Sensitive and Private Data Analysis: A Systematic Review

Syeda Sana e Zainab
Insight Centre for Data Analytics, University College Dublin
Dublin, Ireland
syeda.sanaezainab@insight-centre.org

Tahar Kechadi
Insight Centre for Data Analytics, University College Dublin
Dublin, Ireland
tahar.kechadi@insight-centre.org

ABSTRACT

Each day an extensive amount of data is produced from various organisations, such as e-commerce, IT, hospitals, retail and supply chain, etc. Due to the expansion of computer devices and advances in technology this immense amount of data has been collected and analysed to support decision making. The examination of such data is advancing businesses and contributing advantageously to society in numerous diverse areas. However, serious privacy concerns are raised due to the storage and flow of potentially sensitive data [31]. Strategies that permit the knowledge extraction from the data, while protecting privacy, are known as privacy-preserving data mining (PPDM) techniques. This paper surveys the analysis of private and sensitive data using various PPDM algorithms and techniques. We also highlighted their advantages and limitations within various contexts.

KEYWORDS

privacy, sensitive data, privacy preservation data mining

ACM Reference Format:
Syeda Sana e Zainab and Tahar Kechadi. 2019. Sensitive and Private Data Analysis: A Systematic Review. In *3rd International Conference on Future Networks and Distributed Systems (ICFNDS '19), July 1–2, 2019, Paris, France.* ACM, New York, NY, USA, 11 pages. https://doi.org/10.1145/3341325.3342002

1 INTRODUCTION

The process of extracting, purifying, transforming, and demonstrating data with the objective to gain useful insights, getting conclusions, and supporting decision making is known as *data analysis process.* Data analysis has various aspects and methodologies, including various techniques, and is utilised in various business, banking and science areas. Nowadays, data analysis plays a vital role in data-driven organisations for making decisions and helping them to work more effectively [52]. The approach of data analysis that mainly focuses on modelling and knowledge discovery for prediction is known as *data mining.* In this paper, we are focusing on the analysis of private and sensitive data using data mining techniques.

In this digital era, a large amount of information is generated every moment because of the intensive utilisation of digital technology. The mining and analysis of such information are beneficial in many areas such as banking, health, e-commerce, agriculture and many more. However, the increase of data collections raised data privacy concerns due to the sensitive private nature of the data[11, 30]. In early research, privacy has been perceived as an individual privilege with limited scope[21] (e.g.; at home, with family and friends). The problem arises when privacy comes as a result of the broadness of areas where it applies [65]. The scope of the privacy covers four categories:

- Privacy that concerns the consolidation and handling of personal data is known as *information privacy* or *data privacy.*
- Privacy that concerns the individual physical being against invasive procedures, such as drug testing etc., is known as *bodily privacy.*
- Privacy that concerns the security of any form of communication (e.g.; telephone, email, etc.) is known as *privacy of communication.*
- Privacy that concerns the incursion of physical boundaries is known as *territorial privacy.*

In this paper we are focused on the data privacy. This includes systems that collect, analyse and publish data. First of all let's define the concept of data privacy, as it is a general term used in many areas and having different meanings. In the case of HIPAA[1], the individual tendency to control access to their health information from outside world is known as privacy. Similarly in organisations, privacy contributes a major part to policy-making for data collection, utilisation,

ICFNDS '19, July 1–2, 2019, Paris, France
© 2019 Association for Computing Machinery.
ACM ISBN 978-1-4503-7163-6/19/07…$15.00
https://doi.org/10.1145/3341325.3342002

[1] Health Insurance Portability and Accountability Act

manipulation and how clients are informed and involved in this process. Bertino [8] and Westin [61] described private data as *"the information of individual present in digital databases to be protected from any unauthorised disclosure"*. Ruth Gavison [19] defined three elements of privacy which are anonymity, secrecy and solitude. Walters [60] and Schoeman [51] defined privacy as *"Control of data to release from outside world"*. Moreover, many other privacy definitions have been introduced and the majority of them defined it in terms of control and security[29]. Hence, one can conclude that the principal notion of private data is to control and secure the consolidating and handling of data. In this paper, we are focus on analysis and mining of private-sensitive data which should be protected and cannot be exposed to any un-authorised parties.

In order to preserve the privacy of data numerous research studies have been conducted that resulted in the introduction of many data mining algorithms and techniques. Privacy preserving data mining (PPDM) is one of the techniques for privacy preservation of sensitive information. However, these techniques are used to maximise the utility of the data while preserving its privacy. In recent years, the PPDM techniques have drawn a huge recognition amongst computer scientist researchers. These techniques have been evaluated through several metrics. In this paper we will discuss various PPDM algorithms and techniques, their advantages and limitations. The remainder of this paper is organised as follows. Section 2 discusses the definitions of private and sensitive data along with their identification scenarios. Section 3 describes the PPDM framework and algorithms. In Section 4 PPDM techniques are discussed. While evaluation metrics for such approaches are presented in Section 5. Section 6 concludes the paper.

2 PRIVATE AND SENSITIVE DATA

With the abundance of data in every sector of the economy, many studies, which were not possible in the past, are now become possible. Therefore, many researchers perform data analyses on individuals socioeconomic trends to help in understanding certain phenomena of governing the data. This may lead to privacy concerns as data disclosure should be taken into account. Consider an example of patient data in healthcare. Before performing data mining on such data it is very important that only the right users and tools have access to it. While data mining privacy should be carried out at all levels, both privacy and security are essential yet impediment to data mining. In the following, we first describe the concept of data privacy and security and then its identification into data.

Definition

Although everyone knows what is security and privacy but there is no universally accepted definition. There is always ambiguity between personal, private, and sensitive data. Similarly privacy and security are often used interchangeably in different circumstances, as both are related to each other but they are entirely separate concepts. Figure 1 shows personal, private, and sensitive data in a form of venn diagram. Here we tried to define the concept of personal, private ans sensitive data.

