Advances in Computational Sciences and Technology
ISSN 0973-6107 Volume 10, Number 3 (2017) pp. 395-408

http://www.ripublication.com

A Survey on Big Data & Privacy Preserving Publishing Techniques

Dr. Puneet Goswami
Professor & Head(P.G & Research)
Dept. of Computer science and Engineering, SRM University Delhi-NCR, Sonipat.

Ms. Suman Madan
Research Scholar(SRM) and Astt. Prof(IT),
Jagan Institute of Management Studies, Sec-5,Rohini,Delhi – 85

Abstract

Big data describes very large data sets that have more diverse and complicated structure like weblogs, social media, email, sensors, and photographs. These less structured data and distinctiveness characteristics from traditional databases typically associated with extra complications in storing, analyzing and applying further procedures or extracting results. Big data analytics is the process of inspecting gigantic amounts of complex data in order to find out unseen patterns or recognizing furtive correlations. Since traditional databases systems cannot be used to process the big data, it poses numerous challenges to the research community. Security and privacy are the important concerns with data. However, there exists incongruity between the Big data security and privacy and the extensive use of big data. This paper gives insights on overview of big data, associated challenges, privacy and security concerns and the differentiation between privacy & security requirements in big data. Also, we focused on various privacy models which can be stretched to big data domain, analyzing the benefits and drawbacks of Data anonymity privacy models.

Index Terms: Big data Privacy & Security, Privacy models, Data anonymization

1. INTRODUCTION

Big data refers to pool of large datasets which cannot be processed using traditional computing techniques. Big data is not simply a data but it involves the data generated by variety of gadgets or devices or applications. For example: Black Box Data of planes and helicopters that captures voices of crew, various recordings of microphones, performance information etc., Social Media Data that holds information and views of people around the world, Stock Exchange Data that keeps customer decisions related to buying and selling information of different companies, Power Grid Data which holds information about consumption of different power grid nodes, Transport Data includes data related to vehicle's model, its capacity, distance and availability, Search Engine Data and many more.

The term "BIG DATA" was given by Roger Magoulas from O'Reilly media in 2005[1]. According to him, the gigantic size, complexness and wide range of data sets, it is almost becoming insoluble to handle and manage through traditional data analytical tools. Big data analytics is about joining trusted, internal information with new data types to create value bringing new source of unstructured info to existing core data to create insight about the information that is already existing but we never used it like Email, Blog, Stock Market, Sensors, Mobile Phone GPS etc. Big data analytics has the capacity to process any variety, volume and velocity of information and to derive an insight into data [2]. Various studies later highlighted that the definition of 3V's is not sufficient to describe the current big data scenario. Thus, veracity, validity, value, variability, venue, vocabulary, and vagueness were added to make some complement explanation of big data [3]. The main point in big data is data diversity, i.e., data may contain text, audio, image, or video etc. Big data integration can be used to construct systems that integrate structured, semi-structured and unstructured information from the published data. However, the major concern in data publishing is the privacy constraints. Privacy is a term associated to the right of individuals to control the visibility of their personal information to others. The standard method of data sharing focuses on removing personally recognizing information from the dataset that is published. But this method does not prevent linkage attacks. Many anonymization methods like *k*-anonymity have been proposed for privacy preserving data sharing.

2. BIG DATA CHARACTERISTICS

Doug Laney, an analyst for Gartner, had explained that the big data comprises of three dimensions: high volume, high velocity and high variety. However, there are other "Vs" that help in appreciating the real essence of big data and its effects [4]. The 7V's of big data are explained as follows:

1. *Volume:* The big data is enormous and continuously increasing volume in this digital world is due to the Internet of Things having sensors all over the world in all devices creating data every second. By 2020, it will be almost 50 times

more than the data we had in 2012. The enormous amount of data in social networks is due to the image, video, music and text data that is uploaded by different users, thus the era of a trillion sensors there.

