularly in privacy-preserving mining. Nguyen *et al*. (2012) [42]offered an Enhanced M. Hussein *et al*.'s Scheme (EMHS) for a secure privacy based association rule mining, where horizontally distributed dataset is used. EMHS (developed in 2008) is able to change the efficiency along privacy with growing number of sites. The efficiency of EMHS is determined to be much better than MHS, specifically for datasets with a higher number of sites. A second method is also presented for the other types of datasets. It is significant for solving the collision of combiner and Initiator. Ibrahim *et al*. (2012) [43] proposed a novel cryptographic scheme to compute the KNN categorization over the distributed cloud datasets. Their experiments showed better accuracy than earlier approaches. It is found that such techniques can thwart the users concerns and increases the chances of adoption of cloud computing. The modification of their secure classifier to work in the malicious attack model will be startedaway. Patel et al. (2012) [44] suggested an algorithm to defend the secrecy distributed over K-Means cluster using Shamir's secret sharing scheme. The suggested method calculated the cluster mean collaboratively and avoided the participation of trusted third party. Upon comparison, it is detected that the suggested framework is orders of magnitude faster as related to the homomorphic encryption and polynomial evaluation techniques in terms of reliability and computation cost. It is important to spread the suggested algorithm in vertical partitioning in the presence of malicious attack model. Nix et al. (2012) [45] employed two protocols for the scalar (dot) product of two vectors which are used as sub-protocols in larger data mining activities. Experimental results revealed their low data leakage, high accuracy, and an improved efficiency. The security aspects of these calculations under a security definition are also examined. Compared to the previous definitions these are found to be very effective approximation protocols. It is useful to analyze the use of these dot product protocols in other data mining tasks such as clustering, neural networks, and support vector machines.

3.8 K-anonymity based PPDM methods

For the sake of clearness, it is common to deliver two important definitions of K-anonymity. The first definition states that: QI being a quasi-identifier for a given Table T is a set of attributes {A1 . . . An} if combined literally leads to the identification of an individual, Pi (Sweeney 2002) [4]. The second definition is stated as follows: a table T is said to be satisfying K-anonymity only if for every tuple or record R belonging to table T has, at least, k-1 other tuples ti1ti2 . . . tk1 such that looks the same. (Machanavajjhala *et al.* 2007) [46]. (Wang *et al.* 2004) examined the data mining method called as data masking which is based on privacy protection. The data mining methods are examined in terms of data generalization, where the data mining is done by hiding the original information instead of patterns. After data masking, the common data mining methods are used without any modification. Two key metrics, scalability, and quality are mainly observed. The issue of quality is settled using the trade-off between information loss and privacy. The scalability issue is recognized retaining new data architecture while concentrating on better generalizations. Loukides and Gkoulalasdivanis (2012) [47] suggested a new technique to anonymize the dataset by utilizing the data publishers' utilization necessities facing less information loss. An accurate information loss quantity and an efficient anonymization algorithm are presented to reduce the information losses. Experimental results on medical data shown that the offered technique permitted more consistent query answers than the earlier techniques which are similar in terms of efficiency. Friedman et al. (2008) [48] modified the definitions of K-anonymity to prove that the data mining model does not infringe the K-anonymity of the customers denoted in the learning examples. A tool is presented to find the amount of anonymity achieved during data mining. The suggested method displayed its adoption capability to different data mining problems including clustering, association rule mining and classification. K-anonymity is improved by combining with data mining approach to keep the respondent's identity. Ciriani *et al*. (2008) [2] emphasized the possible threats to K-anonymity, which are stressed via the implementation of mining to gather data and examines of two main techniques to linkK-anonymity in data mining. The different methods adopted to detect K-anonymity breaches are also elaborated. Additionally, the elimination of these methods in classification mining and association rule mining are elaborated. He et al. (2011) [49] suggested an algorithm based on clustering to produce a utility-friendly anonymized version of sensitive personal data. This technique is found to overtake the nonhomogeneous technique where the size of QI-attribute is higher than 3. They performed a clustering-based K-anonymity algorithm, which exposed significant improvement in the utility performance when applied to many real-world datasets. In recent times, K-anonymous privacy preservation is commonly used. Additional modification looked to be even more complex without resolving several issues. Soodejani *et al*. (2012) [50] proposed a version of the chase called as standard chase, which put some restrictions on the constraints and dependencies, such as being conjunctive and positive. The anonymity norm of their technique discloses some similarities to the L-diversity privacy scheme. Examination of other privacy models such as t-closeness may deliver a better privacy model for the suggested method with a higher degree of usefulness. Karim *et al.* (2012) [51] offered a numerical method to extract maximal frequent patterns with privacy preserving capability. This method showed a novel encoded and compressed lattice structure with an efficient data transformation technique via MFPM algorithm. The

