IJCRT.ORG ISSN : 2320-2882

INTERNATIONAL JOURNAL OF CREATIVE RESEARCH THOUGHTS (IJCRT)

An International Open Access, Peer-reviewed, Refereed Journal

DIFFERENT TECHNIQUES FOR PRIVACY PRESERVING IN BIG DATA: COMPARATIVE STUDY

[1]Ms. Himaniben Gajjar, [2]Dr. Nidhi Divecha

[1]Ph.D. Scholar, [2]Assistant Professor
[1]Computer Science,
[1]Kadi Sarva Vishwavidyalaya, Gandhinagar, Gujarat, India

Abstract: Improbable quantities of data is being produced by various organizations like hospitals, banks, e-commerce, retail and supply chain, etc. by virtue of digital technology. Not only individuals but machines also subsidize to data in the form of closed circuit television streaming, web site logs, etc. Tons of data is produced every minute by social media and smart phones. Big data is a word used for very large data sets that have more diverse and composite structure. These features usually correlate with additional problems in storing, analyzing and applying further procedures or mining results. This paper covers uses of privacy by taking existing methods such as k-anonymity, T-closeness and L-diversity and its implementation in business. There have been a various privacy-preserving mechanisms developed for privacy protection at different stages. Although data analytics is useful in decision making, it will lead to serious privacy apprehensions. Hence privacy preserving data analytics became very important. The goal of this paper is to deliver a major review of the privacy preservation mechanisms in big data and present the challenges for existing mechanisms and paper also examines various privacy threats with their limitations in unstructured data.

***Index Terms* - Data, Data analytics, Privacy threats, Privacy preservation, k-anonymity: T-closeness, L-diversity**

I. INTRODUCTION

Big data (Abadi DJ, 2003) (Kolomvatsos K, 2015) precisely refers to data sets that are so large or composite that traditional data processing applications are not adequate. It’s the huge size of data—both structured and unstructured—that overrun a business on a day-to-day basis. Due to current technical development, the quantity of data produced by internet, social networking sites, sensor networks, healthcare applications, and many other enterprises, is radically increasing day by day. All the massive amount of data produced from various sources in multiple formats with very high speed is referred as big data. The large scale data, which also include person specific private and delicate data like gender, zip code, disease, caste, shopping cart, religion etc. is being stored in public domain. The data holder can give this data to a third party data analyst to achieve bottomless insights and classify hidden patterns which are useful in making important decisions that may help in refining businesses, provide value added services to customers (Ducange Pietro, 2018) prediction, forecasting and recommendation (Chauhan Arun, 2017). Amazon, Flip kart also use recommendation systems for suggesting products to customers based on their buying traditions. While Facebook does suggest friend, place to visit based on user interest. In this paper we have deliberate a number of privacy preserving methods available which are being used to protect against privacy threats. Each of these techniques has their own qualities and drawbacks.

II. PRIVACY CONCERN IN BIG DATA

Privacy and security in terms of big data is significant concern. Big data security model is not recommended in the event of composite applications owed to which it gets disabled by default. However, in its nonappearance, data can always be bargained easily. As such, this section emphases on the privacy and security concerns.

Privacy Information privacy is the pleasure to have some control over how the personal info is collected and used. Information privacy is the volume of an specific or group to stop information about themselves from fetching known to persons other than those they give the information to. One serious user privacy issue is the proof of identity of personal information during communication over the Internet (Porambage P, 2016).

2.1 Privacy requirements in big data

Big data analytics attraction in numerous enterprise; a substantial portion of them decide not to utilize these services because of the nonappearance of standard security and privacy protection tools. The basics and development policies of a framework that supports:

- The dimension of privacy policies handling the access to data stored into aim big data platforms.
- The generation of fruitful implementation monitors for these policies, and
- The combination of the generated monitors into the mark analytics platforms. Enforcement methods offered for old-style DBMSs appear insufficient for the big data framework due to the strict execution requirements needed to handle large data volumes, the heterogeneity of the data, and the speed at which data must be analysed.

III. Privacy preserving methods in Big Data

Some of traditional methods for privacy preserving in big data is defined here. These methods being used to provide privacy to a certain amount but their drawbacks led to the initiation of newer methods.