2. *Velocity:* The velocity means speed at which the data is created, stored, analyzed and visualized. In this current scenario of big data, data is produced in almost real-time. With technological advancements, almost all devices and machines- wireless or wired transmit their data the moment it is created. The speed of data generation is almost inconceivable. Every minute around 100 hours of video on YouTube is uploaded, over 200 million emails are sent, around 20 million photos are viewed and almost 300,000 tweets are sent and almost 2.5 million searches on Google are performed. This forces big challenge on organizations to cope with the enormous speed of data generation and use in real-time.
3. *Variety* – In this increasingly digital world, the variety of data generation is unimaginable. To extract the meaningful information from unstructured text, images, audio, video and data from sensors in the IoT world requires ever-increasing algorithmic and computational power.
4. *Veracity* - Veracity means the data is verifiable and truthful. The huge potential of big data is of no use if data analysis is done on inaccurate or incomplete data, especially for automated decision-making, or providing data into an unsupervised machine learning algorithm. The inaccurate data can give catastrophic results. Since the data streams originate from diverse sources in variety of formats with varying signal-to-noise ratios, they may be plentiful accumulated errors that are difficult to sort out when data reaches Big Data analysis stage. Thus cleaning up of data is much required so that the veracity of the final analysis is not degraded.
5. *Value*: To justify the investments made on data collection, we should be able to generate some value out of it, may be by using Big Data or on traditional analytics, data warehouse or business intelligence tools. Recommendations given by various sites based on user preferences and click stream data are best examples to outline this critical characteristic of big data.
6. *Variability*: This refers to the data whose sense is constantly changing. For example: Natural language processing or sentiment analysis of textual data. Since a single word can have multiple meanings, new meanings are generated and old meanings cast-off over time. Here exists a unique decoding challenge for limitless variability of big data to correctly identify the sense of a word by understanding the context.
7. *Visualization*: Once the data is processed, it must be presented in a manner that's readable and accessible- this is where visualization comes in. Techniques should be adopted to represent this enormous volume of data in an efficient manner. One of the best methods is converting it into graphical formats. However, due to the velocity and variety attributes, spreadsheets and/or three-dimensional visualizations are mostly not up to the task. There may be a multitude of spatial and temporal parameters and relationships

between them to condense into visual forms. Dimensionality reduction plays significant role in visualization of massive datasets.

The 3V's i.e. Volume, Velocity and Variety are inherent to Big Data, the other 4 V's i.e. Variability, Veracity, Value and Visualization reflect the gigantic complexity of Big Data regarding who would process, analyze and benefit from it. It does not matter whether one set of V's is Victorious over another set of V's, all of them demand careful consideration and challenge for the researchers.

3. Challenges with Big Data

Big data sizes are continuously increasing from terabytes in 2012 to nearly 44 zettabytes by the year 2020 in a single data set[14]. To unearth unseen patterns and not known correlations for effective decision making, advanced analytical and visualization techniques of big data are applied to large data sets. The big data analysis involves many distinctive phases that include data acquisition and recording, information extraction and cleaning, data integration, aggregation and representation, query processing, data modeling and analysis and interpretation. Each of these phases introduces challenges. Big data mining challenges includes heterogeneity, complexity, scaling, timeliness and privacy.

3.1 Big data processing Framework

The big data challenges are majorly categorized in three tier processing framework: big data mining platform (tier 1), big data semantics (tier 2) and big data mining algorithms (tier 3) as shown in figure 1. The Tier II challenges focus on semantics and domain knowledge for different big data applications. Such information can give extra advantages to the data mining process but at the same time technical barriers to the Big Data access (Tier I) and mining algorithms (Tier III) are added [5][6].

Big Data Mining Platform

The challenges [5] at Tier I focus on data accessing and arithmetic computing procedures as big data are generally spread over different locations and data volumes constantly growing and an effective computing platform will have to take distributed large-scale data storage into consideration for computing. Thus the computing platform should have access to resources - Data and computing processors. For example, all data should be in main memory for data mining algorithms, this pose a technical barrier for big data because moving data across different locations is expensive due to network communication and other IO costs, even though a super large main memory to hold all data for computing is available. Since data scale in big data mining is much more than the capacity of a single personal computer, a typical framework will rely on cluster computers with a high performance computing platform. For data mining task on a large number of clusters or computing nodes, parallel programming tools, such as Map Reduce are deployed. The software component's role is to establish that a single data mining task is split into many small tasks and each small task runs on one or multiple computing nodes.

