Privacy Preservation in Big Data

Anjana Gosain
USICT
Guru Gobind Singh Indraprastha University
Delhi, India

Nikita Chugh
USICT
Guru Gobind Singh Indraprastha University
Delhi, India

ABSTRACT
Big data has brought a revolution in the world of data analytics. Data that was discarded a few years back is now considered a powerful asset. Big data is now being extensively used for knowledge discovery by all sectors of society. It is produced by almost all digital processes and is stored, shared on web. This reliance of big data on web model poses serious security concerns. Traditional security methods cannot be applied to big data due to its large volume, variety and volume. Also since big data contains person specific information, privacy is a major security concern. The three important privacy preservation methods are: data anonymization, notice and consent and differential privacy. In this paper we discuss these privacy preservation methods for big data and how differential privacy is a better solution for big data privacy.

Keywords
Big data, Data privacy, Anonymization, Differential privacy, Notice and consent

1. INTRODUCTION
Big data is among one of the emerging technologies that are bringing revolution in the world of data analytics. It has the power to provide insights into the unseen aspects of data analysis. The term big data is used to indicate large volumes of high variety data being generated at high velocity. It has now become an asset that can be used to make businesses smarter and hence more profitable.

Big data is generated from sources like online transactions, search queries, mobile phones, emails, videos etc. It is stored by partitioning along various servers. This implies that big data sources and storage systems are scattered all around internet. Since it is extensively based on the web model, security issues are a major challenge in big data analytics. Not ensuring secure big data analytics may cause great loses to both people and organizations.

Exposing whole of the big data may yield good analytics results but at the same time can pose great security challenges. The value hidden in big data can be of great value to hackers and invaders. So, there is a tradeoff between big data availability and big data security. One needs to ensure appropriate balance between the two. Hence the need is to protect knowledge hidden in big data during the whole analytics process.

Traditional security mechanisms fail to handle big data due its large volume, variety and velocity. Among various security aspects of big data, privacy is one of the most important issues [1]. This is because big data generally comprises of person specific information. Activities of individuals on social media, search engines etc are recorded and are a part of big data. This information is stored, linked and shared on web. Hence, unsecure big data analytics can lead to exposure of Personal Identifiable Information (PII). This causes customers to lose faith in organizations.

Traditional methods like cryptography can be used but they don't prove to be efficient because of complex nature of data [2].

Data anonymization or de-identification is also helpful in hiding personal information. It is the process of changing data that will be used or published in a way that prevents the identification of key information [3]. There are basically three data anonymization methods that are used in preserving big data privacy. They are: K-Anonymity [3,4], L-Diversity[3], T-Closeness [5].

Another method used for ensuring consumer privacy is Notice and Consent [6]. By this method consumer information is shared only after obtaining consent from the user using a notice. This method is usually used when user uses a new app or a new web service.

Differential privacy is another big data privacy preservation method that is being widely used. It is a method enabling analysts to extract useful answers from databases containing personal information while offering strong individual privacy protections [7].

In this paper we describe various measures that can help to ensure privacy in big data. This paper is worded as follows: Section 2 gives a review of traditional methods like cryptography and encryption to protect data privacy. Section 3 discusses privacy preservation methods for big data. Section 4 concludes our work.

2. TRADITIONAL DATA PRIVACY PRESERVATION METHODS
Cryptography refers to set of techniques and algorithms for protecting data. In cryptography plaintext is converted into cipher text using various encryption schemes. There are various methods based on this scheme like public key cryptography, digital signatures etc.

Cryptography alone can't enforce the privacy demanded by common cloud computing and big data services [2]. This is because big data differs from traditional large data sets on the basis of three V's (velocity, variety, volume) [8, 9]. It is these features of big data that make big data architecture different from traditional information architectures. These changes in architecture and its complex nature make cryptography and traditional encryption schemes not scalable up to the privacy needs of big data.

The challenge with cryptography is all or nothing retrieval policy of encrypted data [10]. The less sensitive data that can be useful in big data analytics is also encrypted and user is not allowed to access it. It makes data inaccessible to those who don't have access to decryption key. Also privacy may be breached if data is stolen before encryption or cryptographic keys are misused.

Attribute based encryption can also be used for big data privacy [11, 12]. This method of securing big data is based on relationships among attributes present in big data. The attributes that need to be protected are identified based on type of big data and company policies.

In nutshell, encryption or cryptography alone can't stand as big data privacy preservation method. They can help us to do data anonymization but cannot be used directly for big data privacy.

3. PRIVACY PRESERVATION FOR BIG DATA

Privacy in big data has raised serious concerns bringing into notice the need for efficient privacy preservation methods. In this section we discuss three privacy preservation methods: data anonymization, notice and consent and differential privacy. Also we look into how these methods can be applied to big data and their limitations when applied to big data.

