International Journal of Science and Research (IJSR)
ISSN (Online): 2319-7064
Index Copernicus Value (2015): 78.96 | Impact Factor (2015): 6.391

Survey of Privacy Preserving Techniques and Upcoming Techniques: A Review

Payal P. Panse[1], P. L. Paikrao[2]

[1]Student, Department of Electronics Engineering, Government College of Engineering, Amravati, India

[2]Assistant Professor, Department of Electronics Engineering, Government College of Engineering, Amravati, India

Abstract: ***In recent years the data-leak instances are growing rapidly. The confidential and private data in many fields should be protected. Many sectors as government organizations, business fields, defence, medical field and educational statistics have sensitive data which should not be leaked and kept confidential. But it is studied from some reports that in last few years the data leak incidents are doubledand thus there is need to have some more protective technique to secure the confidential data. The existing solutions require some more privacy techniques so as to increase the security of the data. Thus a new technique called a Fuzzy-fingerprinting technique is described. It is a privacy preserving data-leak detection solution to solve the privacy issue. Here a special set of sensitive data digests is used. For the confidential data communication its plaintext data must be encrypted, the data is encrypted using fuzzy-fingerprints and corresponding fuzzy-fingerprinting algorithm. In the data communication the data is usually provided to the server or data-leak detection provider (DLD) provider which is of semi-honest nature and where leak can be found out. The described method keeps the sensitive data exposure to the minimum leveland thus it helps to reduce the data-leak instances.***

Keywords: Data leak, network, security, randomization, perturbation, anonymization, cryptography

1. Introduction

1.1 Introduction

According to the survey in research institutions and government organizations showed that the number of data-leak instances has grown rapidly in recent years[13]. Privacy Preserving of sensitive data leakage has become the most important issue in today's world. Many government organizations, research institutions tell that number of data leakage instances have growing rapidly in today's world[14].The most data-leak instances are due to deliberately planned attacks, inadvertent leaks (e.g. forwarding confidential emails to unclassified email accounts), and human mistakes (e.g., assigning the wrong privilege). Data leakage is one of the most concerned security issues that sensitive data has been disclosed to unauthorized entities intentionally or unintentionally. It has become a serious issue to organizations, because a single incident can result in losing customers' loyalty, unexpected lawsuit risks, extra cost of remediation, etc. This issue is getting more serious with the rapid proliferation of mobile devices, widespread use of removable devices, and ubiquitous Internet access. Therefore, many data loss prevention (DLP) systems have been developed to discover, monitor and protect data by deep content inspection. The existing solutions or methods to encrypt the sensitive data before transmission are watermarking and other cryptographic models. The traditionally used techniques have some problems so a new technique called fuzzy-fingerprintingtechnique [1] came into existence. It is an advanced version of Rabin fingerprinting model where the data can be secured and data-leak can be found out by minimal exposure of data to the DLD provider and thus, is a convenient method to find data-leak.

1.2 Need

Symantec reported that more than 232.4 million identities were exposed in 2011. Verizon's data breach investigation report showed that 174 million data records were compromised in a total of 855 data breach incidents in 2011. According to a report from Risk Based Security (RBS)[13], the number of leaked sensitive data records had increased dramatically during the last few years, i.e., from 412 million in 2012 to 822 million in 2013. According to Data Loss DB's statistics, global data leakage incidents in 2012 was 1,529 which was much higher than before. Therefore, many data loss prevention (DLP) systems have been developed to discover, monitor, and protect data by deep content inspection. The most data-leak instances are because of deliberately planned attacks, inadvertent leaks (e.g. forwarding confidential emails to unclassified email accounts), and human mistakes (e.g., assigning the wrong privilege).Due to these data-leak instances the confidential data of the government organizations, defence, etc. can be misused. Thus the data must be secured and stored confidentially and if there is data-leak it can be found out using some cryptographic models. The existing techniques require more complex mathematical operations, are not much reliable and degrade the original quality of the data. So a new technique of fuzzy-fingerprinting came into existence [14][1].

1.3Brief Introduction to Data Mining

As people in each and every field are using internet for various purposes there is growing evidence of proliferation of sensitive information. Security and privacy of data became an important concern. So there is need to learn and study various data security techniques. Privacy preserving data mining techniques are introduced with the aim to extract the relevant knowledge from the large amount of

data while protecting the sensible information simultaneously.