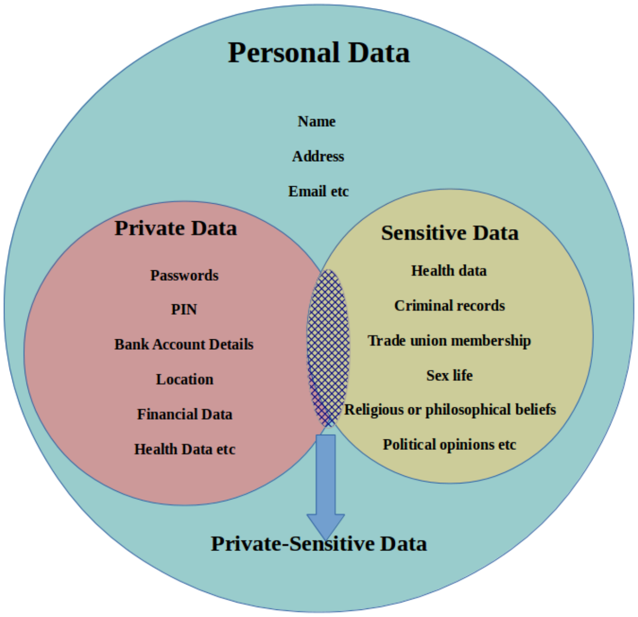

Figure 1: Personal, Private, and Sensitive data.

Personal Data: Any data that explicitly contains information such as name, address, phone numbers, email, etc., is known as personal data [13]. In other words, it is any information about a specific individual ranging from somebody's name to their physical appearance that reveals their identity directly or indirectly.

Private Data: Any personal data that individual don't want to reveal in public is known as private data[13]. Private data is a part or subset of personal data that the individual does not want to share with outside world. Privacy is all about to *"freedom of making decision on your own data that who will see what data"*[6]. It is considered a basic right that is important to safety, security and quality of life.

Sensitive Data / Sensitive Personal Data: Any information that uncovers racial or ethnic source, genetic information, political sentiments, religious or philosophical convictions, bio metric information with the end goal of recognising any individual information that concerns the well being or individual's sexual life and orientation is known as sensitive data or sensitive personal data[40].

Private-Sensitive Data: Apart from personal data, there are several forms of data that are considered as private-sensitive data like organisations data. This information is not subject to a specific individual but it is important that you cannot expose it to the entire world[13]. For example its damaging for an organisation if their sales record fall into wrong hands. Similarly simple thing like hospital appointment may not be expose to everyone. So you should be cautious with both privacy-sensitive and personal data.

In order to explain the above data concepts in mathematical model we are defining D_{per} as a Personal data composed of a set of Private D_{pri} and Sensitive data D_s. Where private-sensitive data D_{ps} is intersection of Private D_{pri} and Sensitive data D_s. Then they are mathematically modelled as:

$$D_{per} = (D_{pri}, D_s)$$
$$D_{pri} \subsetneq D_{per}$$
$$D_s \subsetneq D_{per}$$
$$D_{ps} \iff D_{pri} \cap D_s$$

Private and Sensitive Data Identification

In our environment private and sensitive data exist in many ways and its really challenging to identify them. This section cover techniques to identify private and sensitive data in five of the most common scenarios. But first we will discuss the types of data identifiers

Direct identifiers: Such data that identify any individual directly are known as direct identifiers. For example name, mailing address, phone number, email, social security numbers or driver's license numbers, bio-metric data, IP addresses, photographs, audio recordings etc. Before sharing with the public they are typically removed from data sets.

Indirect identifiers: Such data that is not directly showing identity of some individual but on combining it with other data they can point to someone is known as indirect identifiers. For example doctor's name, gender, rare disease, place of birth, annual income, general geographic indicators like ethnicity, birth year, postal code etc. It is highly recommended that if dataset contains more than three indirect identifiers should be analysed by ethical team to ensure the privacy risk.

In Table 1 we have discussed five scenarios of private and sensitive data identification. We have proposed various powerful techniques for data identification. However, these automated methods are imperfect, and you will want to consider maintaining a governance policy to deal with any sensitive information that remains after scrubbing.

3 PRIVACY AND DATA MINING

Every moment a large amount of digital data has been collected from several domains and than analyse using various data mining techniques. Few of such domains needs to handle and published private sensitive data (e.g. in health care services each day thousands of medical records has been produced), that increases the risk of private data disclosure[15].

Now a days Privacy Preserving Data Mining (PPDM) techniques are becoming popular for knowledge extraction in large dataset(s) while preserving the privacy. By using some of these techniques the original data has been modified or removed in order to preserve privacy[2]. This may reduces the quality of dataset. In order to ensure certain level of privacy PPDM methods are build to maximise the utility of the data by applying effective data mining.

Privacy Preserving Data Mining(PPDM) Framework

In [56] PPDM framework has defined in three levels. In first level raw data is extracted from dataset(s) and transformed as desired. In second level privacy is guaranteed by implication of data mining algorithms and techniques. The result of data mining algorithm is present in third level[3]. Figure 2 shows all levels.

Level 1: This level is all about data collection and transformation. First data is collected from dataset(s) and than transformed according to the desired requirement. Data engineers have done different techniques to extract the raw data and transformed, to make it suitable for analysis purpose while securing the data privacy[20].

Level 2: The data sanitization is a key goal of this level. Data has processed using techniques like suppression, blocking, perturbation, modification, generalisation, sampling etc. After that data mining algorithms are applied for knowledge discovery and securing the privacy.

Level 3: In this level disclosure risks are addressed by further checking the processed data for its privacy and sensitivity concerns before revealing it to the public.

