Privacy Safeguarding techniques in Unstructured Big Data Analytics

[1]Prof.Suneetha V, HOD-MCA(BU), Dayananda Sagar Institutions, Bangalore.

[2]Dr.Y.S. Kumara Swamy, Research Supervisor, Bangalore.

Abstract

Voluminous amounts of data is being generated by various organizations like education, research, health, insurance, financial sectors, e-commerce, retail and supply chain, etc. by virtue of digital technology. Not only humans but machines also contribute to data in the form of streaming, web site logs, etc. The large amounts of data generated from the heterogeneous sources can be processed and analyzed to support decision making. However data analytics is prone to privacy violations. One of the applications of data analytics is recommendation systems which are widely used by ecommerce sites like Amazon, Flip kart, Snap deal for suggesting products to customers based on their buying habits leading to inference attacks. Although data analytics is useful in decision making, it will lead to serious privacy concerns. Big Data Analytics has made open doors for specialists to process immense measure of information yet made a major danger to protection of person. Existing security protecting methods like, anonymization requires having dataset separated in the arrangement of properties like sensitive attributes, quasi identifiers and non sensitive attributes. Hence privacy preserving data analytics became very important. This paper examines various security breaches, privacy preservation techniques and models with their limitations.

Keywords: Data, Data analytics, Privacy threats, Privacy preservation, Data Science.

Introduction

Today, we live in what many call the Information Age, and we are in absolutely no danger of running out of information, particularly in data form. There is a general perception that we are overwhelmed with data, making the ability to store, process, analyze, interpret, consume, and act upon that data a primary concern. For large-scale, multi-national organizations and those in heavily regulated industries— such as finance, healthcare, or those covering multiple industry verticals — the situation becomes even more complex and challenging. Escalating data concerns are rampant in the Internet of Things (IOT) Age, during which growth of data is exceeding the capacity of traditional computing. The question then becomes how do we consume those data sources and transform them into actionable information?

The large scale data, which also include person specific private and sensitive data like gender, pincode, disease, caste, shopping cart, religion etc. is being stored in public domain. The data holder can release this data to a third party data analyst to gain deeper insights and identify hidden patterns which are useful in making important decisions that may help in improving businesses, provide value added services to customers [1], prediction, forecasting and recommendation [2]. One of the prominent applications of data analytics is recommendation systems which are widely used by ecommerce sites like Amazon, Flip kart, Snap deal, Alibaba, Ebay for suggesting products to customers based on their buying habits. Facebook does suggest friends, places to visit and even movie recommendation based on our interest. However releasing user activity data may lead inference attacks like identifying geographical, religion, gender based on user activity [3]. We have studied a number of privacy preserving techniques which are being employed to protect against privacy threats.

Privacy threats in data analytics Privacy is the ability of an individual to determine what data can be shared and employ access control. If the data is in public domain then it is a threat to individual privacy as the data is held by data holder. Data holder can be social networking application, websites, mobile apps, ecommerce site, banks, hospitals etc. Privacy preservation is crucial in ubiquitous computing environment. Without this, users feel uneasy to use and live in the UC environment. The implementation of privacy safeguard or privacy enhancing technologies is going to be a long road. Understanding the challenges & issues of privacy protected in ubiquitous computing, is helpful to design and implement privacy aware system. It is the responsibility of the data holder to ensure privacy of the user's data. Apart from the data held in public domain, knowing or unknowingly users them self contribute to data leakage. For example most of the mobile apps, seek access to our contacts, files, camera etc. and without reading the privacy statement we

agree for all terms and conditions, there by contributing to data leakage. Hence there is a need to educate the smart phone users regarding privacy and privacy threats. Some of the key privacy threats include (1) Surveillance; (2) Disclosure; (3) Discrimination; (4) Personal embracement and abuse.

Surveillance : Many companies including wholesale and retail, e-commerce, etc. study their customers buying habits and try to come up with various offers and value added services [4]. Based on the opinion data, survey data, and sentiment analysis, social media sites does provide recommendations of the new friends, places to visit, people to follow etc. This is possible only when they continuously monitor their customer's transactions. This is a serious threat as no individual accepts surveillance.

Disclosure : Consider a hospital holding patient's data which include (pin, gender, age, disease) [5–7]. The data holder has released data to a third party for analysis by anonymizing sensitive person specific data so that the person cannot be identified. The third party data analyst can map this information with the freely available external data sources like census data and can identify person suffering with some disorder. Data (Candidate name, mobile number, date of birth) provided to entrance tests will be disseminated to various private agencies.

Discrimination is the bias or inequality which can happen when some private information of a person is disclosed. For instance, statistical analysis of electoral results proved that people of one community were completely against the party, which formed the government. Leaders can target the group.

Personal embracement and abuse whenever some private information of a person is disclosed, it can even lead to personal embracement or abuse. For example, a person was privately undergoing medication for some specific problem and was buying some medicines on a regular basis from a medical shop. As part of their regular business model, the medical shop may send some reminder and offers related to these medicines over phone. If any family member has noticed this, it will lead to personal embracement and even abuse [8]. Data analytics activity will affect data Privacy. Many countries are enforcing Privacy preservation laws. Lack of awareness is also a major reason for privacy attacks. For example many smart phones/ PDA / intelligent system users are not aware of the information that is stolen from their devices by many apps. Previous research shows only 19% of smart phone users are aware of privacy threats [9].

