# Security Enhancement and Privacy Preserving Of Big Data

**Taran Singh Bharati**

**Abstract**: ***Big data is gaining the popularity among the data scientist and the people from the Biology related disciplines. Big Data is very huge in volume and comes very fast from various sources. Millions of tweets or posts are generated per second on social networking sites. Big data has many issues of i.e. nature, storing, management, and processing, privacy and security in disclosing of attributes of the sensitive data in data of healthcare etc. For maintaining the privacy there are k-anonymity, p-sensitive k-anonymity, l-diversity, t-closeness, and k-concealment ways. In this paper an anonymity algorithm is proposed which will be used to enhance the security and privacy preserving of sensitive attributes of big data.***

***Index Terms: Privacy, Big Data, Security, Attacks***

## I. INTRODUCTION

Big data is understood by five parameters velocity, variety, volume, veracity, and value. A data is called big if it is huge in volume (expressed in TB or PB) of different kinds and generated at very fast speed. Data that belongs to big data is mixed type of some structured some unstructured, some graph data, streaming data, and some semi-structured. It is said that around 80 percent big data is unstructured data. So it cannot be dealt by ordinary DBMSs therefore we need specialized tools like Hadoop or R language for big data. Transparency, identity, and power are three paradoxes of big data [18], [23]. The analysis of big data is called big data analytics; a new term is generated since 2010 which could be descriptive, predictive, diagnostic, and prescriptive analysis.There are many technologies which assist the big data in its organization and in its working i.e. Hadoop, Hive, distributed file, NoSQL, in memory fabric, search and knowledge discovery system, HDFS, data integration, Pig , data virtualization, cloud computing, IoT, sqoop, polybase, presto, etc [22]. Big data architecture involves a real-time processing, batch processing, interactive exploration, predictive analytics and machine learning workload. Big data has data driven, lamda, and kappa architectures.

Security means protecting resources and it is implemented by its services. Authentication, Access control, integrity, confidentiality, availability and non-repudiation are the services of security. For security services to avail there are some mechanisms i.e. encryption, digital signature, message authentication code, SHA, MD5 etc. algorithms. Unauthorized sniffing or access to system is called attack from unauthorized access. Sometimes these attacks are less dangerous (passive) or sometime they become more dangerous (active). Threats are risk which could be exploited by adversary anytime [24], [25], [26], [27], [28], [29], [44], [45]. One of the goals of security is to detect and remove the attacks if they infected the system. The better countermeasures of attacks and malicious software are to use or deploy the anti-virus, firewall, or complete protection of the system. Privacy is associated with the personal information revealing of the persons. Personal information is very confidential and should neither be disclosed nor be used without the consent of person.

Data received from social networking sites, sensors' data, data from nuclear reactions, data received from satellite, metrological sites, is example of big data. It is very much useful in many places i.e. science, industry, health, space, ATC, weather forecasting [7]. Privacy is the right of individuals and there are protection laws which express the fair information practices (FIPs) [8], [15]. Security is required in all operations so that data can remain confidential, integral authentic, rightly accessed.

Data sources are shared and integrated by the people and data is precious therefore the data should not be tampered, leaked, or misused. There should not be any threat or malicious attacks on it. Security of big data through quantum cryptography for authentication is also proposed [16].

Attacks like background attack, homogeneity attacks, temporal attacks, complementary release attacks, unsorted matching attacks etc. are also recorded [34].