proposed MFPM algorithm and lattice structure made both the search space as well as the searching time lower. The experimental results showed that the MFPM algorithm performed better than PC_Miner and other existing maximal frequent pattern mining algorithms. In addition to the lattice structure, it performed well then the FP-tree and PC-tree algorithm. Loukides *et al.* (2012) [52] projected a rule-based privacy scheme that endorsed data publishers to direct fine-grained protection necessities for both sensitive information and identity disclosure. Based on this technique, they developed two anonymization algorithms. Their first algorithm ran in a top-down fashion, retaining an efficient plan to recursively generalize data with less information loss. On the contrary, the second algorithm used a combination of top-down and bottom-up generalized heuristics and sampling. This significantly enhanced the scalability and maintained less information loss. General experimentations display that these algorithms considerably outpaced the state-of-the-art techniques in the context of recalling data utilization while keeping scalability and good protection. Vijayarani *et al.* (2010b) [53]considered K-anonymity as an interesting approach to protecting sensitive data related to semi-public or public sectors from linking based attacks. The probable threats to K-anonymity approach are discussed in detail. Mostly, the problems related to the approaches and the data are recognized to combine K-anonymity in data mining. Nergiz *et al*. (2009) [5] developed and improved the definitions of K-anonymity to a far greater degree. It is publicized that previously used techniques is either unsuccessful to secure privacy or as a whole minimized data protection and the data utilization in a multiple relations setting. A new clustering algorithm is presented to attain multi-relational anonymity. Experimental results demonstrated that the suggested technique is an efficient approach in terms efficiency and of utility. Support for arbitrary systems with multiple public or private entities is considered. The problem of secure outsourcing of frequent itemset mining on the cloud environments is examined by Tai *et al*. (2013) [54]. Regarding the challenges in big data analysis, they recommended to divide the data into several parts and outsourced each part individually to different cloud provider based on pseudo-taxonomy, anonymization technique, called as KAT. They offered DKNT to ensure the security and privacy for each partial data outsourced to multiple cloud providers. Experimental results confirmed excellent achievement in terms of better computation efficiency and protection as related to those on a single site. Tai *et al*. (2010) [55] presented K-support anonymity, which delivered protection against anexperienced attacker with the detailed support information. To get the K-support anonymity, a pseudo taxonomy tree is presented with the third party mining scheme for the generalized version of frequent itemsets. The building of the pseudo taxonomy tree simplified the concealing of the real items and less the fake items introduced in the encrypted database. The results presented good privacy protection with an improved storage overhead. K-anonymity is additionally extended by Pan *et al*. (2012) [56]. They analyzed and compared the variants of K-anonymity models with their applications. The modified K-anonymity variants such as the L-diversity, (a, K) k-anonymity and (a, L)-diversification K-anonymity overcome some of the existing restrictions associated with privacy. Few K-anonymous approaches are used in attaining the main technology. Based on suppression, Deivanai *et al*. projected a new K-anonymity technique called kactus (Deivanai *et al*. 2011) [57]. In the offered technique, multidimensional suppression is achieved. The values are suppressed in certain records based on other attributes without the use of domain hierarchy trees. Thus, this method recognized the attributes independent of classification of the data records and suppressed these values to conform to K-anonymity. This method is used on a different database to decide its efficiency and accuracy and analyzed with other K-anonymity based techniques. It isstated that in a multiparty scenario, the anonymization can be achieved with perturbation to maintain privacy. A new definition of K-anonymity scheme for an efficient privacy protection of sequential personal data is presented (Monreale *et al*. 2014) [58]. This technique altered the sequential databases into K-anonymized databases, while preserving the utility of data with reference to a variety of analytical properties. A sequence of experimentation on different real-life sequential datasets displayed that the projected method significantly secured the sequential pattern mining results not only in terms of support but also of extracted patterns. Additionally, the results arehighly interestingly regarding dense datasets. Nergiz and Gok (2014) [59] and Nergiz *et al*. (2013) [60] proposed the hybrid generalizations. It not only achieved the generalizations but also included data relocation mechanism. In data process, the location of certain cells is altered to some populated, not recognizable data cells. The relocation process assistedin generating betteranonymization and ensured underlying privacy. The data relocation is a trade-off between reliability and the utilization of the data, where the trade-off is measured by the provider's parameter. The results exposed that a relatively small number of relocations could improve the utility as compared to the query answering accuracy and heuristic metrics. A Hybrid generalizations scheme to relocate the data is presented (Nergiz and Gok 2014) [59]. In data replacement process, data cells are replacedby some populated small groups of tuples which remained different from each other. Again, the data relocation helped to generate better anonymization which confirmed the data privacy. It is established that a small number of rearrangements could extra ordinarily improve the utility. Novelhybrid algorithms can be considered for other privacy metric such as diversity, d-presence, and (a, k) anonymity. This would be central in addressing different types of attackers. Recently, Zhang *et al*. (2014b) [61] presented a two-phase TDS technique based on Map Reduce framework on a cloud environment. Initially, the data sets are partitioned after anonymization and parallel and intermediate results being generated. In the second stage, these intermediate results are combined for additional anonymization to create consistent K-anonymous datasets. The Map Reduce paradigm on the