Taxonomy of Privacy Preservation Data Mining Algorithms

PPDM algorithms are classified into central database and distributed database under the data distribution characteristic. The difference between centralised database and and distributed database is defined as follow:

- The database that is owned by private party is known as centralised database.
- The database that is owned by more than one party who want to perform data mining on combined data is

Table 1: Private and Sensitive Data Identification Scenarios

	Scenario	Identification	Example
1	Private and Sensitive data in columns	Columns that contains private and sensitive information will be identified, secured and documented	User's name, email and mailing address
2	Private and Sensitive data in unstructured text-based data	Regular expression pattern or complex tools like the Google Data Loss Prevention API (DLP API)[44] are used for identification	Credit card numbers, bank account data
3	Private and Sensitive data in free-form unstructured data	**In free text** API such as Cloud Natural Language API[24] can be used to identify private and sensitive data in free text. **In audio recordings** you can use Cloud Speech API[28] which is speech-to-text service for identification. **In images** Cloud Vision API[41] can be used to detect raw text from the image. After that private and sensitive data can easily be identified. **Video data** consist of both audio and image files. Cloud Video Intelligence API[23] can be used for video processing with Cloud Speech API in order to process audio.	Text reports, audio recordings, photographs, or scanned receipts
4	Private and Sensitive data in a combination of fields	Identification required statistical expertise for scrubbing the data to inspect the raw data for potential problems	Combining zip code with address
5	Private and Sensitive data in unstructured content	Standard machine learning algorithms are used to convert the unstructured data into a semi-structured format. After that data mining techniques are used to identify private and sensitive data which we will discuss later in this article.	Chat transcript

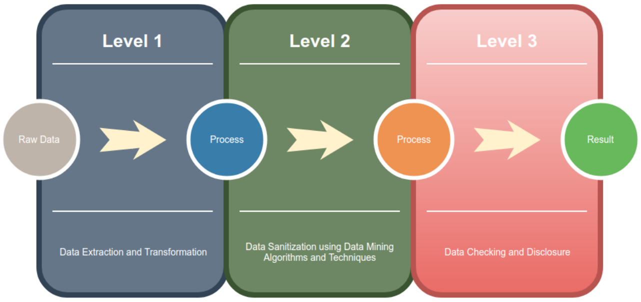
Level 1
Level 2
Level 3
Raw Data
Process
Process
Result
Data Extraction and Transformation
Data Sanitization using Data Mining Algorithms and Techniques
Data Checking and Disclosure

Figure 2: PPDM Framework

known as distributed database. This distribution can be vertical or horizontal.

The classification of PPDM algorithms according to the hiding purpose is lie into data hiding and rule hiding. Whereas data hiding refers to hiding sensitive data from the public disclosure, and rule hiding refers to hiding information derived after applying data mining algorithms.

In order to secure the data PPDM techniques changes the original data. However these techniques are classified into following five dimension[58] which we will explain in the next section.

(1) Data distribution
(2) Data modification
(3) Data mining algorithms
(4) Data or rule hiding
(5) Privacy preservation

In [7] a taxonomy of the existing PPDM algorithms are defined according to above dimensions which are shown in Figure 3. However this is not the complete representation of all PPDM algorithms but it can give an overview of so far proposed PPDM algorithms according to their features. For securing the privacy in centralised databases, heuristic and reconstruction based techniques are used while in distributed databases cryptography based algorithms are designed. using encryption techniques. Many heuristic based algorithms for securing aggregated and raw data are using hiding techniques like blocking, swapping, aggregation , perturbation etc. In reconstruction based algorithms for securing raw data techniques like perturbation based on probability distributions are used. Moreover cryptography based algorithms used encryption techniques for securing raw and aggregated data by using classification, clustering and association rule mining approaches. Now, we briefly describe some of the algorithms proposed in the PPDM area.

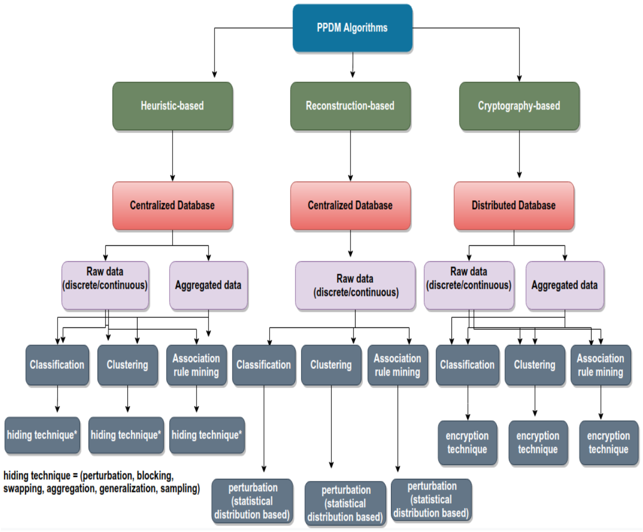

Figure 3: PPDM Algorithms Taxonomy

Heuristic based Algorithms. The complexity problem has raised when we are dealing with various data mining techniques like classification, association rule discovery and clustering, for securing the centralised data. For such complexity issues heuristics can be used.

A heuristic-based algorithm has been proposed by [43] for privacy preservation in mining frequent item-sets. Set of frequent patterns that consist of sensitive information are hide and set of sanitised algorithms are proposed. These algorithms are known as non-perturbative algorithms because they remove data from a transactional database, and doesn't modify the original data by adding noise like perturbative algorithms. To avoid the noise addition and restriction on removal of original data these algorithms utilised item restriction approach. In order to show the effectiveness and the efficiency of their algorithms they have made an evaluation based on following measures.

Hiding Failure: It is measured in terms of the amount of restrictive patterns discovered from the sanitised database.

Misses Cost: It is measured in terms of the amount of non-restrictive patterns that are hidden after the sanitization process.

