Evaluation and Criteria of Privacy Preserving Data Mining Techniques

Muhammad Asghar, Mujtaba Haider, Muhammad Ahsan, Majid Ali
Department of Information Technology/Engineering, National University of Modern Languages (NUML)
Islamabad, Pakistan
Muhammadasghar546@yahoo.com

Abstract__ Big Data refers to the massive amount of data which is impossible to manage and use efficiently using traditional tools and techniques. Big Data is currently one of the fastest growing field in the computing world. It is the large amount of digital information collected by governments and organizations about human beings and our environment. Nobody whether in academia or in the industry deny the huge potential of benefits that can be gained through using Big Data. In order to extract the potential benefits of big data it is necessary to mine the useful information while preserving the individual’s privacy. Hence implementation of Privacy preserving data mining (PPDM) has become the most important need of the time. In this thesis we conduct a systematic literature review (SLR) to identify the privacy issues, their solutions and limitation related to privacy preserving techniques in big data environment that have been proposed by the peers. After obtaining the results from systematic literature review (SLR) and performing the analysis we find it necessary to move towards standardization in privacy preserving data mining (PPDM) techniques. We propose a privacy preserving metric/criteria on which we can evaluate the different privacy preserving data mining techniques and give comparative study of various privacy preserving techniques.

Keywords—Physical Solution; Logical Solution; Decomposition of Function; Thread Analysis;

1) BIG DATA

Big Data refers to the massive amount of data which is impossible to manage and use efficiently using traditional tools and techniques [1, 2, 3]. It is the large amount of digital information collected by governments and organizations about human beings and our environment. Big Data comprises of a detailed description of a large and complicated set of characteristics, practices, techniques and ethical issues that are associated with data [5]. It consists of any voluminous amount of structured, semi-structured and unstructured data that has the potential to be mined for information [1]. Data have become stream flowing in every area of the global economy. [9] Every day we create 2.5 quintillion bytes of data so much that 90 % of the data in the world has been created in the last two years alone. [8] Gartner predicts that enterprise data will grow by 800 times from 2011 to 2015, with 80 percent unstructured (for example, e-mails, documents, video, images, and social media content) and 20 percent structured (for example, credit card transactions and contact information) [4]. It can be seen that Internet of things (IoT) applications will increase the amount of data to an unknown level. The 3Vs that define Big Data are Variety, Velocity and Volume, [1, 5, 6, 7, 4, 3]. Companies generate tremendous volume of transactional data capturing trillions of bytes of information about their clients, suppliers and operations. [10] Millions Of network sensors are present in physical world in devices such as cell phones, smart meters, vehicles and industrial machinery that sense, capture and transmit (communicate) data in the present time of Internet of Things. [10] When Companies and Corporations do their business and interact with people, they are producing a huge amount of information “exhaust data” i.e., data that is generated in the form of by-product of other activities. Social sites (Facebook, Twitter, WhatsApp), Smart phones and various other personal devices like PC’s and laptops have allowed billions of peoples around the globe to contribute in the expansion of Big Data. [11] In present digital world each person also creates its own enormous trails of Big Data while communicating, browsing, buying, sharing and searching etc. [11] As the digital world keeps on growing with every passing moment, so it is the Big Data. The ability to store, manipulate and combine data and then use the outcome or finding for the purpose of analysis has now become possible more than ever as describe by Moore’s law in computing. [12]

A. Stages involved in big data

There are six main stages involved in bid data as discuss below in detail.

1) Data Acquisition

The first step in Big Data is acquiring the data itself. With the growing medium the rate of data generation is rising exponentially. With the introduction of smart devices which are used with a wide array of sensors continuously generate data. The Large Hadrons Collider in Switzerland produces petabytes of data. Most of this data is not useful and can be discarded, however due to its unstructured form; selectively discarding the data presents a challenge. This data becomes more potent in nature when it’s merged with other valuable
data and superimposed. Due to the interconnectedness of devices over the World Wide Web, data is increasingly being collated and stored in the cloud.