Figure 1 shows the framework of the privacy preserving data mining approach

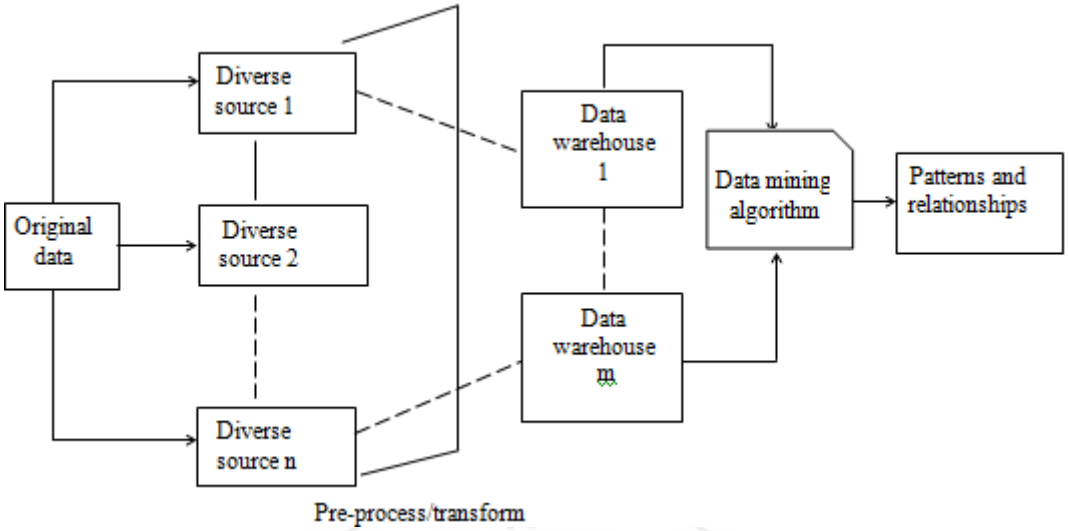

Figure 1: Privacy preserving data-mining approach

As shown in fig.1 the raw material is taken from one or various institutions and databases are stored in the diverse sources. For better analysis the data is transformed or pre-processed in proper format and the modified data is stored in data warehouse. The data mining algorithms serve as filter which filters out valuable nuggets of information from huge amount of data. Privacy cannot be applied at single stage, but needs to be applied at all stages. In stage 1, raw data is collected from diverse sources and is converted into suitable appearance for systematic purposes and stored in the corresponding data warehouses. Privacy techniques are applied at this stage also while collecting data.At level two, in data warehouse, used for reporting and data analysis. They store current and past data and are used to create reports. Data from data warehouses is subjected to get through a number of processes. Blocking, suppression, perturbation, modification, generalization, sampling etc. are these processes. For the discovery of knowledge/information, data mining algorithms are applied to this processed data. Even the data mining algorithms are modified for the purpose of protecting privacy without sacrificing the objectives of data mining. At stage three, the knowledge revealed by data mining algorithms is checked for its sensitiveness towards disclosure risks.

1.3.1 Classification of Privacy Preserving Techniques

Privacy preservation techniques are classified based on the following dimensions[12][2].

1. Data Distribution: It is a type of data storage. Data can be either replicated (centralized) or Fragmented (distributed). Distributed scenario is classifies as horizontal partitioning or vertical partitioning. Horizontal partitioning means records are partitioned without partitioning of attributes in various databases while vertical partitioning is attribute wise data are partitioned in various databases.

2. Data Modification: Data modification means changing of original database to some secure form so that it remains protected and information is not leaked to other. Different methods for data modification are perturbation, blocking, swapping and sampling.

3. Data Mining Algorithm: Many data mining algorithms are used in isolation to one another like classification data mining algorithms of decision trees, Clustering data mining algorithms of K-means, rough sets, Bayesian networks, association rule mining algorithms and many more.

4. Data or Rule Hiding: Data or rule hiding is necessary for security purpose of individual sites.

5. Preservation: In huge amount of data, some data are sensitive and need to be protected so that it does not get leaked to other. Privacy preservation helps to provide security to such sensitive data by various techniques like perturbation, randomization, anonymization, condensation, Blocking, cryptography techniques etc.

1. Randomization:
In this technique the data is randomized by the data providers and the randomized data is sent to the data receiver. The received randomized data is then reconstructed using distribution reconstruction algorithm and original information can be revealed. It is a simple process but has a disadvantage that it treats all its records equally.

2. Perturbation
In Perturbation the original values are replaced with some duplicate data values so that the statistical information computed from the perturbed data does not differ from the statistical information computed from the original data to a larger extent. In perturbation approach, record released is synthetic i.e. it does not correspond to real world entities represented by the original data. So the individual records present in the perturbed data are useless to the human recipient as only statistical properties of the records are preserved. Perturbation can be done by using additive noise or data swapping or synthetic data generation. Since the perturbation method does not reconstruct the original values

but only the distributions, new algorithms are to be developed for mining of data.