cloud is used for data anonymization and a group of data is considered intentionally to concretely attain the particular computation in a scalable way. The results from the implementation of this technique on real-world datasets showed that the presence of competence of TDS and scalability made the performance much better than existing techniques. They have offered an effective quasi-identifier index oriented technique to preserve the privacy of incremental datasets on the cloud. In the suggested technique, QI groups are enumerated using domain values in the present generalization level, which permitted the access only to a small part of records in any dataset rather than accessing of the whole database (Zhang et al. 2013b, c) [62]. In addition, Ding et al. 2013) [63] proposed a distributed anonymization protocol for privacy-preserving data publishing from a set of data providers in a cloud environment. Their technique achieved a personalized anonymization to satisfy each data provider's requirements and the union created a global anonymization to be published. They also proposed a new anonymization algorithm using R-tree index structure.

4. LIMITATIONS OF PPDM METHODS

Limitations of PPDM methods existing, many data mining techniques are available to protect the privacy. Mostly, the privacy preserving techniques are categorized according to data distribution, data distortion, data mining algorithms, anonymization, data or rules hiding, and privacy protection. Table-1 summarizes several techniques that are in use to secure data mining privacy. Exhaustive research findings over the years discovered that the existing privacy-preserving data mining methods still lacks from major incompleteness including the multi semi-honest providers to distributed clients' data, the overhead of global mining computing, incremental data privacy concern in cloud computing, mining result integrity, data utility, scalability and performance overhead. Therefore, efficient, strong and scalable newer models are indispensable to overcome these limitations. Also, proper anonymization of data is required to maintain the privacy of each client before publishing. The association between personal data and personal identification should disappear.

Table-1. Description of PPDM methods.