Artifactual Pattern: It measured in terms of the amount of patterns that are discovered as artifacts.

The percentage of non sanitized sensitive transactions are represented as disclosure threshold ϕ. It allows one to find a balance between the misses cost and hiding failure. Moreover, to measure dissimilarity between the original and sanitized databases Oliveira[43] proposed three different methods(i.e. difference between sizes, content and frequency histogram). Data modification approaches like data blocking is the one which have been used for association rule confusion[12, 50]. In this approach certain attributes present in sensitive data are replaced with a question mark. In some areas(e.g. medical) it is desirable to replace an original data by an unknown data rather than placing wrong data. In [5] the limitations of achieving optimal sanitization has been discussed for the hiding of sensitive large dataset(s) in the context of association rules discovery.

Reconstruction based Algorithms. In [4] based on an Expectation Maximisation (EM) algorithm a reconstruction based approach proposed for distribution reconstruction. This approach meets the maximum possibility calculated of the original distribution on the perturbed data. EM algorithm gives robust estimates of the original distribution when the amount of data is big. Some other work [17, 18, 47] has been proposed for mining association rules from transactions of binary and categorical items, where data randomisation has been used for preserving privacy while keeping the high utility of dataset.

Cryptography based Algorithm. In regarding the privacy preserving data mining algorithms, there are quite numbers of cryptography based approaches have been proposed. Using this approach a work [25] have been proposed that mainly addressed the secure mining issue of association rules over horizontally partitioned data. Using the commutative encryption each party first encrypt its own dataset(s). Then the transmission of data has done in a way that first party send its frequency count, plus a random value to its neighbour. Than the second party adds its frequency count and passes it on to other parties. At the end, a comparison between the first and last party take places to find out if the final result is greater than the threshold plus the random value. The proposed methods are evaluated in terms of communication and computation costs. Another work [57] has been described by that addressed the secure mining issue of association rules over vertically partitioned data. The aim of this approach is to secure the contents of individual transactions by determining the item frequency when transactions are split across different sites.

4 PRIVACY PRESERVATION DATA MINING TECHNIQUES

The main objective of PPDM is to change the real data while keeping the privacy. We have already defined in previous section about the five dimensions on which PPDM techniques are classified[58]. Here we briefly define them.

(1) **Data distribution:** This dimension described the distribution of data. Data can exist in two forms centralised or distributed. There are two types of distributed data: 1. Horizontal distribution[10] in which different data is distributed in different sites, 2. Vertical distribution[16] in which different attributes values are present in different sites.
(2) **Data Modification:** As shown by name that this dimension is related to change the data for public disclosure while preserving its privacy. There are five methods of data modification [26], 1. Perturbation in which attribute value is changing with new value, 2. Blocking: in which is the existing attribute value is replace with a âĂIJ?âĂİ, 3. Swapping in which different records are interchanged, Sampling in which data has been shown to sample of population and Encryption in which data is encrypted using various cryptography techniques.
(3) **Data Mining Algorithms:** This dimension is related to the data mining algorithms like association rule mining, classification mining, Bayesian networks and clustering etc, on which PPDM techniques are classified.
(4) **Data hiding:** This dimension related to hiding raw data or aggregated data as required for privacy concerns.
(5) **Privacy Preservation:** The techniques used for privacy preservation lie in this dimension.

Based on the work[26, 53] and above dimensions, various PPDM techniques are classified into following categories[56, 58]. The evaluation of these techniques are presented in next Section.

Data Anonymization

The risk of identity disclosure has been reduced by using data anonymization. A dataset consist of personal information like name, address that explicitly identifies an individual, sensitive attributes that contains specific individual data e.g. salary, disease, and Quasi identifiers that identify an individual by combining with publicly available data e.g. gender, zip code etc. By using anonymization approach individual private data is to be secure by removing explicit identifiers. But still there is a probability of breaching privacy when quasi identifiers are combined with publicly available data. This is known as linking attack. In order to target such problem for privacy preservation, k-anonymity[48] model has been proposed that uses generalisation and suppression. To implement k-anonymity various algorithms[3, 33, 48] have been proposed using generalisation and suppression approach. Similarly models like t-closeness[34], l-diversity[38], p-sensitive k- anonymity[54], M-invariance[63], (a, k)- anonymity [62] etc. were proposed to handle the problem of k-anonymity.

K-anonymity. Every record present in anonymized table with respect to a set of Quasi attributes(QI), must be dissimilar to at least other k-1 records. In order to get the k-anonymity techniques like suppression or generalisation are used[48]. However this technique is prune towards linking attack, homogeneity attack i.e. anonymized table contains similar sensitive attributes values and background knowledge attack i.e. having knowledge of linking between QI and sensitive attributes.

L-diversity . In order to solve the homogeneity attack caused by K-anonymity technique this technique has been proposed[38] that not only focus on preserving smallest size of K group but also sensitive attributes of each group. By using this technique, every sensitive attribute of each group must consist of l well defined values.

T-closeness . The main disadvantage of L-diversity technique's is that it lacks the background knowledge of values of each attribute and treat them all in similar manner, that causes background knowledge attack. While in real the sensitivity level of each attribute values are different. In this technique[34] t threshold is used which should always remain lesser than the distance between distribution of sensitive attribute in table and its distribution in an anonymized group.

Perturbation

In perturbation the sensitive data values are replaced with some synthetic/false data values so the computed statistical data achieve after perturbation will not much differ from the original data. The privacy has preserved by preserving statistical attributes of data. There are two different approaches to achieve perturbation as shown in Figure 4. These approaches are carried out in both centralised and distributed dataset(s).

Value-Based Perturbation. In **value-based perturbation** a random noise is added to the data values. This approach is known as **random noise addition**. One approach is to add Gaussian noise[4] to the private attributes of dataset(s). While in **randomised response** approach data is jumble up so that the receiver cannot shows the estimated probabilities better than the defined threshold. The receiver has the ambiguity if sender send correct or false data.