## II. RELATED WORK

According to cloud secure alliance (CSA) challenges of privacy are placed into four parts [9]. On cloud, privacy is maintained by storage path, attribute based encryption, holomorphic encryption, access control, and hybrid clouds [20]. To maintain the integrity verification, user has to make sure that he is using the services according to the contract signed. In cloud computing environment few security and privacy challenge categories [1], [2] at different levels i.e. network, authentication, data, generic levels etc. There are some issues and challenges of security and privacy of big data i.e. data preparation; effective online analysis, semantic techniques, and handling of data streams; social analytics, value based governance healthcare analytics; In distributed programming framework secure computations; security best practices; data storage and transmission logs ; end-point filtering; compliance monitoring; secure communication, security and privacy and cryptographically enforced access control; granular access control [3], [4], [5], [6], [9], [10], [12], [19].The concerned to security and privacy of big data are: i) Device manufacturers; ii) Consumers and non-consumers; iii) Government and regulatory body; iv) Third party application developers; v) IoT cloud services and platform providers [13], [14]. The techniques which to protect the big data are such as logging, secure communication, file encryption, key management, masking, access control, and server authentication protocols [30] [31], [36], [11]. The De-identification, t-closeness, k-anonymity, and l-diversity are employed privacy preserving methods. Obvious parameters of privacy fields are identifier attributes. Multiple receiver updates and

**Revised Manuscript Received on August 05, 2019**
**Taran Singh Bharati,** Department of Computer Science**,** Jamia Millia Islamia New Delhi, India.

conditional sharing are also available in which gaming and collision models are developed to thwart the attacks [19]. The t-closenes, l-diversity, k-anonymity become fail at one point then we should go for ($p^+$, α)-sensitive k- anonymity model [32], [33], [42]. Some information may be lost due to generalization which is called cost function [43].

## III. METHODOLOGY

Suppose we use a dummy big data of a health care as shown below whose sensitive information needs to be protected.

**Table1: Micro Row Data**

| | Non-Sensitive Attributes | | | Sensitive Attribute |
|---|---|---|---|---|
| Sr No | Pin Code | Age | Nationality | Disease |
| 1 | 93053 | 29 | Russian | Heart |
| 2 | 93068 | 27 | American | Heart |
| 3 | 93068 | 26 | Japanese | Viral |
| 4 | 93053 | 23 | American | Viral |
| 5 | 94853 | 53 | Indian | Cancer |
| 6 | 94853 | 54 | Russian | Cancer |
| 7 | 93050 | 48 | American | Viral |
| 8 | 93050 | 47 | American | Viral |
| 9 | 93053 | 33 | American | Flu |
| 10 | 93053 | 35 | Indian | Flu |
| 11 | 93068 | 37 | Japanese | Asthma |
| 12 | 93068 | 38 | American | Indigestion |

The initial data called the micro data has the several attribute types:

- Private Attributes: Confidential which keep sensitive personal information which is always supposed to be undisclosed.
- Quasi-Attributes: Which directly do not contain any private personal information that may provide the linkage to identify the people?
- Non-Sensitive Attributes: Which do not disclose the identity of the people?

### A. *ANONYMIZATION*

It is the technique which is used to spread anonymity in the records to hide the sensitive attributes. The k-anonymity property is defined as for every combination of key attribute values, occurring in micro data k times or more. A p-sensitive anonymity property is said to prevail in masked micro data (MM) if for all clusters of records of identical bunch of key attribute values, values of private attribute appear at least p times.

To protect personal information, anonymization is needed and which is done as shown in table 2. This can be achieved with the help of generalization and suppression (figure 1) [17], [21], [35], [39], [40], [41].

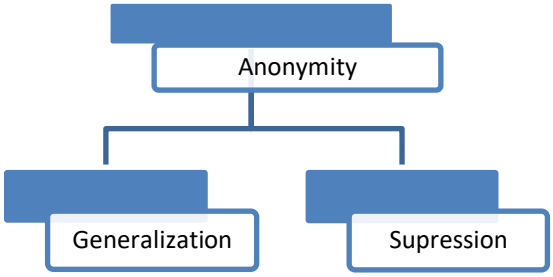



**Figure 1: Generalization and Suppression**

- **Generalization:** It is used to hide the privacy by making a range instead of actual values as shown in table 2.
- **Suppression:** This is used by hiding or putting some unknown (*) character as values in columns.

The attributes are partitioned to be prevented from the attribute disclosure, into different disjoint ordered categories on the basis of the severity level of sensitivity of the attributes. There are three fundamental issues of transportation and storage; for large quantity of data and management; because of variety, veracity, and value it is very difficult to manage big data (figure 2).