PPDM methods	Description
Data distribution	Datacan either be vertically or horizontally partitioned across multiple sites
Data distortion	Distortion did by any one of the approaches such as perturbation, blocking, aggregation, merging, sampling, and swapping
Data mining algorithms	Encloses association rule mining, classification mining, clustering, and Bayesian networks etc.
Data or rule hiding	Denotes to hide main data or rules of innovative data
K-anonymity	Achieve the anonymization via generalization and suppression
L-diversity	Retains the least group size of K and maintains the diversity of the sensitive quasi-attributes within the table
Taxonomy Tree	selectively generalize attributes to limit the information leakage
Randomization	A simple and valuable technique to hide the personal data in PPDM
Privacy protection	keeps the privacy; it should adapt data carefully to manage optimum data utility

Such an anonymization scheme must not only realize essential confidentiality requirements but also secure the data utility. Certainly, variants of K-anonymity are effective schemes of privacy protection in data mining. Conversely, many schemes confirmed that the data processed by these methods sometimes failed to thwart emerging attacks and are still vulnerable to internet phishing. As a result, the future privacy preserving data mining based K-anonymity methods need an advanced data infrastructure to maintain the combination of existing data functionality. This would certainly satisfy the requirements of all types of clients and communities. Also, the existing search algorithms are able to increase the speed of the retrieval process, but they do not scale up to bigger data because of the direct growth of response time with the extent of the searched databases. The suggested techniques for searching of distributed and massive datasets among multiple cloud providers must retain the ability to preserve privacy, must be efficient, scalable, good for integrity as well as utility, compatible.

5. CONCLUSIONS

An extensive overview of PPDM techniques based on k anonymization, randomization distortion, classification, clustering, association rule mining, distribution and outsourcing are discussed. It is well known that PPDM seems to be gradually shared owing to easy sharing of private data for analysis. Important merits and demerits of existing methods are stressed. Large volumes of sensitive data are often publicly available

across sectors such as health, military and are spread across Entities to Entities, Business to Businesses and Government to Government. Therefore, the preservation of privacy against disclosure attacks is of serious concern. Several organizations and governments across the world being totally dependent on information communications through the internet expressed severe concerns over privacy issues. Subsequently, the enormous development of IT faced fresh challenges to PPDM. Data mining retains the capability to extract a high degree of knowledge or interesting patterns from a huge amount of data requires complete security. The main objective of PPDM is to include the traditional data mining techniques in altering the data to hide sensitive information. The major issue is to transform the data and recover its mining result from the transformed one. Furthermore, the incompleteness of earlier studies showed us to engage in an extensive examination of the difficulties of distributed and published data for sharing and mining. So, the overhead for large-scale mining, preserving the privacy of data, the integrity of mining results, the data utility, the scalability and performance in the context of PPDM are surveyed. There is an urgent requirement to develop a robust, effective and scalable model to surmount these issues. In this respect, we recognized the gaps and weaknesses of existing works and examined them for further major enhancements, more vigorous privacy protection, and preservation. This comprehensive and useful review article is expected to serve as a taxonomy for steering and understanding the research advancements towards PPDM.

REFERENCES

[1] Sachan A, Roy D, Arun PV. 2013. An analysis of privacy preservation techniques in data mining. Advances in computing and information technology, Vol 3. Springer. pp. 119-128.

[2] Ciriani V, Vimercati SDC, Foresti S, Samarati P. 2008. K-anonymous data mining: a survey. Journal of data mining, Springer, New York. pp. 105-136.

[3] Aggarwal CC, Yu PS. 2008. A general survey of privacy-preserving data mining models and algorithms. Privacy preserving data mining, Chap 2. Springer, New York, pp. 11-52. http://doi.org/10.1007/978-0-387-48533.

[4] Sweeney L. 2002. Achieving k-anonymity privacy protection using generalization and suppression. International Journal of Uncertain Fuzziness Knowledge-Based System. 10(5): 571-588.