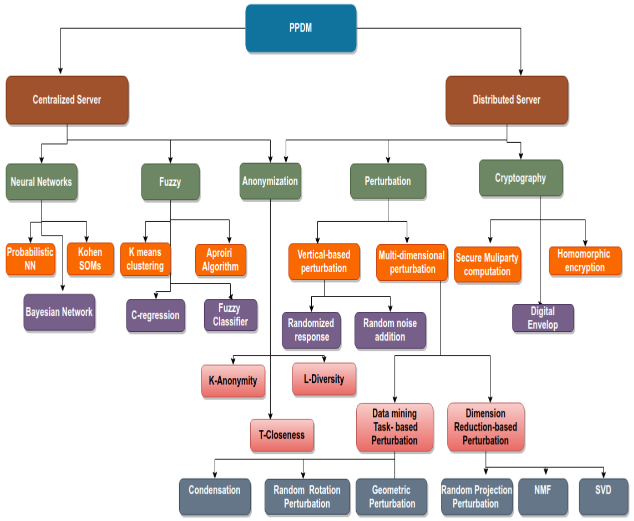

Figure 4: PPDM Techniques Classification

Multidimensional Perturbation. A **multidimensional perturbation** [35] is another technique for satisfying personal privacy. In this technique one of the utilise approach is **data mining task-based perturbation** in which data is modify by applying data mining algorithms directly. Here we discussed few approaches used in data mining task-based perturbation:

- **Condensation** is the one to achieve this task-based perturbation. First the data is condensed in size K groups than the statistical data of each group of records is preserved. After that anonymized data is generated having statistical characteristic similarity as of original dataset. Simple classifier for the K Nearest Neighbor (KNN) [1] can be generated using this technique.
- In multi-dimensional space in order to preserve the euclidean distance, inner product and geometric shape hyper the technique that is used in known as **random rotation perturbation**.
- The technique that combines noise, rotation and translation is known as **geometric perturbation technique**. In this technique limitation of rotation is express with the component ψ and ∇. The addition of noise addressed the distance-inference attack[14] that was unaddressed in random rotation perturbation technique.

While preserving the significant data pattern in dataset(s) by obtaining a compact representation with reduced-rank a technique known as **dimension reduction-based perturbation** has been used. The privacy of dimension and every attribute value of original data has guaranteed in this technique. Here we discussed few approaches used in dimension reduction-based perturbation:

- The idea of **random projection perturbation technique** [37] has been proposed by [18]. It shows the possibility to maintain both distance-related statistical properties with dimension reduction in dataset(s) together. Data mining tasks like euclidean distance estimation, classification, correlation matrix computation, and clustering etc can be obtained using this technique.
- In data mining **singular value decomposition(SVD)** [59] is a known method for dimension reduction.
- In order to obtain data representation using non negative constraints **non-negative matrix factorization (NMF)** [32] technique has used, which is matrix factorization method.

The disadvantages of perturbation can be managed by cryptography technique for privacy preservation.

Data Randomization

Randomization method was proposed by [42] for solving survey problem. In this technique the data collected from all individuals are jumbled. For larger groups of individuals the aggregation of their data gives more accuracy. The result of randomization is the data on which no body can trust if it is true or false data. The advantage of this technique is that its cheap and effective approach for PPDM and usually used in surveys data that contains answers having privacy concerns.

Cryptography

The technique in which data has been secured in a way that only target receiver(s) can interpret and process the private records. Basically it is a secure transmission of data between two parties in presence of third party. Cryptography technique is used in scenarios when results are computed by multiple parties(e.g. multiple hospitals are doing joint research on some disease) because it provides privacy framework as well as several cryptographic algorithms are available to implement PPDM. Several PPDM techniques are introduced under the umbrella of cryptography. Here we are discussing some of them. In [25] a cryptographic technique has been proposed to decreased the data and overhead caused by sharing. This technique aims in association rule mining over horizontally partitioned data. Another cryptographic approach proposed by [64] for horizontal partitioned data to achieve privacy with great accuracy. [36] proposed a cryptographic protocol for ID3 decision trees generation. Although cryptography techniques are efficient and covers the limitation of perturbation techniques but it can also breaches the privacy when multiple parties are involved.

Fuzzy Based

Clustering has been used in fuzzy based technique to achieve anonymization in PPDM for data protection. Similar records are merged into clusters so that each cluster is different from other and records in each cluster is distinct from other cluster records. In [22] k-means clustering has been proposed to achieve anonymization using fuzzy. Another approach [9] used fuzzy based c regression for synthetic data generation after that statistical computation is made with less data loss.

Neural Network

A probabilistic neural network approach has been proposed by [27] for peer-to-peer data mining. Using this approach peer member having best of weight-based is selected. For learning distribution of data [49] proposed a Bayesian network approach using neural networks. A protocol that utilised vertical partitioned data for Bayesian networks is proposed by [55]. The protocol gives better perform ace with minimum overhead.

5 EVALUATION

A lot of work has been done to define various parameters for PPDM algorithms evaluation. Here we defined following parameters based on the work[7, 39, 45, 46, 58].

- **Efficiency:** It is the ability of PPDM algorithm to utilise its resources with better performance;
- **Scalability:** This parameter about the ability of algorithm to deal with the increasing size of dataset(s) on which it is applied to mined the data and ensure privacy;
- **Quality of data:** This parameter shows the quality of original data suffers by the algorithm after the application of algorithm;
- **Hiding failure:** This parameter is evaluated by the unhidden data that remains after applying the algorithm;
- **Privacy level/Uncertainty level:** It is the probability of finding sensitive data after application of PPDM algorithm;
- **Performance:** It is the amount of time taken by algorithm to preserve data privacy;
- **Data utility:** It is the measure of data loss or else the loss in the functionality of the data;
- **Resistance:** The tolerance degree of algorithm against other data mining algorithms and models.