Table 1: Base Dataset

Age	Sex	City	Income
24	M	Delhi	1,00,000
24	M	Gurgaon	18,000
24	M	Gurgaon	25,500
24	M	Delhi	12,000
26	F	Delhi	20,000
26	F	Delhi	50,000
26	M	Delhi	29,000
26	F	Delhi	48,000
32	M	Delhi	26,000
32	F	Gurgaon	45,000
32	F	Gurgaon	34,000
32	M	Delhi	34,000

3.1. Data Anonymization

Data anonymization is the process of changing data that will be used or published in a way that prevents the identification of key information [3]. It is also sometimes referred as data de-identification. In this method key pieces of confidential data are obscured in a way that maintains data privacy [13]. Organizations release data publically by anonymizing it. Anonymization in this case generally refers to hiding identifier attributes (attributes that uniquely identify individuals) like full name, license number, voter id etc. The main problem with data anonymization is that data may look anonymous but re-identification can be done easily by linking it to other external data [4]. In [4], it is shown that re-identification of anonymous medical records can be done using external voter list data. The attributes like gender, date of birth, zip code that can be combined with external data to re-identify individuals are called quasi identifier attributes.

For example, Table 1 represents the data set that needs to be analyzed for obtaining income trends without disclosing individual identity.

Table 2 represents data made anonymous by removing identifier attribute Voter ID. This table may look anonymous but can be linked with external data of to re-identify individuals.

Table 2: Anonymous Dataset

Voter ID	Age	Sex	City	Income
	24	M	Delhi	1,00,000
	24	M	Gurgaon	18,000
	24	M	Gurgaon	25,500
	24	M	Delhi	12,000
	26	F	Delhi	20,000
	26	F	Delhi	50,000
	26	M	Delhi	29,000
	26	F	Delhi	48,000
	32	M	Delhi	26,000
	32	F	Gurgaon	45,000
	32	F	Gurgaon	34,000
	32	M	Delhi	34,000

In the literature, mainly 3 privacy-preserving methods based on data anonymization are discussed: K-Anonymity [3, 4], L-Diversity [3], T-Closeness [5].

3.1.2. K-Anonymity

A dataset is called k-anonymized if for any tuple with given attributes in the dataset there are at least k-1 other records that match those attributes [3, 4]. K-anonymity can be achieved by using suppression and generalization [14]. In suppression, quasi identifiers are replaced or obscured by some constant values like 0,* etc. In generalization, quasi identifiers are replaced by more general values from levels up the hierarchy.

For example, in Table 1, Voter id and name are identifier attributes. Age, DOB, City are quasi identifiers. Income is a sensitive attribute.

Table 3: 2-Anonymized Dataset (Using Suppression)

Age	Sex	City	Income
2*	M	Delhi	1,00,000
2*	M	Gurgaon	18,000
2*	M	Gurgaon	25,500
2*	M	Delhi	12,000
2*	F	Delhi	20,000
2*	F	Delhi	50,000
2*	M	Delhi	29,000
2*	F	Delhi	48,000
3*	M	Delhi	26,000
3*	F	Gurgaon	45,000
3*	F	Gurgaon	34,000
3*	M	Delhi	34,000

Table 3 shows 2-anonymized version of table 2 using suppression. Here, age attribute has been suppressed and k=2.

K-anonymous data can still be vulnerable to attacks like unsorted matching attack, temporal attack, and complementary release attack [4]. Therefore we move towards L-diversity method of data anonymization.

3.1.3. L-Diversity

L-diversity technique of data anonymization tries to bring diversity in the sensitive attribute of data. It ensures that each equivalence class of quasi identifiers has atleast L different values of sensitive attribute [3].

In Table 1 income is a sensitive attribute. For data to be L-diverse there should be atleast L different values of income associated with each equivalence class. Table 4 shows 3-diverse version of table 1 since each equivalence class has atleast 3 different values for sensitive attribute income.

The problem with this method is that it depends upon the range of sensitive attribute. If we want to make data L diverse whereas sensitive attribute has less than L different values, fictitious data is to be inserted. This fictitious data will enhance the security but may result in problems during analysis. Also L-diversity method is prone to skewness and similarity attack and thus can't prevent attribute disclosure [13, 5].

Table 4: 2-Anonymized Dataset (Using Generalization), 3-Diverse Dataset

Age	Sex	City	Income
24	Person	ncr	1,00,000
24	Person	ncr	18,000
24	Person	ncr	25,500
24	Person	ncr	12,000
26	Person	ncr	20,000
26	Person	ncr	50,000
26	Person	ncr	29,000
26	Person	ncr	48,000
32	Person	ncr	26,000
32	Person	ncr	45,000
32	Person	ncr	34,000
32	Person	ncr	34,000

3.1.4. T – Closeness

An equivalence class is said to have t-closeness if the distance between the distribution of a sensitive attribute in this class and the distribution of the attribute in the whole table is no more than a threshold t. A table is said to have t-closeness if all equivalence classes have t-closeness. [5]. The main advantage of t-closeness is that it prevents attribute disclosure.

Data anonymization can be applied to big data but the problem lies in the fact that as size and variety of data increases, the chances of re-identification also increase. Thus, anonymization has a limited potential in the field of big data privacy.