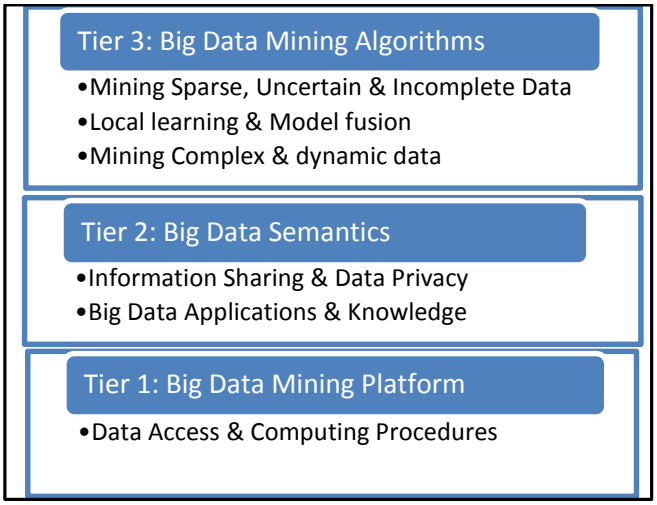

Figure 1: Big Data Processing Framework

3.1.2 Big Data Semantics

The Tier II challenges revolve around semantics and domain knowledge for different big data applications. This information not only provides extra benefits to the mining process but also add technical barriers to the Big Data access (Tier I) and mining algorithms (Tier III). The two most important issues at this tier include:

- Data sharing and privacy – It deals with data maintenance, its access and sharing issues.
- Domain and application knowledge – It deals with the answers about underlying applications and knowledge or patterns users wants to discover from the data.

3.1.3 Big Data Mining Algorithms

The challenges at Tier III concentrate on algorithm designs in dealing with the difficulties arisen due to big data volumes, distributed data distributions, and by complex and dynamic data characteristics. The Tier III contains three stages:

a) Data fusion techniques are used for pre-processing of sparse, uncertain, incomplete, heterogeneous and multi-source data.
b) Mining of complex and dynamic data after pre-processing.
c) Testing the overall knowledge which is gained by local learning and model fusion and giving back relevant information to the pre-processing stage. Then on the basis of feedback, the model and parameters are adjusted.

In the whole process, information sharing is not only a promise of smooth development of each stage, but also a purpose of Big Data processing.

3.2 Security and Privacy Challenges

This is the most important and sensitive challenge with big data which comprises of conceptual, technical and legal significance. Cloud Secure Alliance (CSA) has categorized different security and privacy challenges into four different facets of big data ecosystem.(Big Data Working group, CSA, 2013). These facets and their security challenges are categorized as follows:

3.2.1 Infrastructure Security

The distributed computations and data stores must be secured for secured big data systems. For distributed frameworks, the Map Reduce computation model allows parallel computation of data. The major attack avoidance methods focus on security of mappers and the data when an untrusted mapper is present. Trust establishment and MAC guarantee the reliability of mappers. Challenges are:

a) Having secure computations for the Distributed Programming Frameworks
b) Best security Practices for Non-Relational Data Stores.