We have showed the evaluation of PPDM techniques based on the work[26, 53, 56] in Table 2. The table iterates and summarises the discussion in previous sections on PPDM techniques along with their advantages and limitations. This evaluation shows that each technique is better than others in different ways. For example anonymization technique is good to hide individual identity and its sensitive data but this approach is only for centralised data. In perturbation technique sensitive attributes are preserve independently but it causes loss of original information. Randomization technique is more efficient than cryptography but it is not suitable for multiple sensitive attributes in databases. For private streaming data condensation approach gives better results but it may causes changes in data formats. In cryptography privacy is more preserved as compare to randomization technique but when multiple parties are involved than it breaches the privacy. Hence we conclude that none of any technique gives optimum solution to preserve privacy without information loss.

6 CONCLUSION

Due to the production of extensive amount of digital data each day, businesses and institutions are collecting and perform analysis for decision making purpose. Sometimes it requires publishing or sharing of sensitive data. It is a great challenge to protect the private sensitive data while performing the analysis and mining tasks. However it is really hard to find out the optimum solution to secure the private data without any information loss and disclosure of such data. For this issue Privacy Preserving Data Mining (PPDM) methods allows us to extract the private sensitive data while preserving the individuals privacy.

In this paper we have performed a systematic review of the proposed methods and techniques for analysis of private sensitive data. Moreover PPDM framework along with proposed taxonomy of algorithms are addressed. We have made an effort to review a decent number of existing PPDM algorithms and techniques. Finally, we come to this point that there does not exists a single privacy preserving data mining algorithm that outperforms all other algorithms on all possible criteria like efficiency, scalability, data quality, performance etc. Different algorithm may perform better than another on one particular criterion. Similarly none of any PPDM technique provides optimum solution to preserve privacy without changing the originality of data that causes the information loss. Hence this is real life issue to address the deficiency in performance and practical implementation of privacy preserving algorithms and techniques. Standardisation of the evaluation criteria/parameters for PPDMs also requires further investigation as the current criteria/parameters utilised by researchers is not uniform.

ACKNOWLEDGMENTS

This work is supported by Science Foundation Ireland under grant number SFI/12/RC/2289.

Table 2: PPDM Techniques Evaluation

Technique	Methods Utilised	Data Distribution	Data Mining Tasks	Advantages	Limitations
Anonymization	Generalisation Suppression, Permutations	Centralised data	Association Rule, Clustering, Classification	Individual identity and associated sensitive data are to be hidden	Linking attack
Perturbation	Adding Noise, Data Swapping, Micro aggregation	Centralised and Distributed data	Association Rule, Clustering, Classification	This technique allows to preserve sensitive attributes independently	Original data values cannot be regenerated hence information loss
Randomization	Adding Noise, Scrambling, Re sampling	Centralised and Distributed data	Classification	Better efficiency compare to cryptography	This technique is not suitable for multiple sensitive attribute databases
Condensation	Aggregation	Centralised and Distributed data	Classification	Use pseudo data rather than altered data. Provide better results of stream data	Data format problems
Cryptography	Secure Multi-party Computation (SMC), Encryption	Centralised and Distributed data	Association Rule	Transformed data are exact and protected. Provides better privacy as compare to randomized approach	Breaches the privacy when multiple parties are involved

REFERENCES

[1] Charu C Aggarwal and S Yu Philip. 2004. A condensation approach to privacy preserving data mining. In *International Conference on Extending Database Technology*. Springer, 183–199.

[2] Charu C Aggarwal and S Yu Philip. 2008. A general survey of privacy-preserving data mining models and algorithms. In *Privacy-preserving data mining*. Springer, 11–52.

[3] Charu C Aggarwal and S Yu Philip. 2008. *Privacy-preserving data mining: models and algorithms*. Springer Science & Business Media.

[4] Dakshi Agrawal and Charu C Aggarwal. 2001. On the design and quantification of privacy preserving data mining algorithms. In *Proceedings of the twentieth ACM SIGMOD-SIGACT-SIGART symposium on Principles of database systems*. ACM, 247–255.

[5] Mike Atallah, Elisa Bertino, Ahmed Elmagarmid, Mohamed Ibrahim, and Vassilios Verykios. 1999. Disclosure limitation of sensitive rules. In *Proceedings 1999 Workshop on Knowledge and Data Engineering Exchange (KDEX'99)(Cat. No. PR00453)*. IEEE, 45–52.

[6] David Banisar and Simon Davies. 1999. Global trends in privacy protection: An international survey of privacy, data protection, and surveillance laws and developments. *J. Marshall J. Computer & Info. L.* 18 (1999), 1.

[7] Elisa Bertino, Igor Nai Fovino, and Loredana Parasiliti Provenza. 2005. A framework for evaluating privacy preserving data mining algorithms. *Data Mining and Knowledge Discovery* 11, 2 (2005), 121–154.

[8] Elisa Bertino, Dan Lin, and Wei Jiang. 2008. A survey of quantification of privacy preserving data mining algorithms. In *Privacy-preserving data mining*. Springer, 183–205.

[9] Isaac Cano and Vicenç Torra. 2009. Generation of synthetic data by means of fuzzy c-Regression. In *2009 IEEE International Conference on Fuzzy Systems*. IEEE, 1145–1150.

[10] Stefano Ceri, Mauro Negri, and Giuseppe Pelagatti. 1982. Horizontal data partitioning in database design. In *Proceedings of the 1982 ACM SIGMOD international conference on Management of data*. ACM, 128–136.