[5] Nergiz ME, Christopher C, Ahmet EN. 2009. Multi-relational k-anonymity. IEEE Transactions on Knowledge Data Engineering. 21(8): 1104-1117.

[6] Matwin S. 2013. Privacy-preserving data mining techniques: survey and challenges. Discrimination and privacy in the information society, Springer, Berlin, Heidelberg. pp. 209-221.

[7] Vatsalan D, Christen P, Verykios VS. 2013. Taxonomy of privacy-preserving record linkage techniques. Journal of Information Systems. 38(6): 946-969.

[8] Qi X, Zong M. 2012. An overview of privacy-preserving data mining. Procedia Environmental Science. 12(Icese 2011): 1341.-1347.

[9] Raju R, Komalavalli R, Kesavakumar V. 2009. Privacy maintenance collaborative data mining: a practical approach. Proceedings of 2nd international conference on emerging trends in engineering and technology (ICETET). pp. 307-311. http://doi. org/10.1109/ICETET.2009.184.

[10] Malina L, Hajny J. 2013. Efficient security solution for privacy-preserving cloud services. Proceedings of 36th international conference on telecommunications and signal processing (TSP). pp. 23-27. http://doi.org/10.1109/TSP.2013.6613884.

[11] Mukkamala R, Ashok VG. 2011. Fuzzy-based methods for privacy-preserving data mining. Proceedings of IEEE eighth international conference on information technology: new generations (ITNG).

[12] Kamakshi P. 2012. Automatic detection of the sensitive attribute in PPDM. Proceedings of IEEE international conference on computational intelligence and computing research (ICCIC).

[13] Islam MZ, Brankovic L. 2011. Privacy preserving data mining: a noise addition framework using a novel clustering technique. Journal of Knowledge-Based Systems. 24(8): 1214-1223.

[14] Kamakshi P, Babu AV. 2010. Preserving privacy and sharing the data in distributed environment using the cryptographic technique on perturbed data. 2(4).

[15] Li W, Liu J. 2009. Privacy preserving association rules mining based on data disturbance and inquiry limitation. Proceedings of 2009 Fourth International Conference on Internet Computer Science Engineering. pp. 24-29.

[16] Wang ET, Lee G. 2008. An efficient sanitization algorithm for balancing information privacy and

knowledge discovery in association patterns mining. Journal of Data Knowledge Engineering. 65(3): 463-484

[17] Belwal R, Varshney J, Khan S. 2013. Hiding sensitive association rules efficiently by introducing new variable hiding counter. Proceedings of IEEE international conference on service operations and logistics, and informatics.

[18] Naeem M, Asghar S, Fong S. 2010. Hiding sensitive association rules using central tendency. Proceedings of 6th international conference on advanced information management and service (IMS). pp. 478-484.

[19] Li X, Liu Z, Zuo C. 2009. Hiding association rules based on relative-non-sensitive frequent itemsets. Proceedings of 2009 8th IEEE International Conference Cognitive Informatics. pp. 384-389.

[20] Weng C, Chen S, Lo H. 2008. A novel algorithm for completely hiding sensitive association rules. Proceedings of Eighth international conference on intelligent systems design and applications, ISDA'08. 3: 202-208.

[21] Dehkordi MNM, Badie K, Zadeh AKA (2009), "A novel method for privacy preserving in association rule mining based on genetic algorithms", Journal of Software engineering 4(6):555–562.

[22] Kasthuri S, Meyyappan T. 2013. Detection of sensitive items in market basket database using association rule mining for privacy preserving. Proceedings of IEEE international conference on pattern recognition, informatics and mobile engineering (PRIME).

[23] Quoc H, Arch-int S, Xuan H, Arch-int N. 2013. Computers in industry association rule hiding in risk management for retail supply chain collaboration. Journal of Computing Technology. 64(7): 776-784.

[24] Xiong L, Chitti S, Liu L. 2006. k nearest neighbor classification across multiple sites. Proceedings of the 15th ACM international conference on information and knowledge management-CIKM'06. pp. 840-841.