3.1. K anonymity

In Anonymization process before it is given for data analytics it first adjusting the data (Iyengar S., 2002), so that de identification is not possible and will lead to K indistinguishable records if an effort is made to de identify by plotting the anonymized data with external data sources. K anonymity is inclined to two attacks namely homogeneity attack and back ground knowledge attack. Some of the algorithms applied include, Incognito (LeFevre K, 2005), Mondrian (LeFevre K, 2006) to confirm Anonymization. K anonymity is functional on the patient data shown in Table 1. The table shows data before anonymization. There are five attributes along with ten records in this data. There are two regular techniques for accomplishing k-anonymity for some value of k.

Table 3.1: A Non-anonymized database consisting of the patient records

Name	Age	Gender	Zipcode	Disease
Nikhil	45	Male	234587	Cancer
Nikunj	50	Male	245782	Heart-Related
Salini	24	Female	245783	TB
Joshna	25	Female	245785	Viral Infection

1. Suppression In this method, certain values of the attributes are supplanted by an asterisk '*'. All or some of the values of a column may be replaced by '*'. In the anonymized Table 1, replaced all the values in the 'Name' attribute and each of the values in the 'Religion' attribute by a '*'.

2. Generalization In this method, individual values of attributes are replaced with a broader category. For example, the value '28' of the attribute 'Age' may be supplanted by '≥25', the value '50' by '35 < age ≤ 50', etc.

Table 2 has 2-anonymity with respect to the attributes 'Age' and 'Gender' since for any blend of these attributes found in any row of the table there are always no less than two rows with those exact attributes. The attributes that are available to an adversary are called "quasi-identifiers". Each "quasi-identifier" tuple arises in at least k records for a dataset with k-anonymity. K-anonymous data can still be helpless against attacks like unsorted matching attack, temporal attack, and complementary release attack (Samarati P, 1998) (Sweeney L, 2002).

Table 3.2: 2-anonymity with respect to the attributes 'Age', 'Gender' and 'State of domicile'

Name	**Age**	**Gender**	**Zipcode**	**Disease**
*	30 < Age ≤34	Male	2345**	Cancer
*	35 < Age ≤50	Male	2457**	Heart-Related
*	25≥Age	Female	2457**	TB
*	25≥Age	Female	2457**	Viral Infection

3.2. T closeness

It is an additional enhancement of l-diversity group based anonymization that is used to preserve privacy in data sets by reducing the granularity of a data demonstration. This reduction is a trade-off that results in some loss of capability of data management or mining algorithms in order to gain some privacy. The t-closeness model(Equal/Hierarchical distance) (Li N, 2007) encompasses the l-diversity model by giving the values of an attribute definitely by considering into account the distribution of data values for that attribute. Similar class is said to have t-closeness if the distance between the transmission of a sensitive attribute in this class and the distribution of the attribute in the whole table is less than a threshold t. A table is said to have t-closeness if all similar classes have t-closeness. The main benefit of t-closeness is that it interrupts attribute expose. The issue lies in t-closeness is that as size and variability of data increases, the odds of re-identification too increases.

3.3. Randomization

Randomization is the process of addition noise to the data which is usually done by probability distribution [21]. Randomization is applied in surveys, sentiment analysis etc. Randomization does not required knowledge of other records in the data. It can be functional during data collection and pre-processing time. There is no anonymization up above in randomization. However, put on randomization on large datasets is not probable because of time complexity (P. Ram Mohan Rao, 2018).

3.4. Homomorphic encryption

In (Sangeetha,M, 2014), a homomorphic technique is developed which is basically a form of encryption that allows performing some specific computations on ciphertext and encrypted results are obtained. The decrypted results are then matched to the results of operations that are performed on plain text. This approach is useful to deal with the entrusted party because neither the input is unveiled nor the internal state of the encrypted data.

3.5. Data distribution

Data distribution method, the data is distributed across many sites. Distribution of the data can be done in two ways:

i. Horizontal distribution of data
ii. Vertical distribution of data

Horizontal distribution Data is distributed across many sites with same attributes, the distribution is said to be horizontal distribution which is described in Fig. 1. Horizontal distribution of data can be functional only when some aggregate functions or operations are to be applied on the data without actually sharing the data. For example, if a car dealer showroom wants to analyse their sales across branches, they may hire some analytics which does computations on aggregate data. Data holder may share the data with third party analyst which may prompt to privacy breach. Classification and Clustering algorithms does not ensure privacy while it is applied on distributed data.