2) Data Extraction
All of the data generated and acquired is not of use. It contains a large amount of redundant or unimportant data. For instance, a simple CCTV camera, constantly polls sensor to gather information of the user’s movements. However, when the user is in a state of inactivity, the data generated by the activity sensor is redundant and of no use. The challenges presented in data extraction are twofold: firstly, due to nature of data generated, deciding which data to keep and which to discard increasingly depends on the context in which the however it is important not to discard similar data in a case where it is being generated by a heart-rate sensor.

3) Data Collection
Data from a singular source often is not enough for analysis or prediction. More than one data sources are often combined to give a bigger picture to analyze. For example a health monitor application often collects data from the heart-rate sensor, pedometer, etc. to summarize the health information of the user.

4) Data Structuring
Once all the data is aggregated, it is very important to present and store data for further use in a structured format. The structuring is important so queries can be made on the data. Data structuring employs methods of organizing the data in a particular schema. Various new platforms, such as NoSQL, can query even on unstructured data and are being increasingly used for Big Data Analysis. A major issue with big data is providing real time results and therefore structuring of aggregated data needs to be done at a rapid pace.

5) Data Visualization
Once the data is structured, queries are made on the data and the data is presented in a visual format. Data Analysis involves targeting areas of interest and providing results based on the data that has been structured. For instance, data containing average temperatures are shown alongside water consumption rates to calculate a relation in between them. This analysis and presentation of data makes it ready for consumption for users. Raw data cannot be used to gain insights or for judging patterns, therefore “humanizing” the data becomes all the more important.

6) Data Interpretation
The ultimate step in Big Data processing includes interpretation and gaining valuable information from the data that is processed. The information gained can be of two types: Retrospective Analysis includes gaining insights about events and actions that have already taken place. For instance, data about the television viewership for a show in different areas can help us judge the popularity of the show in those areas. Prospective Analysis includes judging patterns and discerning trends for future from data that is already been generated. Weather Prediction using big data analysis is an example of prospective analysis.

B. Today’s Threat Environment
Today threat environment emerges from the three Vs of the Big Data. [13] Each aspect must be understood carefully in order to manage the threats.

1) Volume
As the volume of Big Data is growing so is the threat environment. In the 1990’s, on average every PC user received one or two spam messages daily.[13] But as of August 2010 up to 200 billion spam messages are being sent per day. [13] Similar trend is being followed in the file transfer and web page retrieval. [13]

2) Variety
The temptation of financial benefits has encouraged cybercriminal to come up with more innovative and lethal methods in order to avoid detection. [13] For example, malware created today goes through several quality control procedures. [13] It is tested on various machine and platforms. [13] Similarly, server-side polymorphic threats that evolve and spread rapidly are very difficult to detect through traditional methods. [13] Moreover, distribution points for malware, viruses, spam and other malicious tools used by cybercriminals are increasing rapidly. [13]

3) Velocity
The large volume and variety of threats must be managed and maintained on regular basis due to which velocity becomes a growing challenge. [13] Due to the rapidness of the internet now threats do not remain localize to any one location and expand drastically hence, threats have become very complex. [13] Unlike a physical address, changing IP address on internet is very trivial, swift and difficult to trace. [13] Determining whether a specific website contains malicious content is fluid over time as well. [13] Cybercriminal routinely change sites into corrupt sites instantly. [13]

C. Security Goals
Having defined the adversary we want to protect against, we need to describe the security goals. The three most fundamental security goals are confidentiality, integrity, and availability collectively known as the CIA triad.

1) Confidentiality
Confidentiality is the goal of keeping all sensitive data secret from an adversary. More formally, traditional definitions of confidentiality guarantee that an adversary should learn no information about the sensitive data, other than its length. Confidentiality is critical in big data applications to guarantee that sensitive data is not revealed to the wrong parties.

2) Integrity
Integrity is the goal that any unauthorized modification of data should be detectable. That is, a malicious adversary should not be able to modify such data without leaving a trace. This is
very important to help guarantee the veracity of data collected in big data applications.

3) Availability

Availability is the goal of always being able to access one’s data and computing resources. In particular, an adversary should not be able to disable access to critical data or resources. This is a very important security goal in big data processing, as the sheer volume and velocity of the data make guaranteeing constant access a difficult task. However, in today's big data systems, availability is typically guaranteed via non-cryptographic means such as replication.