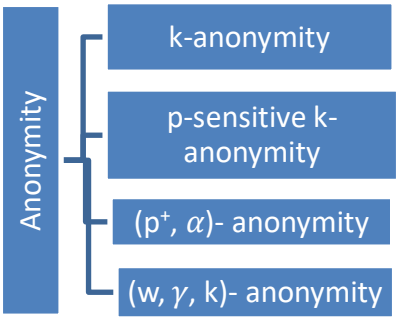



**Figure 2: Anonymization**

Bu above principle the above table 1 can be converted into table 2. In (w, $\gamma$, k) - anonymity the average weight of a partition must be at least equal to w and similarity must not be more than $\gamma$.In raw micro data ($p^+$, $\alpha$) is used if it is k-anonymous where each partition has p different categories and its weight should not exceed $\alpha$.

**Table 2: Micro Data With 4-Anonymity**

| | Non-Sensitive Attributes | | | Sensitive Attribute |
|---|---|---|---|---|
| Sr No | Pin Code | Age | Nationality | Disease |
| 1 | 930** | <35 | * | Heart |
| 2 | 930** | <35 | * | Heart |
| 3 | 930** | <35 | * | Viral |
| 4 | 930** | <35 | * | Viral |
| 5 | 948** | ≥ 45 | * | Cancer |
| 6 | 948** | ≥ 45 | * | Cancer |
| 7 | 930** | ≥ 45 | * | Viral |

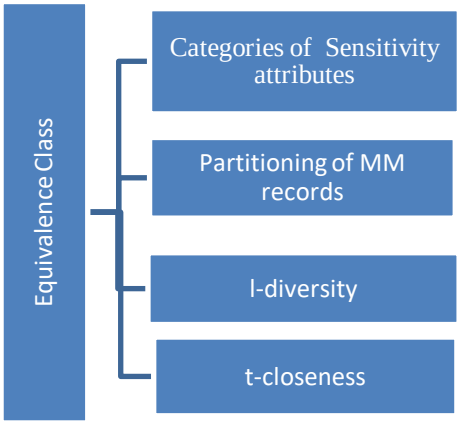
| | | | | |
|---|---|---|---|---|
| 8 | 930** | ≥ 45 | * | Viral |
| 9 | 930** | 3* | * | Flu |
| 10 | 930** | 3* | * | Flu |
| 11 | 930** | 3* | * | Asthma |
| 12 | 930** | 3* | * | Indigestion |

The anonymity table is partitioned into clusters / equivalence class on the basis of similarity in record of table. Every record has some distance from other records. For example table1 can be grouped into different categories as below (Table 3):

**Table 3: Partitioned Table**

| Level | Sensitive attributes | Weight |
|---|---|---|
| 1 | Heart, Cancer | 0 |
| 2 | Asthma | 1 |
| 3 | Indigestion | 2 |
| 4 | Viral, Flu | 3 |

Above we partitioned the attributes to be prevented from the attribute disclosure, into different disjoint ordered categories on the basis of the severity level of sensitivity of the attributes [37].




### B. ALGORITHM

This algorithm lets us find the partitions of fixed similarity and fixed p and k anonymity which enhances the privacy of sensitive attributes in big data which in turn makes difficult to attacker to guess the sensitive attributes of some individuals.

*Given: Micro data, domain of sensitive attributes= ($S_1$,$S_2$, ...$S_n$), α,* p.

1. microdata must be k-anonymized
2. return true
3. compute the weights of sensitive attributes
4. for all groups in microdata
5. Let d be distance and α' be the total weight
6. if p > d or α′ < α
   return false
7. else
   break loop
   return true
8. return false