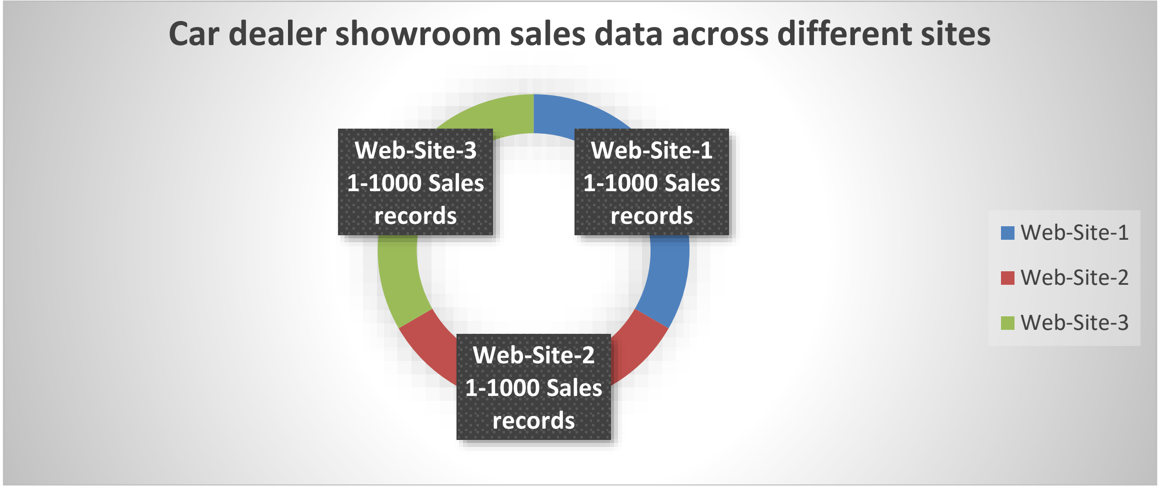

Fig. 1 Distribution of Car dealer showroom sales data across different web-sites

Vertical distribution of data Map and reduce tasks are performed in the public cloud using public data as the input, shuffle in-between data amongst them, and store the result in the public cloud. This work is done in the private cloud with private data. The jobs are processed in isolation as shown in Fig. 2. For example, Insurance organizations, the officials would like to know details of a particular persons which include health, profession, financial, personal etc. All this information may not be available at one site. This type of distribution is called vertical distribution in which each site holds few set of attributes of a person. When some analytics has to be done data has to be shared in from all these sites and there is a vulnerability of privacy breach.

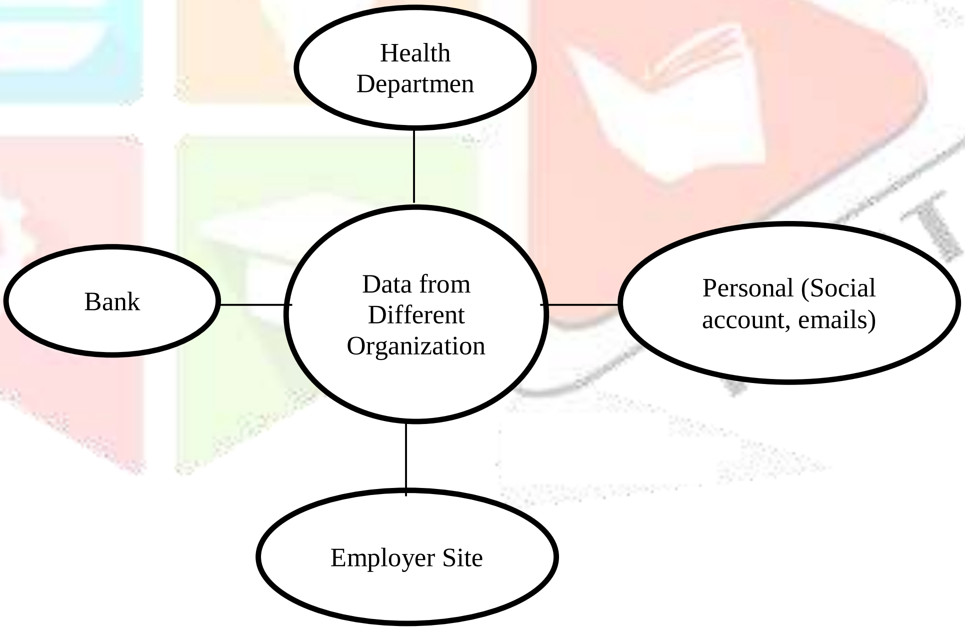
Health Departmen
Bank
Data from Different Organization
Personal (Social account, emails)
Employer Site

Fig. 2 Vertical distribution of person specific data

3.6. Cryptographic techniques

The data holder may encrypt the data before discharging the same for analytics. Such process of encrypting large data using old encryption methods is highly complex and must be functional only during data gathering time. Differential privacy techniques have already been applied where some aggregate calculations on the data are done without actually sharing the inputs. For example, if a and b are two data items then a function F (a, b) will be computed to gain some aggregate info from both x and y without actually sharing a and b. This can be applied on when a and b are held with various parties as in the case of vertical distribution. However, if the data is at single location under the guardian of a single enterprise, then differential privacy cannot be employed. Another similar technique called secure multiparty computation has been used but showed to be insufficient in privacy preservation. Data utility will be less if encryption is used during data analytics. Thus encryption is not only difficult to implement but it reduces the data utility (Jiang R, 2018).