D. Big Data with data mining

Mostly Big data states to a group of huge volumes of data and these data are produced from numerous sources such as internet, social media, business organizations etc. From these data some useful information can be mined with the help of data mining. Data mining is a method for determining interesting patterns as well as expressive, understandable models from vast scale data [31].

E. HACE Theorem

Big Data starts with large-volume, heterogeneous, autonomous sources with distributed and decentralized control, and seeks to explore complex and evolving relationships among data [33,34]. These characteristics make it an extreme challenge for discovering useful information from big data. In connection with this scenario, let us imagine a scenario where blind people are asked to draw the picture of an elephant. The information collected by each blind person will be such that they may think the trunk as a “wall”, leg as a “tree”, body as a “wall” and tail as a “rope”. In this case one blind man can exchange information with other which may be biased.

F. Data mining tools

Data mining tools can be categorized into following three different types.

1) Traditional Data Mining Tools

These type of tools helps companies to construct data patterns and structures by using different algorithms and methods normally installed on an isolated computer. Windows and UNIX versions are available in majority, they usually control data using offline tools [30].

2) Dashboards

Reflect data changes enabling the user to see how the business is performing.

3) Text mining tools

Text mining tools mines data from different kinds of texts, example, Microsoft word, acrobat PDF, and text files. These tools scan the content and convert the selected data into format that is compatible with tool’s database [30].

2) RELATED WORK

Privacy preserving data mining is one of the main property that an information system must fulfill. For this purpose, many researches have been conducted to incorporating privacy preserving techniques with data mining algorithms in order to stop the disclosure of sensitive information during the knowledge discovery. The present privacy preserving data mining (PPDM) techniques can be classified according to the different dimensions by Verykios et al. (2004) that include Data distribution (centralized or distributed), the modification applied to the data (encryption, perturbation, generalization, and so on) in order to sanitize them, the data mining algorithm which the privacy preservation technique is designed for, the data type (single data items or complex data correlations) that needs to be protected from disclosure and the approach adopted for preserving privacy (heuristic or cryptography-based approaches) [50]. There are different privacy preserving data mining (PPDM) techniques that have been developed in the recent years but there is a strong need of moving toward standardization in this new research area, as discussed by Oliveira and Zaiane [51]. One step toward this essential process is to provide a quantification approach for PPDM algorithms to make it possible for the evaluation and comparison of such algorithms.[52] However, due to the variety of features of PPDM algorithms, it is often the case that no privacy preserving algorithm exists that outperforms all the others on all possible criteria. [52]

To evaluate the privacy preserving data mining algorithms (PPDM), many researchers have proposed different metrics related to privacy preserving and to set the criteria that measures the effectiveness of algorithms from one to another. Elisa Bertino et al. identified a criteria based on which a PPDM algorithm can be evaluated that was based on Privacy level, Hiding failure, Data quality and Complexity. [52,55] In August 2005 ELISA BERTINO and IGOR NAI FOVINO identified the set of criteria based on which a PPDM algorithm can be evaluated as efficiency, scalability, data quality, hiding failure and privacy level [54].

3) PRIVACY PRESERVING DATA MINING

Privacy Preserving Data Mining is one major issue in recent years due to the advancement in the big data architecture e.g. storage, Processing, and increase in the optimization of the data mining techniques. Various techniques such as Data Modification (Association Rules, Randomized Noise, Correlated Noise, Randomized Response), secure Multi-party computation (SMC). Later on these techniques are discussed in detail.

A. Requirements of Privacy Preserving Data Mining

There are many privacy preserving data mining requirements and some of them are given below.

1) Accuracy

The accuracy is closely related to the information loss resulting from the hiding strategy. The less is the information loss, the better is the data quality. Always a PPDM techniques has to maintain high accuracy to reduce information loss (Aggarwal et al. 2008).
2) **Completeness and Reliability**
Completeness evaluates the degree of missed data in the sanitized database. Incomplete data has a significant impact on data mining results and impairs the data mining techniques from providing an accurate representation of the underlying data.
Reliability is related to the semantic constraints holding on the data and it measures how many of these constraints are still satisfied after the sanitization (Kantarcioglu et. al 2007).

