

# Data Security in Big Data

Er. Jyoti Arora\* and Misha Gupta\*\*

Assistant Professor\* and M.Tech(CSE) Scholar\*\*

Desh Bhagat University, Mandi Govindgarh (Punjab)

## Abstract

*The emergence of big data and the potential to undertake complex analysis of very large data sets is, essentially, a consequence of recent advances in the technology that allow this. Government agencies will need to manage these new requirements efficiently in order to deliver net benefits through the adoption of new technologies. In this paper we have tried to safeguard our data from public use by identifying the legitimate user and attacker. We have also tried to make different sensitivity levels for different users*

**Keywords:** Data Provider, Data miner, Data collector, Decision maker and legitimate user.

## I. Introduction

Data mining, the extraction of hidden predictive information from large databases, is a powerful new technology with great potential to help companies focus on the most important information in their data warehouses. Data mining tools predict future trends and behaviors, allowing businesses to make proactive, knowledge-driven decisions. The automated, prospective analyses offered by data mining move beyond the analyses of past events provided by retrospective tools typical of decision support systems. Data mining tools can answer business questions that traditionally were too time consuming to resolve. They scour

databases for hidden patterns, finding predictive information that experts may miss because it lies outside their expectations.

Data mining techniques are the result of a long process of research and product development. This evolution began when business data was first stored on computers, continued with improvements in data access, and more recently, generated technologies that allow users to navigate through their data in real time. Data mining takes this evolutionary process beyond retrospective data access and navigation to prospective and proactive information delivery. Data mining is ready for application in the business community because it is supported by three technologies that are now sufficiently mature:

- Massive data collection
- Powerful multiprocessor computers
- Data mining algorithms

To deal with the privacy issues in data mining, a subset of data mining, referred to as privacy preserving data mining (PPDM) has gained a great development in recent years. The objective of PPDM is to safeguard sensitive information from unsolicited or unsanctioned disclosure, and meanwhile, preserve the utility of the data. The consideration of PPDM is two-fold. First, sensitive raw data, such as individual's ID card number and cell phone

number, should not be directly used for mining. Second, sensitive mining results whose disclosure will result in privacy violation should be excluded.

We present an overview of the many approaches which can help to make proper use of sensitive data and protect the security of sensitive information discovered by data mining. We use the term "sensitive information" to refer to privileged or proprietary information that only certain people are allowed to see and that is therefore not accessible to everyone. If sensitive information is lost or used in any way other than intended, the result can be severe damage to the person or organization to which that information belongs. The term "sensitive data" refers to data from which sensitive information can be extracted. Throughout the paper, we consider the two terms "privacy" and "sensitive information" are interchangeable. In this paper, we develop a user-role based methodology to conduct the review of related studies. Based on the stage division in KDD process, we can identify four different types of users, namely four user roles, in a typical data mining scenario.

**Data Provider:** the user who owns some data that are desired by the data mining task.

**Data Collector:** The user who collects data from data providers and then publishes the data to the data miner.

**Data Miner:** The user who performs data mining tasks on the data.

**Decision Maker:** The user who makes decisions based on the data mining results in order to achieve certain goals.

## II. Literature Review

*Thuraisingham, B. (2011)* Data mining has many applications in security including for national security as well as for cyber security. The threats to national security include attacking buildings, destroying critical infrastructures such as power grids and telecommunication systems. Data mining techniques are being investigated to find out who the suspicious people are and who is capable of carrying out terrorist activities. Cyber security is involved with protecting the computer and network systems against corruption due to Trojan horses, worms and viruses. Data mining is also being applied to provide solutions such as intrusion detection and auditing. Classification may be used to group various cyber attacks and then use the profiles to detect an attack when it occurs. Prediction may be used to determine potential future attacks depending in a way on information learnt about terrorists through email and phone conversations.

*Lei Xu (2014)* An emerging research topic in data mining, known as privacy-preserving data mining (PPDM), has been extensively studied in recent years. The basic idea of PPDM is to modify the data in such a way so as to perform data mining algorithms effectively without compromising the security of sensitive information contained in the data. Current studies of PPDM mainly focus on how to reduce the privacy risk brought by data mining operations, while in fact, unwanted disclosure of sensitive information may also happen in the process of data collecting, data publishing, and information (i.e., the data mining results) delivering. In this paper, we view the privacy issues related to data mining from a wider perspective and investigate various approaches that can help to protect

sensitive information. In particular, we identify four different types of users involved in data mining applications, namely, data provider, data collector, data miner, and decision maker.

*Nasrin Irshad Hussain, Bharadwaj Choudhury and Sandip Rakshit (2014)* privacy-preserving data mining has been emerged as a popular research area for the security of sensitive information in the network. In this field we study the extraction of knowledge or pattern from big data maintaining the commercial or legislative privacy constraints. Privacy preserving mining of distributed data has diverse applications. We have numerous algorithmic techniques for privacy preserving data mining. This paper presents a privacy preserving method for big data. We proposed a novel method for preserving privacy in mining big databases. Our goal is to mining the data while preserving data privacy and confidentiality.