3.7. Multidimensional Sensitivity Based Anonymization (MDSBA)

Bottom up Generalization (Wang K, 2004) and Top down Generalization (Fung BCM, 2005) are the old methods of Anonymization which were useful on well signified structured data records. However, applying the same on large scale data sets is very complex principal to issues of scalability and information loss. Multidimensional Sensitivity Based Anonymization is an improved version of Anonymization and more effective than old Anonymization method. MDSBA is an enhanced Anonymization (Zhang X, 2013) method such that it can be functional on large data sets with decreased loss of information and predefined quasi identifiers. As part of this technique Apache MAP REDUCE (Zhang X, 2014) framework has been used to handle large data sets. In conventional Hadoop Distributed Files System, the data will be separated into blocks of either 64 MB or 128 MB each and distributed across various nodes without bearing in mind the data inside the blocks. As part of Multidimensional Sensitivity Based Anonymization (Al-Zobbi M, 2017) technique the data is split into different bags based on the probability distribution of the quasi identifiers by making use of filters in Apache Pig scripting language.

IV. PRIVACY TECHNIQUES CHALLENGES

Challenges faced by various privacy techniques are delineated in Table 4. The analysis results in that no single method is reliable in all spheres. Each method performs in a different way depending on the size of data and the type of application.

Table 4.1: Privacy techniques and challenges

Techniques	Challenges
K-anonymity	Gives no consideration of the links between sensitive data Not able to protect against attacks based on background knowledge Not applicable for high-dimensional data
L diversity	Modeling background knowledge of adversaries and attacks about social network data is much more challenging than that about relational data. Measuring information loss in anonymizing social network data is much more challenging than that in anonymizing relational data.
Randomization	Does not reconstruct the original data values
Cryptographic technique	Difficult to apply for large databases Difficult to scale when more events are involved Non-sensitive data is also encrypted that can be useful for analytics
Homomorphic encryption	Computational overhead increased Not applicable for large datasets
Multidimensional Sensitivity Based Anonymization (MDSBA)	Suitable for large scale data but only when the data is at rest. It cannot be applied for streaming data.

V. COMPARISON OF DIFFERENT PRIVACY TECHNIQUES

Table 5 shows a comparative analysis of some of the privacy-preserving techniques based on parameters linkage property, information loss, type of data, and privacy preserved.

Table 5.1: Comparison of different privacy techniques

Techniques	Parameters			
	Linkage Property	Information Loss	Type of data	Privacy Preserved
K-anonymity	High	Low	Micro Data	Low
L diversity	High	Low	Micro Data	Low
Randomization	Low	High	High dimensional	High
Cryptographic technique	Low	Low	Micro data	High
Homomorphic encryption	Low	Low	Micro data	High
Multidimensional Sensitivity Based Anonymization (MDSBA)	High	Low	High dimensional	High

VI. CONCLUSION AND FUTURE WORK

No concrete result for unstructured data has been established yet. Conventional data algorithms can be applied for classification and clustering problems but cannot be used in privacy preservation especially when dealing with person specific information. Machine learning and soft computing techniques can be used to develop new and more suitable solution to privacy problems which contain individuality disclosure that can lead to personal embarrassment and abuse.

There is a strong need for law implementation by governments of all countries to ensure individual privacy. European Union (TCS white paper, 2016) is making an attempt to enforce privacy preservation law. One of the main privacy threats is smart phone. Personal information like contacts, messages, chats and files are being accessed by many apps running in our smart phone without our knowledge. People needs to read privacy statement before installing any application. There is a strong need to educate people on the various liabilities which can donate to leakage of private information.

In big data it is not possible to carry out the operations without compromising the privacy. Business organizations hold sensitive information about their clients and this information is considered as a big asset to them. To safeguard this information against unauthorized access, few techniques are proposed in the literature but have limitations. So, the authors believe that more such techniques and mechanism need to be developed that will help in preserving privacy during data analysis process, for the reason that if privacy about an individual is violated it may have catastrophic significance on someone’s life.