## IV. RESULT ANALYSIS AND DISCUSSIONS

### A. FOR ANONYMITY ANALYSIS

In generalization process some information is lost which is called cost function [43]. The k-concealment is a secure form of the k-anonymity therefore every k-anonymity model is k-concealment while revers may not be true.
Equivalence class is the records' set which has the same values of the quasi attributes. There are different types of l-diversities:

i) Distinct l-diversity: All equivalence classes have at least l-diversity.
ii) Entropy l-diversity: Equivalence class E is said to be entropy of l-diversity, if

$$E \geq \log l.$$

Entropy l-diversity is more powerful than l-diversity. Entropy of any class is specified as:

$$\text{Entropy (E)} = -\sum_{s \epsilon S} p(E) \log p(E, s)$$

where S is the domain of sensitive attributes and p (E,s) record fraction containing sensitive attribute s. The l-diversity is difficult and redundant and is also not enough to prevent attribute disclosure attacks.

iii) Recursive (c, l)-diversity: It is used to check that the most frequent and least frequent values do not occur most/least frequently.

t-closeness: The closeness in distributions is describe by the distance between the distributions. Suppose P = ($a_1$, $a_2$, $a_3$,.., $a_m$) and Q = ($b_1$, $b_2$, $b_3$,.., $b_m$) are two distributions, their variational distance can be specified as:

$$|P, Q| = \sum_{i=1}^{m} \frac{1}{2} |a_i - b_i|$$, and Kullback- Leibler (KL) distance

$$|P, Q| = \sum_{i=1}^{m} a_i \log \frac{a_i}{b_i} = H(P) - H(P, Q)$$

where H (P) is the entropy of P distribution and H (P, Q) is equivocation of the distributions Q and P. Note that k-anonymity protects the identity (attribute) disclosure.Distance between two attributes d ($S_1$, $S_2$) is the distance in categories in which these two sensitive attributes fall.

$$\text{Weight of any attribute w } (S_i) = \begin{cases} 0, & if\ i = 1 \\ \frac{i-1}{m-1}, & if\ 1 < i < m \\ 1, & if\ i = m \end{cases}$$

### B. FOR ALGORITHM

Grouping is done on the basis of l-diversity and t-closeness. The complexity of making groups of a table having n records is:

$$\sum_{i=1}^{i=n} n_{C_i} = 2^n$$

### C. FOR SECURITY ANALYSIS

**Attacks on Anonymity:** Two types of attribute disclosure attacks [34], [38] exist as below:

- Homogeneity Attack: If two persons are neighbours when one goes to hospital by ambulance. Since neighbour knows many personal details like pin code, age, etc. therefore he can get record and know the sensitive attributes of disease.
- Back ground attack: If person is familiar with patient's background like pin code, age, he can guess the group of records identifying sensitive attributes.

## V. CONCLUSIONS

Big data in recent days have become more needy and popular specially in the fields of bioinformatics and data sciences. The data of healthcare may contain many private or sensitive attributes which need to be de-identified. For de-identifying the sensitive private attributes, anonymity with some precision is used in the form of generalization and suppression on quasi-identifier attributes. There are some attacks for re-identification which must be protected. Data is collected from many heterogeneous sources and then it is shared among the people. Among the so many issues of big data this paper focused on security and privacy issues. This paper employed k, p, ($p^+$, $\alpha$)-anonymity to preserve the identity of individuals and proposed the algorithm for anonymizing the sensitive attributes.