*Jaseena K.U. (2014)* The Big Data introduce unique computational and statistical challenges including scalability and storage bottlenecks, noise accumulation, and spurious correlation and measurement errors. These challenges are distinguished and rewire new computational and statistical paradigm. This paper presents the literature review about Big data Mining and the issues and challenges with emphasis on the distinguished features of big data. It also discusses some method to deal with big data.

*Russom, Philip (2013)* Big Data analytics – the process of analyzing and mining Big Data – can produce operational and business knowledge at an unprecedented scale and specificity. The need to analyze and leverage trend data collected by

businesses is one of the main drivers for Big Data analysis tools. The goal of Big Data analytics for security is to obtain actionable intelligence in real time. Although Big Data analytics have significant promise, there are a number of challenges that must be overcome to realize its true potential.

*Ghinita, Gabriel (2008)* Privacy Preservation for high dimensional database is important. There are two popular data anonymization technique Generalization and Bucketization. These techniques are designed for privacy preserving microdata publishing. Our Proposed work include a slicing technique which is better than generalization and bucketization for the high dimension data sets. Slicing preserves better data utility than generalization and can be used for membership disclosure protection. Another important advantage of slicing is that it can handle high-dimensional data.

### III. Proposed Work

With increase in the size of data every year by 40% globally data mining and its consequences both have become important factors of research. The tech giants such as Google, Facebook etc. are looking for patterns in data for enhancing the user experience. With data mining for user experience enhancement there are some negative outcomes that happened with this. People have reported of leakage of their private information and content. So this privacy control along with advancement in data mining models has become important factor of research. In the study we did on the topic we decided to give our contribution to this research because this factor is of very importance and touches million lives. Doing research on topic and studying various factors we found that there is a requirement of Internet model

which prevents the expose of user private data and sensitive information on all types of user levels i.e. data provider, data collector, data miner and decision maker. On the basis of this problem we encountered we decided following objectives.

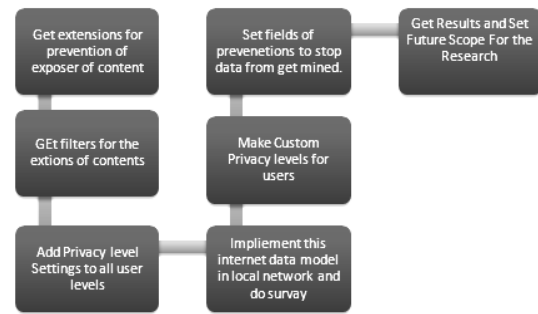
The data-mining model designed will give extensions for prevention and expose to every user level. The extensions provided will be on data such as:

- Phone Numbers.
- Personal Shopping details
- Address
- Family Details

The above factors are the privacy fields we will design system, which will not allow access to these sensitive information with restriction, enable feature for all four types of users. While users end is providing personal information of content browsing we set field of prevention, This will make sure that information respective to private fields will not shared anywhere except for the user and service provider.

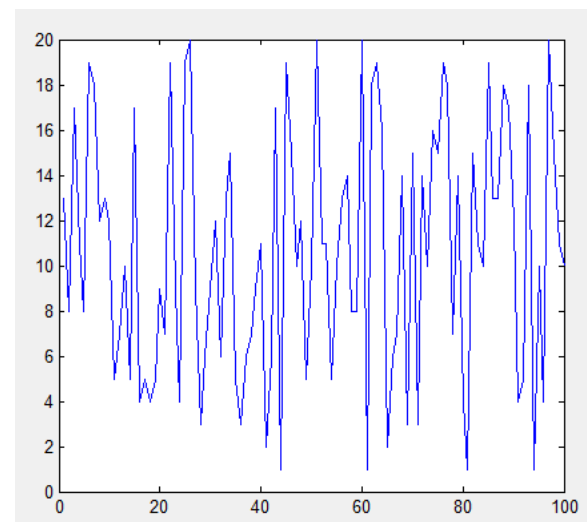
As we know that privacy level varies from person to person and this should be implemented in research also so we will make sure that in our model every user can set his privacy level. He can prevent his any information to be shared or dragged, or he can allow his all information to be shared. This implementation is done on all four users to provide extra robustness to the security.

This Internet net model will be used on local university connection and in friend zones to make survey on the success rate of this approach on the Internet. And with help of live information we will perform improvements to our systems and set future scopes for the research.