REFERENCES

[1] Abadi DJ, Carney D, Cetintemel U, Cherniack M, Convey C, Lee S, Stone-braker M, Tatbul N, Zdonik SB. Aurora: a new model and architecture for data stream manag ement. VLDB J. 2003;12(2):120–39.

[2] Kolomvatsos K, Anagnostopoulos C, Hadjiefthymiades S. An efficient time optimized scheme for progressive analytics in big data. Big Data Res. 2015;2(4):155–65.

[3] Big data at the speed of business, [online]. http://www-01.ibm.com/soft-ware/data/bigdata/2012.

[4] Ducange Pietro, Pecori Riccardo, Mezzina Paolo. A glimpse on big data analytics in the framework of marketing strategies. Soft Comput. 2018;22(1):325–42.

[5] Chauhan Arun, Kummamuru Krishna, Toshniwal Durga. Prediction of places of visit using tweets. Knowl Inf Syst. 2017;50(1):145–66.

[6] Porambage P, et al. The quest for privacy in the internet of things. IEEE Cloud Comp. 2016;3(2):36–45.

[7] Iyengar S. Transforming data to satisfy privacy constraints. In: Proceedings of the eighth ACM SIGKDD international conference on knowledge discovery and data mining. New York: ACM; 2002.

[8] LeFevre K, DeWitt DJ, Ramakrishnan R. Incognito: efficient full-domain k-anonymity. In: Proceedings of the 2005 ACM SIGMOD international conference on management of data. New York: ACM; 2005.

[9] LeFevre K, DeWitt DJ, Ramakrishnan R. Mondrian multidimensional k-anonymity. In: Proceedings of the 22nd international conference (ICDE'06) on data engineering, 2006. New York: ACM; 2006.

[10] Samarati P, Sweeney L. Protecting privacy when disclosing information: k-anonymity and its enforcement through generalization and suppression. Technical Report SRI-CSL-98-04, SRI Computer Science Laboratory; 1998.

[11] Sweeney L. K-anonymity: a model for protecting privacy. Int J Uncertain Fuzz. 2002;10(5):557–70

[12] Li N, et al. t-Closeness: privacy beyond *k*-anonymity and *L*-diversity. In: Data engineering (ICDE) IEEE 23rd international conference; 2007.

[13] Samarati P, Sweeney L. Protecting privacy when disclosing information: k-anonymity and its enforcement through generalization and suppression. Technical Report SRI-CSL-98-04, SRI Computer Science Laboratory; 1998.

[14] P. Ram Mohan Rao, S. Murali Krishna and A. P. Siva Kumar. Privacy preservation techniques in big data analytics: a survey, Journal of Big data (Springer); 2018

[15] Sangeetha,M., Anishprabu, P.,&Shanmathi, S. Homomorphic Encryption Schema for Privacy Preserving Mining of Association Rules. *International Journal of Innovation Research Science Engineering.*

[16] Jiang R, Lu R, Choo KK. Achieving high performance and privacy-preserving query over encrypted multidimensional big metering data. Future Gen Comput Syst. 2018;78:392–401.

[17] Wang K, Yu PS, Chakraborty S. Bottom-up generalization: A data mining solution to privacy protection. In: Fourth IEEE international conference on data mining, 2004 (ICDM'04). Piscataway: IEEE; 2004.

[18] Fung BCM, Wang K, Yu PS. Top-down specialization for information and privacy preservation. In: Proceedings 21st international conference on data engineering, 2005 (ICDE 2005). Piscataway: IEEE; 2005.

[19] Zhang X et al. A MapReduce based approach of scalable multidimensional anonymization for big data privacy preservation on cloud. In: Third international conference on cloud and green computing (CGC), 2013. Piscataway:IEEE; 2013.

[20] Zhang X, et al. A scalable two-phase top-down specialization approach for data anonymization using mapreduce on cloud. IEEE Trans Parallel Distrib Syst. 2014;25(2):363–73.

[21] Al-Zobbi M, Shahrestani S, Ruan C. Improving MapReduce privacy by implementing multi-dimensional sensitivitybased anonymization. J Big Data. 2017;4(1):45.

[22] TCS. Emphasizing the need for government regulations on data privacy; 2016. https ://www.tcs.com/conte nt/dam/ tcs/pdf/techn ologi es/Cyber -Secur ity/Abstr act/Stren gthen ing-Priva cy-Prote ction -with-the-Europ ean-Gener al-Data- Prote ction -Regul ation .pdf.