## REFERENCES


1. Inukollu VN, Arsi S, Ravuri SR. Security issues associated with big data in cloud computing. International Journal of Network Security & Its Applications. 2014 May 1;6 3):45.
2. Terzi DS, Terzi R, Sagiroglu S. A survey on security and privacy issues in big data. In2015 10th International Conference for Internet Technology and Secured Transactions (ICITST) 2015 Dec 14 (pp. 202-207). IEEE.
3. Mora AC, Chen Y, Fuchs A, Lane A, Lu R, Manadhata P. Top ten big data security and privacy challenges. Cloud Security Alliance. 2012; 140.
4. Kaisler S, Armour F, Espinosa JA, Money W. Big data: Issues and challenges moving forward. In2013 46th Hawaii International Conference on System Sciences 2013 Jan 7 (pp. 995-1004). IEEE.
5. Matturdi B, Zhou X, Li S, Lin F. Big Data security and privacy: A review. China Communications. 2014 Apr; 11 (14):135-45.
6. Patil HK, Seshadri R. Big data security and privacy issues in healthcare. In2014 IEEE international congress on big data 2014 Jun 27 (pp. 762-765). IEEE.
7. Saurabh pandey, rashmi pandey, Medical (Healthcare) Big Data Security and Privacy Issues, International Journal of Scientific & Engineering Research Volume 9, Issue 2, feb-2018.
8. Cavoukian A, Jonas J. Privacy by Design in the Age of Big Data. In Guide to Big Data Applications 2012 (pp. 29-48). Springer, Cham. Web site: www.ipc.on.ca, Privacy by Design: www.privacybydesign.ca
9. Moura J, Serrão C. Security and privacy issues of big data. In Handbook of research on trends and future directions in big data and web intelligence 2015 (pp. 20-52). IGI Global.
10. Lu R. Zhu H., Liu X. et al., towards efficient and privacy-preserving computing in Big Data era, IEEE Network, july-august, 2014 pp. 46-50.
11. Gahi Y, Guennoun M, Mouftah HT. Big data analytics: Security and privacy challenges. In2016 IEEE Symposium on Computers and Communication (ISCC) 2016 Jun 27 (pp. 952-957). IEEE.
12. Alguliyev R, Imamverdiyev Y. Big data: big promises for information security. In2014 IEEE 8th International Conference on Application of Information and Communication Technologies (AICT) 2014 Oct 15 (pp. 1-4). IEEE.
13. Jain P, Gyanchandani M, Khare N. Big data privacy: a technological perspective and review. Journal of Big Data. 2016 Dec; 3(1):25.
14. Perera C, Ranjan R, Wang L, Khan SU, Zomaya AY. Big data privacy in the internet of things era. IT Professional. 2015 May; 17 (3):32-9.
15. Zhang D. Big data security and privacy protection. In8th International Conference on Management and Computer Science (ICMCS 2018) 2018 Oct 20. Atlantis Press.
16. Thayananthan V, Albeshri A. Big data security issues based on quantum cryptography and privacy with authentication for mobile data center. Procedia Computer Science. 2015 Jan 1; 50: 149-56.
17. Yu S. Big privacy: Challenges and opportunities of privacy study in the age of big data. IEEE access. 2016; 4: 2751-63.
18. Demchenko Y, Ngo C, de Laat C, Membrey P, Gordijenko D. Big security for big data: Addressing security challenges for the big data infrastructure. In Workshop on Secure Data Management 2013 Aug 30 (pp. 76-94). Springer, Cham.
19. Hu J, Vasilakos AV. Energy big data analytics and security: challenges and opportunities. IEEE Transactions on Smart Grid. 2016 Sep; 7 (5):2423-36.
20. Mehmood A, Natgunanathan I, Xiang Y, Hua G, Guo S. Protection of big data privacy. IEEE access. 2016; 4:1821-34.
21. Gadepally V, Hancock B, Kaiser B, Kepner J, Michaleas P, Varia M, Yerukhimovich A. Computing on masked data to improve the security of big data. In2015 IEEE International Symposium on Technologies for Homeland Security (HST) 2015 Apr 14 (pp. 1-6). IEEE.
22. Sharma PP, Navdeti CP. Securing big data hadoop: a review of security issues, threats and solution. Int. J. Comput. Sci. Inf. Technol. 2014; 5 (2):2126-31.
23. Richards NM, King JH. Three paradoxes of big data. Stan. L. Rev. Online. 2013; 66: 41.
24. Bharati, T. S. (2015). Enhanced Intrusion Detection System for Mobile Adhoc Networks using Mobile Agents with no Manager. International Journal of Computer Applications, 111(10).
25. Bharati, T. S., & Kumar, R. (2015, March). Secure intrusion detection system for mobile adhoc networks. In Computing for Sustainable Global Development (INDIACom), 2015 2nd International Conference on (pp. 1257-1261). IEEE.
26. Bharati, T. S., & Kumar, R. (2015). Intrusion Detection System for MANET using Machine Learning and State Transition Analysis. International Journal of Computer Engineering & Technology (IJCET), 6(12), 1-8.
27. Bharati, T. S., & Kumar, R. (2016). Enhanced Key Distribution for Mobile Adhoc Networks. International Journal of Engineering Science, 6(4), 4184-4187.
28. Bharati T. S. (2017). Agents to Secure MANETS. International Journal of Advanced Engineering and Research Development, 4(11), 1267-1273.
29. Bharati T.S. (2018). MANETs and Its' Security. International Journal of Computer Networks and Wireless Communication, 8(4), 166-171.
30. Jaseena KU, David JM. Issues, challenges, and solutions: big data mining. CS & IT-CSCP. 2014 Dec 27; 4 (13):131-40.
31. Abouelmehdi K, Beni-Hessane A, Khaloufi H. Big healthcare data: preserving security and privacy. Journal of Big Data. 2018 Dec 1; 5(1):1.
32. Li N, Li T, Venkatasubramanian S. t-closeness: Privacy beyond k-anonymity and l-diversity. In2007 IEEE 23rd International Conference on Data Engineering 2007 Apr 15 (pp. 106-115). IEEE.
33. Sweeney L. Achieving k-anonymity privacy protection using generalization and suppression. International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems. 2002 Oct; 10 (05):571-88.
34. Sweeney L. k-anonymity: A model for protecting privacy. International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems. 2002 Oct; 10 (05):557-70.
35. Samarati P. Protecting respondents identities in microdata release. IEEE transactions on Knowledge and Data Engineering. 2001 Nov; 13 (6):1010-27.