3) **Scalable**
It is another important aspect to assess the performance of a PPDM technique. In particular, scalability describes the efficiency trends when data sizes increase. Such parameter concerns the increase of both performance and storage requirements as well as the costs of the communications required by a data mining technique with the increase of data size (Bertino et. al 2008).

4) **Data quality**
It is an important aspect of PPDM. High quality data that has been prepared specifically for data mining tasks will result in useful data mining models and output. Alternatively, low quality data has a significant negative impact on the utility of data mining results (Bettino et. Al 2009).

5) **Security**
It is the degree of protection against danger, damage, loss, and crime. There are two main approaches regarding how to deal with the problems of privacy that arise today. The first is a legal and policy approach whereby organizations are limited in how they store and use data based on privacy law and public policy. It typically works by evaluating scenarios and deciding if the privacy breach caused by using the data in a given way is justified or not. The second approach is technological, and provides enforced privacy guarantees through cryptographic means. This approach has the capability of enabling the data to be used while preventing privacy breaches (Nan Zhang et. al 2009).

B. **Classification Scheme**
Each privacy preserving data mining (PPDM) scheme is developed and suitable for a specific scenario. Techniques of privacy preserving data mining classified according to the given characteristics.

1) **5.2.1 Data Mining Scenario**
Two major data mining scenarios are
1) Organization allow unrestricted access to data for data mining and data modification technique is used in such scenario.
2) Data mining of data is allowed without releasing the data set and cryptographic techniques are used in such conditions.

2) **Data Mining Task**
Data mining tasks like evaluation analysis, clustering, association rule mining, outlier analysis and classification are used to take out different type of patterns in data set [6]. Privacy preserving techniques are categorized base on above mentioned tasks.

3) **Data Distribution**
For data mining used data can be
1) Distributed: Data shared between two or more organization and they cannot trust each other but want to perform joint data mining. Distributed data can be partitioned vertically or horizontally.
2) Centralized: A single organization owns the data which either can be available at the computation or site.
Distribution based classification is given below in figure 1.

![Distribution based data classification](image)

4) **Data Types**
Two major types of data are
1) Numerical Data
2) Categorical Data: Special case of Categorical Data is Boolean data.

5) **Privacy Definition**
The privacy definition vary in scenario to scenario. In some context data values of individuals are private and similarly in other classification rules are data associations are private.

C. **Evaluation Criteria**
Evaluation criteria for privacy preserving data mining techniques is given as

1) **Flexibility**
Ability of the data mining techniques to address data mining tasks, different types of data and different privacy requirements. Flexibility consists upon:
- Private data or sensitive patterns.
- Distributed data or centralized data.
- Data mining tasks.

2) **Hiding Failure**
Disclosure risks means sensitive information can be extracted from data by an illegal data miner. Level of privacy and hiding failure are inversely proportional to each other. Minimizing of hiding failure is a major objective of privacy preserving data mining techniques so its evaluation is really essential.
3) Cost
Two major types of cost are
- Computation cost
- Communication cost

If two or more organizations are performing joint data mining then communication cost is the most important factor for evaluation. Computation cost and communication cost of data mining are inversely proportional to the efficiency of data mining techniques.

4) Information Loss
Information loss increases as the added noise increases or the level of privacy get higher. To maintain the data quality is an important factor for data mining techniques. If technique cannot maintain data quality the privacy of data mining is useless.

5) Data type
It refers different types of data including numerical, binary or categorical.

D. Comparative Study of Privacy preserving data mining techniques

In current scenarios, the most important issue is privacy of data. Mostly distributed systems contain the most important and valuable information. In mining, information found can be confidential or may be used by unauthorized person in bad manners. For centralized database current privacy preserving techniques can be classified into three main groups based on their usage such as:

(i) query restriction
(ii) output perturbation
(iii) data modification.[7]

1) Data Modification
Data modification is a straightforward technique to implement among above mentioned techniques. In data modification, modifies the data set before release of a data set for many data mining tasks and analysis but the quality of released data remains high. After modification of data we can use any off the shelf software such as See5, multimedia, spread sheet software to manage or analyze the data it’s not with the case of query restriction and output perturbation. Due to simplicity of this technique it used mostly in the case of statistical database in data mining. There are many number of ways of performing modification such as suppression, swapping, aggregation and noise addition. The basic idea of these techniques is given below.