[11] Yoan Chabot, Aurélie Bertaux, Christophe Nicolle, and M-Tahar Kechadi. 2014. A complete formalized knowledge representation model for advanced digital forensics timeline analysis. *Digital Investigation* 11 (2014), S95–S105.

[12] LiWu Chang and Ira S Moskowitz. 2002. An integrated framework for database privacy protection. In *Data and Application Security*. Springer, 161–172.

[13] Gauthier Chassang. 2017. The impact of the EU general data protection regulation on scientific research. *ecancermedicalscience* 11 (2017).

[14] Keke Chen, Gordon Sun, and Ling Liu. 2007. Towards attack-resilient geometric data perturbation. In *proceedings of the 2007 SIAM international conference on Data mining*. SIAM, 78–89.

[15] Lorrie Cranor, Tal Rabin, Vitaly Shmatikov, Salil Vadhan, and Daniel Weitzner. 2016. Towards a privacy research roadmap for the computing community. *arXiv preprint arXiv:1604.03160* (2016).

[16] Dan S Decasper, Zubin Dittia, Prashanth Mundkur, and Rajib Ghosh. 2009. Advanced content and data distribution techniques. US Patent 7,555,532.

[17] Alexandre Evfimievski. 2002. Randomization in privacy preserving data mining. *ACM Sigkdd Explorations Newsletter* 4, 2 (2002), 43–48.

[18] Alexandre Evfimievski, Ramakrishnan Srikant, Rakesh Agrawal, and Johannes Gehrke. 2004. Privacy preserving mining of association rules. *Information Systems* 29, 4 (2004), 343–364.

[19] Ruth Gavison. 1980. Privacy and the Limits of Law. *The Yale Law Journal* 89, 3 (1980), 421–471.

[20] Ahmed HajYasien. 2007. *Preserving privacy in association rule mining*. Ph.D. Dissertation. Ph. D Thesis, University of Griffith.

[21] Lynne M Healy. 2008. Exploring the history of social work as a human rights profession. *International social work* 51, 6 (2008), 735–748.

[22] Katsuhiro Honda, Arina Kawano, Akira Notsu, and Hidetomo Ichihashi. 2012. A fuzzy variant of k-member clustering for collaborative filtering with data anonymization. In *2012 IEEE International Conference on Fuzzy Systems*. IEEE, 1–6.

[23] Hossein Hosseini, Baicen Xiao, Andrew Clark, and Radha Poovendran. 2017. Attacking automatic video analysis algorithms: A case study of google cloud video intelligence api. In *Proceedings of the 2017 on Multimedia Privacy and Security*. ACM, 21–32.

[24] Adnan Rashid Hussain, Mohd Abdul Hameed, and Nagaratna P Hegde. 2011. Mining twitter using cloud computing. In *2011 World Congress on Information and Communication Technologies*. IEEE, 187–190.

[25] Murat Kantarcioglu and Chris Clifton. 2004. Privacy-preserving distributed mining of association rules on horizontally partitioned data. *IEEE Transactions on Knowledge & Data Engineering* 9 (2004), 1026–1037.

[26] MohammadReza Keyvanpour and Somayyeh Seifi Moradi. 2011. Classification and evaluation the privacy preserving data mining techniques by using a data modification-based framework. *arXiv preprint arXiv:1105.1945* (2011).

[27] Yiannis Kokkinos and Konstantinos Margaritis. 2013. Distributed privacy-preserving P2P data mining via probabilistic neural network committee machines. In *IISA 2013*. IEEE, 1–4.

[28] Patrick Lange and David Suendermann-Oeft. 2014. Tuning Sphinx to outperform GoogleâĂŹs speech recognition API. In *Proc. of the ESSV 2014, Conference on Electronic Speech Signal Processing*. 1–10.

[29] Marc Langheinrich. 2016. Privacy in Ubiquitous Computing. In *Ubiquitous computing fundamentals*. Chapman and Hall/CRC, 109–174.

[30] Nhien-An Le-Khac, Lamine M Aouad, and M-Tahar Kechadi. 2007. Knowledge Map: Toward a new approach supporting the knowledge management in Distributed Data Mining. In *Third International Conference on Autonomic and Autonomous Systems (ICAS'07)*. IEEE, 67–67.

[31] Nhien An Le Khac and M-Tahar Kechadi. 2010. Application of data mining for anti-money laundering detection: A case study. In *2010 IEEE International Conference on Data Mining Workshops*. IEEE, 577–584.

[32] Daniel D Lee and H Sebastian Seung. 2001. Algorithms for non-negative matrix factorization. In *Advances in neural information processing systems*. 556–562.

[33] Kristen LeFevre, David J DeWitt, and Raghu Ramakrishnan. 2005. Incognito: Efficient full-domain k-anonymity. In *Proceedings of the 2005 ACM SIGMOD international conference on Management of data*. ACM, 49–60.

[34] Ninghui Li, Tiancheng Li, and Suresh Venkatasubramanian. 2007. t-closeness: Privacy beyond k-anonymity and l-diversity. In *2007 IEEE 23rd International Conference on Data Engineering*. IEEE, 106–115.

[35] Yaping Li, Minghua Chen, Qiwei Li, and Wei Zhang. 2012. Enabling multilevel trust in privacy preserving data mining. *IEEE Transactions on Knowledge and Data Engineering* 24, 9 (2012), 1598–1612.

[36] Yehuda Lindell and Benny Pinkas. 2000. Privacy preserving data mining. In *Annual International Cryptology Conference*. Springer, 36–54.

[37] Kun Liu, Hillol Kargupta, and Jessica Ryan. 2006. Random projection-based multiplicative data perturbation for privacy preserving distributed data mining. *IEEE Transactions on knowledge and Data Engineering* 18, 1 (2006), 92–106.