[25] Singh MD, Krishna PR, Saxena A. 2010. A cryptography based privacy preserving solution to my cloud data. Proceedings of third annual ACM Bangalore conference.

[26] Baotou T. 2010. Research on privacy preserving classification data mining based on random perturbation. Xiaolin Zhang Hongjing Bi. pp. 1-6.

[27] Vaidya J, Clifton C, Kantarcioglu M, Patterson AS. 2008. Privacy-preserving decision trees over vertically partitioned data. ACM Trans Knowledge Discovery Data. 2(3): 1-27.

[28] Kantarcioglu M, Vaidya J. 2003. Privacy preserving naive Bayes classifier for horizontally partitioned data. Proceedings of IEEE ICDM workshop on privacy preserving data mining. pp. 3-9.

[29] Yi X, Zhang Y. 2013. Equally contributory privacy-preserving k-means clustering over vertically partitioned data. Journal of Information Systems. 38(1): 97-107.

[30] Giannotti F, Lakshmanan LVS, Monreale A, Pedreschi D, Wang H. 2013. Privacy-preserving mining of association rules from outsourced transaction databases. IEEE Journal of Computer Systems. 7(3): 385-395.

[31] Worku SG, Xu C, Zhao J, He X. 2014. Secure and efficient privacy-preserving public auditing scheme for cloud storage. Journal of Computer and Electrical Engineering. 40(5): 1703-1713.

[32] Arunadevi M, Anuradha R. 2014. Privacy preserving outsourcing for frequent item-set mining. International Journal of Innovative computing and Communication Engineering. 2(1): 3867-3873.

[33] Kerschbaum F, Julien V. 2008. Privacy-preserving data analytics as an outsourced service. Proceedings of the 2008 ACM workshop on secure web services.

[34] Li Y. 2013. Privacy-Preserving and Reputation System in Distributed Computing with Untrusted Parties. Pro-Quest LLC (2013). Copyright in the Dissertation held by the Author. Microform Edition © ProQuest LLC. All rights reserved.

[35] Dev H, Sen T, Basak M, Ali ME. 2012. An approach to protect the privacy of cloud data from data mining based attacks. Proceedings of IEEE 2012 SC companion high-performance computing, networking, storage and analysis (SCC).

[36] Tassa T. 2014. Secure mining of association rules in horizontally distributed databases. IEEE Transactions on Knowledge and Data Engineering. 26(4): 970-983.

[37] Dong R, Kresman R. 2009. Indirect disclosures in data mining. Proceedings of the Fourth international conference on the frontier of computer science and technology, FCST'09.

[38] Inan A, Saygin Y. 2010. Privacy preserving spatiotemporal clustering on horizontally partitioned data. Lecture Notes in Computer Science (including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), 6202 LNAI. pp. 187-198.

[39] Nanavati N, Jinwala D. 2012. Privacy preservation for global cyclic associations in distributed databases. Procedia Technology. 6: 962-969.

[40] Agrawal R, Srikant R. 2000. Privacy-preserving data mining. Proceedings of the 2000 ACM SIGMOD international conference on management of data-SIGMOD '00. 29(2): 439-450.

[41] Wang H, Hu C, Liu J. 2010. Distributed mining of association rules based on privacy-preserved method. Proceedings of International symposium on information science and engineering (ISISE). pp. 494-497.

[42] Nguyen XC, Le HB, Cao TA. 2012. An enhanced scheme for privacy-preserving association rules mining on horizontally distributed databases Proceedings of IEEE RIVF international conference on computing and communication technologies, research, innovation, and vision for the future. pp. 1-4.

[43] Ibrahim A, Jin H, Yassin AA, Zou D. 2012. Towards privacy-preserving mining over distributed cloud databases. Proceedings of IEEE second international conference on cloud and green computing (CGC).