2) Data Swapping

Dalenius and Reiss first introduced data swapping technique in 1982, for categorical values modification in the context of secure statistical database. According to their work this technique can be used to both produce micro data and release statistical tabulations so that confidentiality is not violated [8].

The main concept of the technique was it keeps all original value in the data set, while at the same time makes the record re-identification very complex. An overview to existing data swapping technique can be found in [9][10].

3) Aggregation
Generalization or global recording are the other names of aggregation explore preserving the anonymity by the use of generalizations and suppressions on the potentially identifying portions of the data. It is use to protect individual privacy in a released data set before its releasing by perturbing the original data set. Aggregation change k no. of records of a data by representative records. Another method of aggregation or generalization is transformation of attribute values. For ex- an exact birth date can be changed by the year of birth. Such a generalization makes an attribute value less informatics. For ex- if exact birth date is changed by the century of birth then the released data can became useless to data miners [11].

4) Suppression
In this technique sensitive data value are deleted or suppressed prior to the release of a micro data. Suppression is used to protect an individual privacy from intruders attempt to accurately predict a suppressed value. An important issue in suppression is to minimize the information loss by minimizing the number of values suppressed. Suppression technique is also been used for association and classification rule confusion [12], [13].

5) Noise Addition in Data Mining
Noise addition is usually adds a random number (noise) to numerical attributes. This random number is generally drawn from a normal distribution with zero mean or standard deviation. Noise is added in a controlled way so as to maintain variance, co-variance and means of the attributes of a data set. Due to the absence of natural ordering in categorical values, addition of noise in categorical attributes is not straightforward as like addition of noise in numerical attributes. Many techniques proposed for noise addition in data mining. Evfimievski et al. proposed a novel noise addition technique for privacy preserving association rule mining in 2002 [14]. Agarwal and Srikant proposed a noise addition technique in 2000 which is based on addition of random noise to attribute values in such a way that the distributions of data values belonging to original and perturbed data set were very difficult [5]. Du and Zhan presented a decision tree building algorithm which is used to perturb multiple attributes [15]. In 2004 Zhu and Liu [16] proposed a general framework for randomization using a well-studied statistical model called mixture model. According to this scheme data are generated from a distribution that depends on some factors including original data as well. Their randomization framework supports privacy preserving density estimation.

i. Association Rule Mining
Association rule mining technique is developed to find out the combinations of products or services which a buyer gets from seller and the relationship among these provided products or services. These relationships among products are services are called association rules. And at industrial level it is highly important to find out these association rules to increase the business or to minimize loss risks. [1][2]

Association rules mining consist upon two major steps
1) Find frequent Items:

Asghar, Mujtaba, Ahsan, Majid|Big Data 48
In this step it is determined that a specific item from item set is transacted how many times or what is the frequency of transaction of a specific item.

2) Generate strong association rule:
In this step of association rule the relationship among transacted items is generated. Ex if an item a transacted then how many times item B is also transacted with item A. [1][3]

ii. Random Noise Addition Technique
Random noise addition technique is developed based on the challenge that sometimes data mining need distribution of data. The basic idea of this technique is given as below. Assume X1, X2, X3, ................., Xn are the original variables of distributed data set. The function to distribute the data set is represented as FX and we get variables W1, W2, W3, ................., Wn after addition of noise variables Y1, Y2, y3, ................., Yn. The noise and original data are added as Wi = Xi + Yi

In this technique we can observe that a large level of noise added and it can causes information loss in data for data mining and the other limitation is that the noise can be abstracted using spectral techniques on randomized data. Thus to manage balance between data mining accuracy and information loss is difficult. To mitigate information loss limitation a new technique called decision tree noise developed. [5][4][6][7]

iii. Correlated Noise
In 1986 Kim suggested a new technique for noise addition called correlated noise. In correlated noise technique if the data is X11, X22, X33 . . . . Xij. In this statement X is the original data with ith record and jth attribute. We add noise at each record and attribute level. Example,

\[ Yij = Xij + Eij \] \hspace{1cm} (1)

In equation (1) the yij is noised data, and Eij is the added noise. In this technique the information loss is reduced in numerical data.