3.2.2 Data Privacy

This challenging area in big data domain focuses on securing the data itself. To do this, privacy preservation is done before information exchange and circulation and the sensitive data must be secured cryptographically. In today's digital world, privacy preserving data mining and sharing are important areas because it provides maximum utility of published dataset without enquiring individual's privacy. Challenges are:

a) Privacy Preserving Data Mining and Analytics
b) Cryptographically Enforced Data Centric Security
c) Granular Access Control

3.2.3 Data Management and Integrity

For securing the data storage, managing massive datasets need efficient solutions. Granular access control mechanisms inhibit unauthorized users to access data elements. Audit information is however another important characteristic that provides better security by employing query auditing mechanisms. Challenges are:

a) Secure Data Storage and Transaction Logs
b) Granular Audits
c) Data Provenance

3.2.4 Reactive Security

It generally includes real-time security monitoring, end point input validation and filtering. Real-time security monitoring is related to checking the big data infrastructure and its applications. Since data is produced enormously, input data validation is a great challenge in big data. Challenges are:

a) End-point Validation and Filtering
b) Real Time Security Monitoring

These security and privacy challenges cover the entire spectrum of the Big Data lifecycle (Figure 2) i.e. data production source or devices, the data, data processor, storage of data, and data transport and data usage on different devices [13].

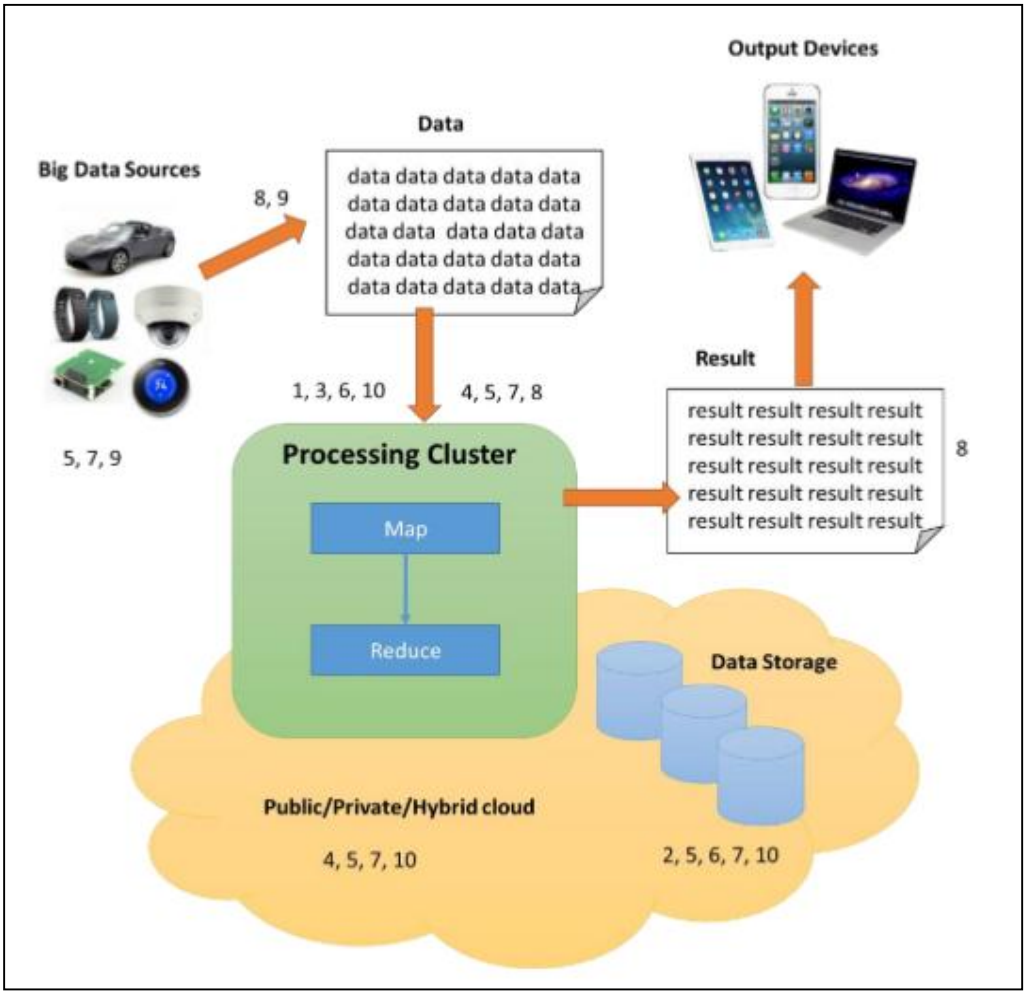

Figure 2: Security and Privacy challenges in Big Data system (adapted from CSA, 2013)

3.1

3.2

3.3 Difference between Security and Privacy

Security and Privacy in big data is an important issue. Security focusses on protecting data from pernicious attacks and stealing data for profit [7]. Data privacy focusses on the use and governance of individual's personal data like making policies to ensure that consumers' personal information is being collected, shared and utilized in right ways. Although security is vital for protecting data but it's insufficient for addressing privacy. Table 1 focuses on additional difference between security and privacy.