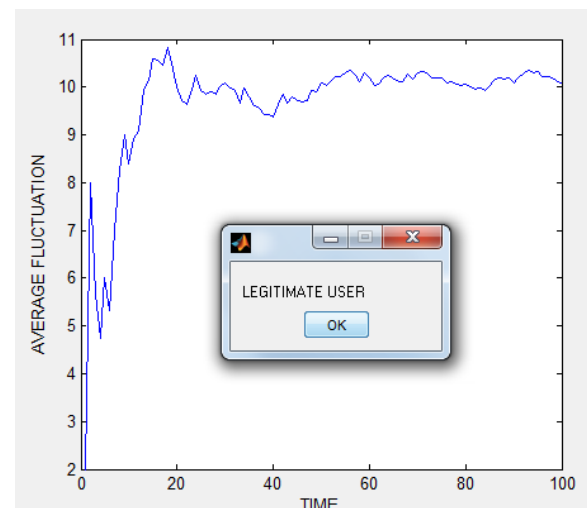


#### IV. Results and Discussions

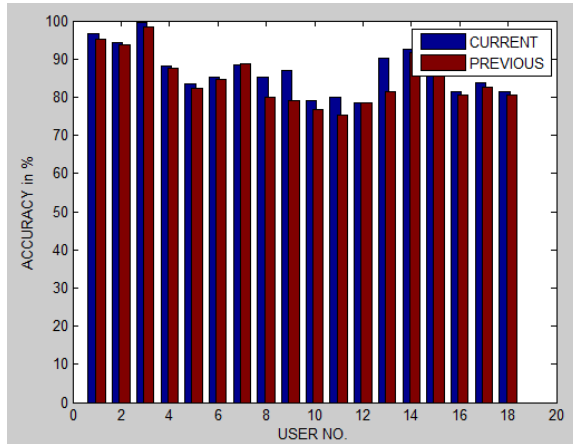
The experiment is setup and performed in Matlab. The results are collected and presented here.



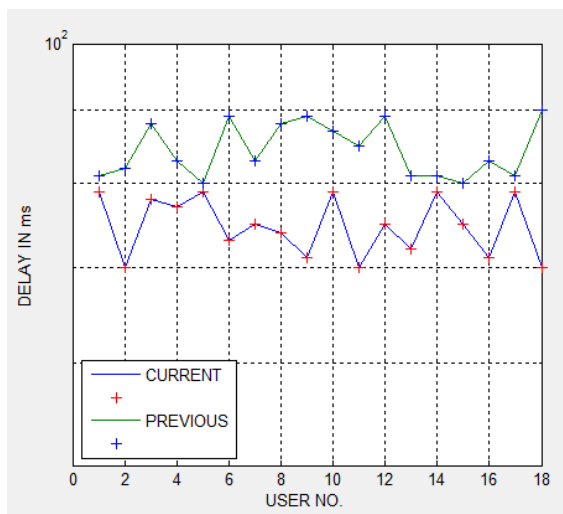
The graph above shows the data monitoring and processing. The graph is showing the time vs. data which is processed which is enumerated.



The above graph shows the average fluctuation vs. time. It is like the signature to determine if the user is legitimate or fake.



This is the comparison of the accuracy vs. number of users in previous work as well as in our work. We can see that the accuracy in our work is higher in comparison to previous work.



This graph shows the processing delay vs. number of users for the two works, the base work and the proposed work. We can see that in proposed work, we have lower delays.

## V. Conclusion

The emergence of big data and the potential to undertake complex analysis of very large data sets is, essentially, a

consequence of recent advances in the technology that allow this. Government agencies will need to manage these new requirements efficiently in order to deliver net benefits through the adoption of new technologies .Here in this work we have tried to safeguard our data from public use by identifying the legitimate user and attacker. We have also tried to make different sensitivity levels for different users.

## References

- [1] Xu, Lei, et al. "Information Security in Big Data: Privacy and Data Mining." 2014
- [2] IrshadHussain, Nasrin, Bharadwaj Choudhury, and Sandip Rakshit. "A Novel Method for Preserving Privacy in Big-Data Mining." International Journal of Computer Applications ,2014
- [3] Jaseena, K. U., and Julie M. David. "ISSUES, CHALLENGES, AND SOLUTIONS: BIG DATA MINING.",2008
- [4] Thuraisingham, Bhavani. "Big Data Security and Privacy." Proceedings of the 5th ACM Conference on Data and Application Security and Privacy. ACM, 2015.
- [5] Russom, Philip. "Big data analytics." TDWI Best Practices Report, Fourth Quarter (2011).
- [6] Ghinita, Gabriel, Yufei Tao, and Panos Kalnis. "On the anonymization of sparse high-dimensional data." IEEE 2008.
- [7] Botta, Marco, et al. "Query languages supporting descriptive rule mining: a comparative study." Database Support for Data Mining Applications. Springer Berlin Heidelberg, 2004. 24-51.

- [8] Khan, Danish. "CAKE–Classifying, Associating and Knowledge DiscovEry- An Approach for Distributed Data Mining (DDM) Using PArallel Data Mining Agents (PADMAs)." Web Intelligence and Intelligent Agent Technology, 2008. WI-IAT'08. IEEE/WIC/ACM International Conference on. Vol. 3. IEEE, 2008.
- [9] Shahriar, Md Sumon, and Sarawat Anam. "Quality Data for Data Mining and Data Mining for Quality Data: A Constraint Based Approach in XML." Future Generation Communication and Networking Symposia, 2008. FGCNS'08. Second International Conference on. Vol. 2. IEEE, 2008.
- [10] Bao, Feng, Xu He, and Fengzhi Zhao. "Applying data mining to the geosciences data." Computer, Mechatronics, Control and Electronic Engineering (CMCE), 2010 International Conference on. Vol. 5. IEEE, 2010.