3. Anonymization

Anonymization refers to hiding the sensitive information or identity of record owners. Though explicit identifiers are removed there is a danger of privacy intrusion due to quasi identifiers. The attributes in the quasi identifiers such as DOB, gender or other general information may help to identify the owner potentially. These types of records are found in voter list or medical field and when linked with publically available data can be used to infer the identity of the corresponding individual with high probability. Thus using this technique, a value is replaced with less semantic consistent value called generalization and in suppression values are blocked. When such data is combined with publically available data, the risk of identification is reduced due to data mining. Although the anonymization method ensures that the transformed data is true but suffers heavyinformation loss.

4. Condensation

Using this technique constrained clusters are built in the data set and then pseudo-data is produced. The name of the method is due to the condensation of data into multiple groups. Dynamic data update such as stream problems use condensation technique. If each group has "k" size, then it is also referred to as the level of that privacy preserving approach. The higher the level, the high is the amount of privacy. The statistics from each group is used in order to generate the corresponding pseudo-data. This is a simple privacy preservation approach but it is not efficient because it leads to loss of the information.

5. Cryptographic

Cryptographic technique is used to encrypt the data. This technique is used when two or more than two parties are involved in sharing of data but simultaneously are concerned about their privacy of the sensitive data. Various cryptographic algorithms are implemented to maintain privacy. The algorithms in the cryptographic techniques may be less complex mathematically than others and are more accurate. So mostly used in defence and educational fields.

2. Literature Review

In 1982, Andrew C. Yao from University of California presented a paper on Protocols for Secure Computations. He carried out different mathematical analysis and designed a protocol which was used in different applications as Secret Voting, Oblivious Negotiation, Private Querying of Database, Probabilistic Computations, etc.[3].

HillolKargupta, Qi Wang et al. described randomized data distortion techniques to encrypt the data for providing privacy to the sensitive data. This methodology attempts to hide the sensitive data by randomly modifying the data values often using additive noise. A random matrix-based spectral filtering technique is developed to retrieve original data from the dataset, distorted by adding random values. They demonstrated that in many cases random data distortion preserve very little data privacy. They also discussed possible avenues for the development of new privacy-preserving data mining techniques like exploiting multiplicative and colored noise for preserving privacy in data mining applications[4].

In the paper "A Decision Tree Based Categorical Value Clustering and Perturbation Technique for Preserving Privacy in Data Mining" the privacy preserving technique named perturbation was discussed[5].

In the Cornell University, Ashwin Machanavajjhalaet al. described *l*-diversity technique over *k*-anonymity technique. There were two types of attacks; one who can discover sensitive attributes and other knowing some background information, for these both types of attacks *k*-anonymity did not guarantee privacy against attackers. *l*-diversity was discovered as more privacy preserving than the *k*-anonymization technique[6].

"t-Closeness: Privacy Beyond *k*-Anonymity and *l*-Diversity", the paper discussed a new anonymization technique called t-closeness over *k*-anonymity and *l*-diversity. *k*-anonymity protects against identity disclosure, it does not provide sufficient protection against attribute disclosure. *l*-diversity attempts to solve the problem by requiring that each equivalence class has at least well-represented values for each sensitive attribute, but *l*-diversity has number of limitations, so t-closeness was discovered, which requires that the distribution of a sensitive attribute in any equivalence class is close to the distribution of the attribute in the overall table. It almost reduces the risk of data leakage[7].

In the Liberty University and University of Minnesota a dot-product protocol was discussed. It is an efficient cryptographic technique to provide privacy in multi-sharing environment. Introducing the semi-trusted party increases gain and performance of the system and as per the involvement of the organizations the protocol can be extended[8].
Whereas in the country the study over privacy preserving is made as:

In 2007, Durgesh Kumar Mishra studied an arithmetic cryptography protocol which is used for multi-party communication over anonymization technique. Anonymization technique is not that trustworthy, the arithmetic protocol uses polynomial function for the data encryption and thus the probability of leakage is quite high[9].

In 2011, an approach was made by Anita A. Parmar for classification rule hiding to preserve privacy. A blocking based approach was made so that the mediator should not learn the sensitive patterns. Also in the Anna University of Technology, Tirunelveli a suppression based technique was studied to achieve privacy using k-anonymity[10].

And in the recent years a survey is continuing on the privacy preservation techniques where the better techniques for privacy preservation are being explored and studied[11].Various soft computing techniques using fuzzy

logic and neural network are being studied for the better privacy accuracy and reduced data leakage.

Xiaokui Shu et al. described a new privacy preserving technique called Fuzzy-fingerprinting technique which enhanced the data privacy but also helped to find out the true data leak. There are various data-leak cases, one of the causes of which is human mistakes etc. The common approaches do not provide that much security to the data owner and user. The method described provides better reliability to the sensitive data. To implement the method, various shingles are formed, then the fingerprints are made and the sensitive data is encrypted and outsourced. The mediator or server cannot even interpret the exact data due to enhanced security. Thus the probability of data-leak is reduced and so the sensitive data get enhanced security due to this method[1].