Table 1: Differentiation between security and privacy

S.No	Security	Privacy
1.	It is the "confidentiality, integrity and availability" of data	It is the appropriate use of user's information
2.	It may provide for confidentiality or to protect an enterprise	It concerns with consumer's right to safeguard their information from any other parties
3.	Various techniques like Encryption, Firewall etc. are used in order to prevent data compromise from technology or vulnerabilities in the network of an organization	The organization can't sell its customer/user's information to a third party without prior consent of user
4.	It offers the ability to be confident that decisions are respected	It is the ability to decide what information of an individual goes

4. Privacy preserving data publishing

Privacy preservation is a major issue for big data mining applications. The Privacy Preserving Data Publishing has two phases: data collection and data publish. In data collection phase, the dataset is collected by data publisher from data owner. Then the raw datasets collected is processed and in the data publishing phase, the processed dataset is sent to data recipient, process shown in figure 3.

Figure 3: PPDP phases

Before a data set is out for other parties, the possibility of identifying sensitive information about individuals is reduced by some privacy-preserving technique. This is called the disclosure-control problem [5][6]. There are two approaches to preserve privacy:

i. *Role based access control:* Restricting access to the data by adding certification or access control to the data entries so sensitive information is accessible to a limited user groups only. Challenge here is that no sensitive information can be misconduct by unauthorized individuals and thus secured certification or access control mechanisms must be designed.

ii. *Cryptographic Technique:* Sensitive information fields should be anonymized so that they cannot pinpoint to an individual record. The main challenge is to inject randomness into the data to ensure a number of privacy goals [5].

Various proposals have been designed for privacy preserving of data while publishing. These proposals can be divided into two categories: One is to achieve the privacy preservation by utilizing the methods of probability or statistics in the case of the statistical properties of the final data and classification properties are unchanged, such as clustering, randomization, sampling, cell suppression, data swapping and perturbation have been designed for data publishing. The other is based on Data anonymity model where the frequently used method is using the non-specific information instead of more sensitive and specific information i.e. the generalization of the information. In this paper, focus is on Data anonymization Method.

5. DATA ANONYMIZATION

Anonymization of data removes identifying attributes like aadhar No or names from the database. It is also referred as data de-identification. Table 2 shows base dataset that is to be analyzed for income trends of population. Aadhar card and name can be removed by data collectors. Income tax department may remove income attribute also. The records in the database may be categorized as follows:

i. *Explicit identifiers*- These are the attributes that helps in identifying an individual uniquely. For e.g.: Name, Aadhar No etc.
ii. *Quasi identifiers*- These are the attributes which are harmless but can be combined with other information for identifying an individual from the people's group. For e.g.: gender, age, city etc.
iii. *Sensitive identifiers*- These are the attributes with sensitive value with respect to data owner. This data is generally released and required by researchers. For e.g.: Income in table 2.

Data may look anonymous in data anonymization but major problem is re-identification that may be done effortlessly by combining it to other external data [8]. The data looks anonymous when we remove identifier attribute Aadhar No and Name, as shown in Table 3 but can be linked with external data of to re-identify individuals.