[44] Patel S, Garasia S, Jinwala D. 2012. An efficient approach for privacy preserving distributed K-means clustering based on Shamir's algorithm. pp. 129-141.

[45] Nix R, Kantarcioglu M, Han KJ. 2012. Approximate privacy-preserving data mining on vertically partitioned data. Journal of Data and applications security and privacy, Springer, Berlin, Heidelberg. pp. 129-144.

[46] Machanavajjhala A, Kifer D, Gehrke J, Venkitasubramaniam M. 2007. L-diversity: privacy beyond k-anonymity. ACM Transactions on Knowledge Discovery in Databases. 1(1):3

[47] Loukides G, Gkoulalas-divanis A. 2012. Expert systems with applications utility-preserving transaction data anonymization with low information loss. Journal of Expert Systems and Application. 39(10): 9764-9777.

[48] Friedman A, Wolff R, Schuster A. 2008. Providing k-anonymity in data mining. VLDB J 17(4):789-804.

[49] He Y, Siddharth B, Jeffrey FN. 2011. Preventing equivalence attacks in updated, anonymized data. Proceedings of IEEE 27th international conference on data engineering (ICDE).

[50] Soodejani AT, Hadavi MA, Jalili R. 2012. K-anonymity-based horizontal fragmentation to preserve privacy in data outsourcing. Data and applications security and privacy XXVI. Springer, Berlin, Heidelberg. pp. 263-273.

[51] Karim R, Rashid M, Jeong B, Choi H. 2012. In Transactional Databases. pp. 303-319.

[52] Loukides G, Gkoulalas-Divanis A, Shao J. 2012. Efficient and flexible anonymization of transaction data. Journal of Knowledge and Information Systems. 36(1): 153-210.

[53] Vijayarani S, Tamilarasi A, Sampoorna M. 2010. Analysis of privacy preserving k-anonymity methods and techniques. Proceedings of IEEE international conference on communication and computational intelligence (INCOCCI). pp. 540-545.

[54] Tai C-H, Huang J-W, Chung M-H. 2013. Privacy preserving frequent pattern mining on multi-cloud environment. Proceedings of 2013 international symposium on biometrics and security technologies (ISBAST).

[55] Tai C-H, Yu PS, Chen M-S. 2010. k-Support anonymity based on pseudo taxonomy for outsourcing of frequent item-set mining. Proceedings of the 16th ACM SIGKDD international conference on knowledge discovery and data mining.

[56] Pan Y, Zhu X, Chen T. 2012. Research on privacy preserving on K-anonymity. Journal o Software. 7(7): 1649-1656.

[57] Deivanai P, Nayahi J, Kavitha V. 2011. A hybrid data anonymization integrated with suppression for preserving privacy in mining multi-party data.

Proceedings of IEEE international conference on recent trends in information technology (ICRTIT).

[58] Monreale A, Pedreschi D, Pensa RG, Pinelli F. 2014. Anonymity-preservingsequential pattern mining. Journal of Artificial Intelligence. 22(2): 141-173.

[59] Nergiz ME, Gok MZ. 2014. Hybrid k-anonymity. Journal of Computer Security. 44: 51-63.

[60] Nergiz ME, Gok MZ, ozkanli U. 2013. Preservation of utility through hybrid k-anonymization. Proceedings of Trust, privacy and security in digital business. Springer, Berlin, Heidelberg. pp. 97-111.

[61] Zhang X, Yang LT, Member S, Liu C, Chen J. 2014. A scalable two-phase top-down specialization approach for data anonymization using Map Reduce on thecloud. 25(2): 363-373.

[62] Zhang X, Liu C, Nepal S, Chen J. 2013. An efficient quasi-identifier index-based approach for privacy preservation over incremental data sets on thecloud. Journal of Computer Systems. 79(5):542-555.

[63] Zhang X, Liu C, Nepal S, Chen J. 2013. An efficient quasi-identifier index-based approach for privacy preservation over incremental data sets on the cloud. Journal of Computer Systems. 79(5): 542-555.