36. Truta TM, Vinay B. Privacy protection: p-sensitive k-anonymity property. In22nd International Conference on Data Engineering Workshops (ICDEW'06) 2006 (pp. 94-94). IEEE.
37. Xiao X, Tao Y. Anatomy: Simple and effective privacy preservation. InProceedings of the 32nd international conference on Very large data bases 2006 Sep 1 (pp. 139-150). VLDB Endowment.
38. Xiao X, Tao Y. Personalized privacy preservation. InProceedings of the 2006 ACM SIGMOD international conference on Management of data 2006 Jun 27 (pp. 229-240). ACM.
39. Machanavajjhala A, Gehrke J, Kifer D, Venkitasubramaniam M. l-diversity: Privacy beyond k-anonymity. In22nd International Conference on Data Engineering (ICDE'06) 2006 Apr 3 (pp. 24-24). IEEE.
40. Iyengar VS. Transforming data to satisfy privacy constraints. InProceedings of the eighth ACM SIGKDD international conference on Knowledge discovery and data mining 2002 Jul 23 (pp. 279-288). ACM.
41. Bayardo RJ, Agrawal R. Data privacy through optimal k-anonymization. In21st International conference on data engineering (ICDE'05) 2005 Apr 5 (pp. 217-228). IEEE.
42. Sun X, Wang H, Li J, Truta TM, Li P. (p+, α)-sensitive k-anonymity: A new enhanced privacy protection model. In2008 8th IEEE International Conference on Computer and Information Technology 2008 Jul 8 (pp. 59-64). IEEE.
43. Tassa T, Mazza A, Gionis A. k-Concealment: An Alternative Model of k-Type Anonymity. Trans. Data Privacy. 2012 Apr 1; 5(1):189-222.
44. Bharati T.S. (2019). Trust Based Security of MANETs. International Journal of Innovative Technology and Exploring Engineering, 8(8), 792-795.
45. Bharati T.S. (2019). Security and Privacy of Internet of Things. International Journal of Innovative Technology and Exploring Engineering, 8(8), 2740-2743.

## AUTHOR PROFILE

Author has done B.Tech, Master of Engineering, and Ph.D. in Computer Science Stream from, Kanpur, Gwalior, and New Delhi respectively. He has around 18+ years of experience at time of writing this paper. He has served at different positions in various Universities and Engineering Colleges. His area of interests includes Security, IoT, Big Data, Theoretical Computer Science, Data Science etc.