3.2. Notice and Consent

The most common privacy preservation method for web services is notice and consent [6]. Every time an individual accesses a new application or service, a notice stating privacy concerns is displayed. The consumer needs to consent the notice before using the service. This method empowers an individual to ensure his privacy rights. It puts the burden of privacy preservation on the individual [6].

When applied to big data, this method poses numerous challenges [6]. In most of the cases uses of big data are unexpected or unknown at the time when notice and consent is given. This requires the notice to change every time big data is used for a different purpose. Also big data is collected and processed so rapidly that it creates burden on consumers to consent the notice.A method by which notice and consent can be modified for big data is the use of third parties offering a choice of different privacy profiles.

3.3. Differential Privacy

Differential Privacy is a method enabling analysts to extract useful answers from databases containing personal information while offering strong individual privacy protections [7, 15]. It aims to minimize the chances of individual identification while querying the data. The method of differential privacy is shown in fig. 1.

As opposed to anonymization, data is not modified in differential privacy. Users don't have direct access to the database. There is an interface that calculates the results and adds desired inaccuracies. It acts as a firewall. These inaccuracies are large enough that they protect privacy, but small enough that the answers provided to analysts and researchers are still useful [7].

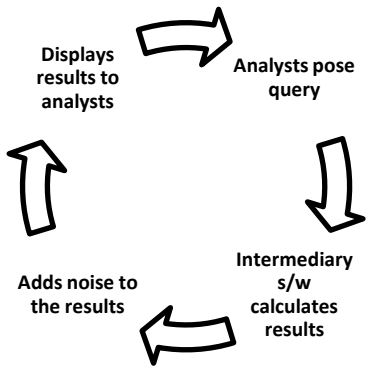

Fig. 1. Differential privacy process

The advantages of differential privacy over anonymization are:

- The original data set is not modified at all. There is no need for suppression or generalization.
- Distortion is added to the results by mathematical calculations based on the type of data, type of questions etc.
- The distortion is added in such a way that value hidden is useful to analysts.

4. CONCLUSION

Big data privacy has become an important issue since it is directly related to customers. It is now essential for an organization to promise privacy in big data analytics. Privacy measures should now focus on the uses of data rather than collection of data. They should be modified with respect to the size and unexpected uses of big data. Techniques like anonymization have limited potential when applied to big data. Notice and consent method also burdens the customer for ensuring privacy. Differential privacy may be seen as a viable solution for big data privacy. One problem with this method is that analyst should know the query before using the differential privacy model. When modified and applied to big data, it may ensure privacy without actually modifying the data.

5. REFERENCES

[1] X. Zhang, C. Liu, S. Nepal, C. Yang, J. Chen, "Privacy Preservation over Big Data in Cloud Systems," Security, Privacy and Trust in Cloud Systems, pp 239-257, Springer.

[2] M. V. Dijk, A. Juels, "On the impossibility of cryptography alone for privacy-preserving cloud computing," Proceedings of the 5th USENIX conference on Hot topics in security, August 10, 2010, pp.1-8.

[3] J. Sedayao, "Enhancing cloud security using data anonymization", White Paper, Intel Coporation.

[4] L. Sweeney, "k-anonymity: A model for protecting privacy," International Journal on Uncertainty, Fuzziness and Knowledge-Based Systems, pp. 557–570, 2002.

[5] N. Li, T. Li, S. Venkatasubramanian, "t-Closeness: Privacy Beyond k-Anonymity and l-Diversity**,** " IEEE 23rd International Conference on Data Engineering, 2007, pp. 106 - 115.

[6] F. H. Cate, V. M. Schönberger, "Notice and Consent in a World of Big Data," Microsoft Global Privacy Summit Summary Report and Outcomes, Nov 2012.

[7] J. Salido, "Differential privacy for everyone," White Paper, Microsoft Coproration, 2012.

[8] S. Sagiroglu and D. Sinanc, "Big Data: A Review," Proc. International Conference on Collaboration Technologies and Systems, 2013, pp. 42- 47

[9] Y. Demchenko, P. Grzsso, C. De Laat, P. Membrey, "Addressing Big Data Issues in Scientific Data Infrastructure," Proc. International Conference on Collaboration Technologies and Systems, 2013, pp. 48-55.

[10] Top Ten Big Data Security and Privacy Challenges, Technical report, Cloud Security Alliance, November 2012

[11] S. H. Kim, N. U. Kim, T. M. Chung, "Attribute Relationship Evaluation Methodology for Big Data Security," Proc. International Conference on IT Convergence and Security (ICITCS), 2013, pp. 1-4.

[12] S.H. Kim, J. H. Eom, T. M. Chung, "Big Data Security Hardening Methodology Using Attributes Relationship," Proc. International Conference on Information Science and Applications (ICISA), 2013, pp. 1-2.

[13] Big Data Privacy Preservation, Ericsson Labs, http://labs.ericsson.com/blog/privacy-preservation-in-big-data-analytics

[14] P. Samarati and L. Sweeney, "Protecting privacy when disclosing information: k-anonymity and its enforcement through generalization and cell suppression," Technical report, SRI International, 1998.

[15] O. Heffetz and K. Ligett, "Privacy and data-based research," NBER Working Paper, September 2013.