Table 2: Base Dataset

Aadhar No.	**Name**	**Age**	**Gender**	**City**	**Income**
765423458976	Raman	24	Male	Delhi	2,00,000
567812356521	Harish	27	Male	Delhi	1,50,000
723415437895	Zaika	24	Female	Gurugram	20,000
654398761234	Salman	36	Male	Delhi	36,500
456712369823	Priyanka	36	Female	Noida	45,000

543267891234	Meena	44	Female	Gurugram	75,000
876452341876	Rakesh	47	Male	Delhi	1,00,000

Table 3: Anonymous Dataset

Aadhar No.	**Name**	**Age**	**Gender**	**City**	**Income**
		24	Male	Delhi	2,00,000
		27	Male	Delhi	1,50,000
		24	Female	Gurugram	20,000
		36	Male	Delhi	36,500
		36	Female	Noida	45,000
		44	Female	Gurugram	75,000
		47	Male	Delhi	1,00,000

All the explicit identifiers will be removed and only the quasi identifiers and sensitive attributes are published during the data publishing phase. Modification of the dataset should be done before data publishing. This is done by execution of variety of anonymization operations on the dataset. The anonymization includes following approaches:

i. *Generalization & Suppression* – This is generally used to replace the specific values with more general ones that leads to many tuples will be having duplicate values for quasi identifiers. The term equivalence class can be defined as the set of tuples that have the same value for quasi identifiers. In suppression, we replace quasi identifiers by some constant values like 0,* etc.
ii. *Anatomization & Permutation*- The objective of this is de-linking the relation between quasi-identifiers and the sensitive attributes.
iii. *Perturbation* –It involves addition of some noise to the original data before giving that to the user

There are mainly three privacy-preserving methods based on data anonymization are discussed: K-Anonymity [8, 9], L-Diversity [9], and T-Closeness [10].

5.1 K-Anonymity

A dataset is k-anonymized when for any tuple having certain attributes in the dataset, there are at least k-1 other records that match those attributes [8,9]. It can be attained by using suppression and generalization [11]. Table 4 shows k-anonymity when k=2 and age attribute is suppressed using equivalence class

Table 4: 2-Anonymized Dataset

S.No.	Age (Equivalence class)	Age (after suppression)	Gender	City	Income
1	24 (20-30)	2*	Male	Delhi	2,00,000
2	27 (20-30)	2*	Male	Delhi	1,50,000
3	24 (20-30)	2*	Female	Gurugram	20,000
4	36 (31-40)	3*	Male	Delhi	36,500
5	36 (31-40)	3*	Female	Noida	45,000
6	44 (41-50)	4*	Female	Gurugram	75,000
7	47 (41-50)	4*	Male	Delhi	1,00,000

k-anonymity has drawbacks:

a) *Homogeneity attack (or Attribute linkage)* – It happens when the sensitive attribute lacks diversity. For example in S.No. 1 & 2 of table 3, age, gender and city are same.

b) *Background attack (or Record linkage)* – It occurs when the challenger has some background knowledge about the individual. *K*-anonymity approach does not help in preventing attribute disclosure which means that challenger will get additional insights about an individual even without combining it to any item in the published table.

5.2 L-Diversity

L-diversity method prevents homogeneity and background attacks of k-anonymity method. An equivalence class has L-diversity when there are at least "L well represented" values for the sensitive attribute. To obtain "L well represented" values, each equivalence class has at least L distinct values for the sensitive field. This is called Distinct L- diversity. Table 5 shows 2 diverse version of Table 1 since each equivalence class has at least 3 different values of income.

Table 5: 3- Diverse Dataset

Aadhar No.	Name	Age	Gender	City	Income
		24	Person	NCR	2,00,000
		27	Person	NCR	1,50,000
		24	Person	NCR	20,000
		36	Person	NCR	36,500
		36	Person	NCR	45,000

		44	Person	NCR	75,000
		47	Person	NCR	1,00,000

L-diversity has drawbacks:

a) *Skewness attack* – It occurs when each block of quasi identifiers or equivalence class has equal probability for positive and negative values of sensitive attributes.

b) *Similarity attack* – It happens when the values of sensitive attributes are actually similar in meaning but seems to be different.

5.3 T-Closeness

T-closeness method incorporates the *k*-anonymity and *l*-diversity approaches. If the *t*-closeness principle holds for a dataset, then the dataset confirms *k*-anonymity and *l*-diversity principles as well. 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. [10]. The main advantage of t-closeness is that it prevents attribute disclosure.