[38] Ashwin Machanavajjhala, Johannes Gehrke, Daniel Kifer, and Muthuramakrishnan Venkitasubramaniam. 2006. l-diversity: Privacy beyond k-anonymity. In *22nd International Conference on Data Engineering (ICDE'06)*. IEEE, 24–24.

[39] Majid Bashir Malik, M Asger Ghazi, and Rashid Ali. 2012. Privacy preserving data mining techniques: current scenario and future prospects. In *2012 Third International Conference on Computer and Communication Technology*. IEEE, 26–32.

[40] Karen Mc Cullagh. 2008. What is' private'data?. In *23rd BILETA Conference. Glasgow Caledonian University*.

[41] Davide Mulfari, Antonio Celesti, Maria Fazio, Massimo Villari, and Antonio Puliafito. 2016. Using Google Cloud Vision in assistive technology scenarios. In *2016 IEEE Symposium on Computers and Communication (ISCC)*. IEEE, 214–219.

[42] Gayatri Nayak and Swagatika Devi. 2011. A survey on privacy preserving data mining: approaches and techniques. *International Journal of Engineering Science and Technology* 3, 3 (2011).

[43] Stanley RM Oliveira and Osmar R Zaiane. 2002. Privacy preserving frequent itemset mining. In *Proceedings of the IEEE international conference on Privacy, security and data mining-Volume 14*. Australian Computer Society, Inc., 43–54.

[44] Annie Pearl, Andy Kiang, and Joel Bailon. 2016. Data loss prevention (DLP) methods by a cloud service including third party integration architectures. US Patent 9,473,532.

[45] Xinjun Qi and Mingkui Zong. 2012. An overview of privacy preserving data mining. *Procedia Environmental Sciences* 12 (2012), 1341–1347.

[46] Ronica Raj and Veena Kulkarni. 2015. A Study on Privacy Preserving Data Mining: Techniques, Challenges and Future Prospects. *International Journal of Innovative Research in Computer and Communication Engineering* 3, 11 (2015).

[47] Shariq J Rizvi and Jayant R Haritsa. 2002. Maintaining data privacy in association rule mining. In *Proceedings of the 28th international conference on Very Large Data Bases*. VLDB Endowment, 682–693.

[48] Pierangela Samarati and Latanya Sweeney. 1998. *Protecting privacy when disclosing information: k-anonymity and its enforcement through generalization and suppression.* Technical Report. technical report, SRI International.

[49] Saeed Samet and Ali Miri. 2009. Privacy-preserving bayesian network for horizontally partitioned data. In *2009 International Conference on Computational Science and Engineering*, Vol. 3. IEEE, 9–16.

[50] Yücel Saygin, Vassilios S Verykios, and Ahmed K Elmagarmid. 2002. Privacy preserving association rule mining. In *Proceedings Twelfth International Workshop on Research Issues in Data Engineering: Engineering E-Commerce/E-Business Systems RIDE-2EC 2002*. IEEE, 151–158.

[51] Ferdinand David Schoeman. 1984. *Philosophical dimensions of privacy: An anthology.* Cambridge University Press.

[52] Belle Selene Xia and Peng Gong. 2014. Review of business intelligence through data analysis. *Benchmarking: An International Journal* 21, 2

(2014), 300–311.

[53] Alpa Shah and Ravi Gulati. 2016. Privacy preserving data mining: Techniques classification and implicationsâĂŤA survey. *Int. J. Comput. Appl* 137, 12 (2016), 40–46.

[54] Traian Marius Truta and Bindu Vinay. 2006. Privacy protection: p-sensitive k-anonymity property. In *22nd International Conference on Data Engineering Workshops (ICDEW'06)*. IEEE, 94–94.

[55] S Tsiafoulis, Vasilis C Zorkadis, and Dimitris A Karras. 2010. A Neural-Network Clustering-Based Algorithm for Privacy Preserving Data Mining. In *Grid and Distributed Computing, Control and Automation*. Springer, 269–276.

[56] Hina Vaghashia and Amit Ganatra. 2015. A survey: privacy preservation techniques in data mining. *International Journal of Computer Applications* 119, 4 (2015).

[57] Jaideep Vaidya and Chris Clifton. 2002. Privacy preserving association rule mining in vertically partitioned data. In *Proceedings of the eighth ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM, 639–644.

[58] Vassilios S Verykios, Elisa Bertino, Igor Nai Fovino, Loredana Parasiliti Provenza, Yucel Saygin, and Yannis Theodoridis. 2004. State-of-the-art in privacy preserving data mining. *ACM Sigmod Record* 33, 1 (2004), 50–57.

[59] Michael E Wall, Andreas Rechtsteiner, and Luis M Rocha. 2003. Singular value decomposition and principal component analysis. In *A practical approach to microarray data analysis*. Springer, 91–109.

[60] Gregory J Walters. 2002. *Human rights in an information age: A philosophical analysis*. University of Toronto Press.

[61] Alan F Westin. 1968. Privacy and freedom. *Washington and Lee Law Review* 25, 1 (1968), 166.

[62] Raymond Chi-Wing Wong, Jiuyong Li, Ada Wai-Chee Fu, and Ke Wang. 2006. (α, k)-anonymity: an enhanced k-anonymity model for privacy preserving data publishing. In *Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM, 754–759.

[63] Xiaokui Xiao and Yufei Tao. 2007. M-invariance: towards privacy preserving re-publication of dynamic datasets. In *Proceedings of the 2007 ACM SIGMOD international conference on Management of data*. ACM, 689–700.

[64] Zhiqiang Yang, Sheng Zhong, and Rebecca N Wright. 2005. Privacy-preserving classification of customer data without loss of accuracy. In *Proceedings of the 2005 SIAM International Conference on Data Mining*. SIAM, 92–102.

[65] Shui Yu. 2016. Big privacy: Challenges and opportunities of privacy study in the age of big data. *IEEE access* 4 (2016), 2751–2763.