Table 1: Brief Survey Of Comparison Between The Data Privacy Preservation Techniques

Sr. no	Name	Technique	Advantages	Disadvantages
1.	Integer partitioning based encryption	Encryption	Provides greater amount of protection	Complex mathematical computations
2.	Additive Perturbation	Randomization	Identification of data is not possible directly	Weak and less reliable
3.	Perturbation by random projection technique	Randomization	Simple and original data value cannot be easily guessed	Reconstruction of data leads to leakage of data privacy
4.	Oblivious Transfer	Cryptography	Multi-party sharing without revealing sensitive data	Breach of trust may take place
5.	k-anonymous method	k-anonimity	Granularity reduced	Susceptible more to attacks
6.	Cooperative Model	Cryptography Approach	Major workload can be set on small set of servers	Used in multi-party communication so trust may breach

3. Conclusions

In today's world, privacy is the major concern to protect the sensitive data. People are very much concerned about their sensitive information which they don't want to share. From the survey on existing literature about privacy preserving it can be concluded that there is no single technique which is consistent in all domains. All methods perform in a different way depending on the type of data as well as the type of application or domain. The proposed system uses simple mathematical computations, less code so less time consumption, does not degrade much quality of the original data and has higher accuracy.

References

[1] XiaokuiShu, Danfeng Yao, Member, IEEE, and Elisa Bertino, Fellow, IEEE, "Privacy-Preserving Detection of Sensitive Data Exposure", *IEEE TRANSACTIONS ON INFORMATION FORENSICS AND SECURITY*, VOL. 10, NO. 5, MAY 2015

[2] Majid Bashir Malik, M. Asger Ghazi, Rashid Ali, "Privacy Preserving Data Mining Techniques: Current Scenario and Future Prospects", *Third International Conference on Computer and Communication Technology IEEE,* 2012

[3] Andrew C. Yao, "Protocols for Secure Computations", *University of California Berkeley,* California 94720, 1982

[4] HillolKargupta and Souptik Datta, Qi Wang and Krishnamoorthy Sivakumar, "On the Privacy Preserving Properties of Random Data Perturbation Techniques", *Proceedings of the Third IEEE International Conference on Data Mining (ICDM'03),* IEEE, 2003

[5] Md. Zahidul Islam and Ljiljana Brankovic, "DETECTIVE: A Decision Tree Based Categorical Value Clustering and Perturbation Technique for Preserving Privacy in Data Mining", *3rd IEEE International Conference on Industrial Informatics (INDIN),*2005

[6] AshwinMachanavajjhala Johannes Gehrke Daniel Kifer Muthuramakrishnan Venkitasubramaniam, "-Diversity: Privacy Beyond k-Anonymity", *Proceedings of the 22nd International Conference on Data Engineering (ICDE'06),*2006

[7] Ninghui Li, Tiancheng Li, Suresh Venkatasubramanian, "t-Closeness: Privacy Beyond k-Anonymity and l-Diversity", *IEEE*,2007

[8] Mark Shaneck,Yongdae Kim, "Efficient Cryptographic Primitives for Private Data Mining", *Proceedings of the 43rd Hawaii International Conference on System Sciences,* IEEE,2010

[9] Durgesh Kumar Mishra, Manohar Chandwani, "Arithmetic Cryptography Protocol for Secure Multi-party Computation", *IEEE*,2007

[10] Anita A. Parmar,Udai Pratap Rao, Dhiren R. Patel, "Blocking based approach for classification Rule hiding to Preserve the Privacy in Database", *International Symposium on Computer Science and Society IEEE,*2011

[11] SavitaLohiya, LataRagha, "Privacy Preserving in Data Mining Using Hybrid Approach", *Fourth International Conference on Computational Intelligence and Communication Networks, IEEE,*2012

[12] Ms. Dhanalakshmi.M, Mrs.Siva Sankari.E, "Privacy Preserving Data Mining Techniques-Survey", *IEEE,* 2014

[13] Risk Based Security. (Feb. 2014). *Data Breach Quick-View: An Executive's Guide to 2013 Data Breach Trends*. [Online] Available:https://www.riskbasedsecurity.com/reports/2013-DataBreachQuickView.pdf, accessed Oct. 2014

[14] Ponemon Institute. (May 2013). *2013 Cost of Data Breach Study: Global Analysis*.[Online] Available:https://www4.symantec.com/mktginfo/whitepaper/053013_GL_NA_WP_Ponemon-2013-Cost-of-a-Data-Breach-Report_daiNA_cta72382.pdf, accessed Oct. 2014.