### Table 1.

<table>
<thead>
<tr>
<th>Name of Techniques</th>
<th>Flexibility</th>
<th>Data mining tasks</th>
<th>Hiding failure</th>
<th>Cost</th>
<th>Information Loss</th>
<th>Data type</th>
</tr>
</thead>
<tbody>
<tr>
<td>Association Rules</td>
<td>Private data/sensitive pattern</td>
<td>Distributed/centralized data.</td>
<td>Association Rules</td>
<td>Very Low</td>
<td>Moderate</td>
<td>None</td>
</tr>
<tr>
<td>Randomized Noise</td>
<td>Data</td>
<td>Both</td>
<td>Classification</td>
<td>Low</td>
<td>Low</td>
<td>Moderate</td>
</tr>
<tr>
<td>Correlated Noise</td>
<td>Data</td>
<td>Centralized</td>
<td>Clustering, Classification</td>
<td>Moderate</td>
<td>Low</td>
<td>Low</td>
</tr>
<tr>
<td>Secure Multiparty Computation (SMC)</td>
<td>Data</td>
<td>Distributed (Vertical)</td>
<td>Association Rules</td>
<td>Very Low</td>
<td>Moderate</td>
<td>None</td>
</tr>
<tr>
<td>Randomized Response</td>
<td>Data</td>
<td>Both</td>
<td>Density estimation</td>
<td>Low</td>
<td>Low</td>
<td>Moderate</td>
</tr>
</tbody>
</table>

iv. Randomized Responses Technique
In this technique data is scrambled so that while data mining any person could not extract the hidden information received by source. This technique consists upon two major models. Related question and unrelated question. In related question interviewer asks questions related to each other from every respondent and get the opposite reply from each other. Ex 1) He has gone
2) He has not gone
In this case the interviewer receives random answer is yes and no form and he don’t know for which question the answer is received. And in this way the hiding failure risk reduced. [8][5]

6) Secure Multiparty Computation (SMC)
A Secure Multi-party Computation (SMC) technique encrypts the data sets, while still allowing data mining operations. SMC techniques are not supposed to disclose any new information other than the final result of the computation to a participating party. These techniques are typically based on cryptographic protocols and are applied to distribute data sets. Parties involved in a distributed data mining encrypt their data and send to others parties. These encrypted data are used to compute the aggregate data, belonging to the joint data set, which is used for data mining purpose. Secure Multiparty Computation was originally introduced by Yao in 1982 [17]. Basically, SMC is supposed to reveal to a party just the result of the computation and the data owned by the party. There are various SMC algorithms developed. Most of the algorithms make use of some primitive computations such as secure sum, secure set union, secure size of set intersection and secure scalar product.

E. Comparative Study Results
Comparative study of some privacy preserving data mining techniques performed here and these techniques are evaluated based on the criteria discussed above. The results of comparative study are given below in table 12.
4) CONCLUSION
In this paper, different techniques of privacy preserving data mining are explained. We have proposed an evaluation criteria based on which we have evaluated data modification and secure multi-party computation techniques. Through it we can identify which privacy preserving technique is suitable in any given environment. In future, many other privacy metrics or attributes can be added in order to further optimize this criteria. We may also use it in evaluating other privacy preserving data mining techniques. One major future work is to examine this proposed work in the industry so that its effectiveness and usability can be measured.

5) REFERENCES
[17] WORKSHOP REPORT “BIG DATA SECURITY AND PRIVACY” Sponsored by the National Science Foundation September 16-17.
[26] Salini S. “ Survey on Data Privacy in Big Data with K- Anonymity”
[34] Xindong Wu, “Data Mining with Big Data”,
[38] Charu C. Aggarwal, “A General Survey of Privacy-Preserving Data Mining Models and Algorithms”
[52] Elisa Bertino, Dan Lin, and Wei Jiang : “A Survey of Quantification of Privacy Preserving Data Mining Algorithms”
[65] Rongxing Lu, Hui Zhu, Xinmeng Liu, Joseph K. Liu, and Jun Shao”Toward Efficient and Privacy-Preserving Computing in Big Data”