6. CONCLUSION

Big data privacy is a critical component in today's digital world where data is generated, accessed and shared widely with each other. It is now mandatory to promise privacy in big data analytics. Privacy measures should now give emphasis on the uses of data instead of data collection. Techniques like 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. This paper gives a good insight on big data, privacy issues and approaches.

REFERENCES

[1] http://www.researchtrends.com/wp-content/uploads/2012/09/Research_Trends_Issue30.pdf

[2] https://www-950.ibm.com/events/wwe/grp/grp037.nsf/vLookupPDFs/Calgary_Keynote_%20David_%20Corrigan%20-%20v1/$file/Calgary_Keynote_%20David_%20Corrigan%20-%20v1.pdf

[3] Tsai C-W, Lai C-F, Chao H-C, Vasilakos AV. Big data analytics: a survey. J Big Data Springer Open J. 2015.

[4] Katal, A., Wazid, M. and Goudar, R.H. (2013) 'Big data: issues,challenges, tools and good practices', in *2013 Sixth International Conference on Contemporary Computing (IC3)*,IEEE, pp.404–409.

[5] Data Mining With Big Data Xindong Wu, Fellow, IEEE, Xingquan Zhu, Senior Member, IEEE, Gong-Qing Wu, And Wei Ding, Senior Member, IEEE, IEEE Transactions On Knowledge And Data Engineering, Vol. 26, No. 1, January 2014.

[6] Review on Data Mining with Big Data" Vitthal Yenkar, Prof.Mahip Bartere, IJCSMC, Vol. 3, Issue. 4, April 2014

[7] Jing Q, et al. Security of the internet of things: perspectives and challenges. Wirel Netw. 2014;20(8):2481–501

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

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

[10] 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.

[11] 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.

[12] https://go.oracle.com/LP=35781?elqCampaignId=47670&src1=ad:pas:go:dg:bd&src2=wwmk160603p00065c0007&SC=sckw=WWMK160603P00065C0007&mkwid=s6IiPCV9C|pcrid|160705555388|pkw|big%20data%20privacy|pmt|e|pdv|c|sckw=srch:big%20data%20privacy

[13] Big Data Working Group, Cloud Security Alliance (2013) Expanded Top Ten Big Data Security and Privacy challenges

[14] 'Executive summary, data growth, business opportunities, and the IT imperatives' (2014) [online] http://www.emc.com/leadership/digitaluniverse/2014iview/executive-summary.htm (accessed 2015)

[15] Ke Wang, Benjamin C. M. Fung, Anonymizing sequential releases, In Proceedings of the 12th ACM SIGKDD Conference, 414-423 (2006).

[16] Raymond Chi-Wing Wong, Jiuyong Li, Ada Wai-Chee Fu, KeWang, (a, k)-anonymity: An enhanced k-anonymity model for privacy preserving data publishing, In Proceedings of the 12th ACM SIGKDD, 754-759 (2006).

[17] Qing Zhang, Koudas N., Srivastava D., Ting Yu, Aggregate query answering on anonymized tables, In Proceedings of the 23rd IEEE International Conference on Data Engineering (ICDE), 116-125 (2007).

[18] Ninghui Li, Tiancheng Li, Venkatasubramanian S., tcloseness: privacy beyond k-anonymity and l-diversity, In Proceedings of the 21st IEEE International Conference on Data Engineering (ICDE), 106-115 (2007).

[19] Xiaokui Xiao, Yufei Tao, Personalized privacy preservation, In Proceedings of the ACM SIGMOD Conference, 29-240 (2006).

[20] Mehmet Ercan Nergiz, Maurizio Atzori, Chris Clifton,Hiding the presence of individuals from shared databases, In Proceedings of ACM SIGMOD Conference, 665-676 (2007).

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

[22] 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.

[23] O. Heffetz and K. Ligett, “Privacy and data-based research,” NBER Working Paper, September 2013.