Privacy preservation methods Many Privacy preserving techniques were developed, but most of them are based on anonymization of data. The list of privacy preservation techniques is given below.

• K anonymity

• L diversity
• T closeness
• Randomization
• Data distribution
• Cryptographic techniques
• Multidimensional Sensitivity Based Anonymization (MDSBA).

k-anonymity has been treated with great interest as an anonymization technique ensuring privacy in big data when we are dealing with quasi identifier attributes. Despite the fact that many algorithms of *k*-anonymity have been proposed, most of them admit that the threshold *k* of *k*-anonymity has to be known before anonymizing the data set. Here, a novel way in applying *k*-anonymity for quasi identifier attributes is presented. It's a new algorithm called "*k*-anonymity without prior value of the threshold *k*".

K anonymity is applied on the patient data shown in Table 1. The table shows data before anonymization. K anonymity algorithm is applied with k value as 3 to ensure 3 indistinguishable records when an attempt is made to identify a particular person's data. K anonymity is applied on the two attributes viz. Zip and age shown in Table 1. The result of applying anonymization on Zip and age attributes is shown in Table 2.

Slno	Pincode	Age	Disease
1	515001	29	Diabetic Problem
2	515275	22	Diabetic Problem
3	500094	27	Diabetic problem
4	524369	43	Dermotology
5	524362	52	Diabetic Problem
6	524333	47	Kidney
7	522365	30	Diabetic problem
8	522466	36	Kidney
9	522236	32	Kidney

Table 2 After applying anonymization on Pin and age

Slno	Pin	Age	Disease
1	515**	2*	Diabetic Problem
2	515**	2*	Diabetic Problem
3	500**	2*	Diabetic problem
4	5243*	>40	Dermatology
5	5243*	>40	Diabetic Problem
6	5243*	>40	Kidney
7	522**	3*	Diabetic problem
8	522**	3*	Kidney
9	522**	3*	Kidney

Table 3 L diversity privacy preservation technique

Slno	pin	Age	Salary	Disease
1	515**	2*	5k	Diabetic Problem
2	515**	2*	6k	Diabetic Problem
3	515**	2*	7k	Diabetic Problem
4	5243*	>40	20k	Dermatology
5	5243*	>40	22k	Diabetic problem
6	522*	>40	24k	Kidney

The above technique has used generalization [14] to achieve Anonymization. Suppose if we know that John is 27 year old and lives in

500094 pincodes then we can conclude John to have Diabetic problem even after anonymization as shown in Table 2. This is called Homogeneity attack. For example if John is 36 year old and it is known that John does not have Kidney, then definitely John must have Diabetic problem. This is called as background knowledge attack. Achieving K anonymity [15, 16] can be done either by using generalization or suppression. K anonymity can optimized if the minimal generalization can be done without huge data loss [17]. Identity disclosure is the major privacy threat which cannot be guaranteed by K anonymity [18]. Personalized privacy is the most important aspect of individual privacy [19].

L-diversity was proposed to conquer the limitations of k- anonymity. As an extension to k-anonymity, they have introduced a novel method, which can ensure data privacy even without identifying the enemy's background knowledge to avoid attribute disclosure. This approach revolves around the notion that the sensitive attributes in each group are "well- represented". This technique is a modification of k-anonymity by incorporating the k- anonymity principle

When overall distribution of data is skewed into few equivalence classes attribute disclosure cannot be ensured. For example if the entire records are distributed into only three equivalence classes then semantic closeness of these values may lead to attribute disclosure. Also L diversity may lead to similarity attack.

From Table 3 it can be noticed that if we know that John is 27 year old and lives in 5243 zip, then definitely John is under low income group because salaries of all three persons in 515** zip is low compare to others in the table. This is called as similarity attack.

A betterment of l-diversity is a t-closeness technique by decreasing the granularity of the interpreted data. The observer's extent of knowledge on a specific data is limited while the knowledge is not limited to the overall table containing the datasets. Therefore, this reduces the correlation between the quasi-identifier attributes and the sensitive attributes. From Table 4 it can be observed that if we know John is 27 year old, still it will be difficult to estimate whether John has Diabetic problem or not and he is under low income group or not. T closeness may ensure attribute disclosure but implementing T closeness may not give proper distribution of data every time.

Randomization technique Randomization is the process of adding noise to the data which is generally done by probability distribution [21]. Randomization is applied in surveys, sentiment analysis etc. Randomization does not need knowledge of other records in the data. It can be applied during data collection and pre processing time. There is no anonymization overhead in randomization. However, applying randomization on large datasets is not possible because of time complexity and data utility which has been proved in our experiment described below. We have loaded 10k records

from an employee database into Hadoop Distributed File System and processed them by executing a Map Reduce Job.

• More number of Mappers and Reducers were used as data volume increased. • Results before and after randomization was significantly different. • Some of the records which are outliers remain unaffected with randomization and are vulnerable to adversary attack. • Privacy preservation at the cost of data utility is not appreciated and hence randomization may not be suitable for privacy preservation especially attribute disclosure.

Table4 T closeness privacy preservation technique

Slno	Pin	Age	Salary	Disease
1	515**	2*	5k	Diabetic Problem
2	515**	2*	16k	Kidney
3	515**	2*	9k	Dermatology
4	5243*	>40	20k	Dermatology
5	5243*	>40	42k	Diabetic problem
6	5243*	>40	8k	Flu

Data distribution technique

In this technique, the data is distributed across many sites. Distribution of the data can be done in two ways:

i. Horizontal distribution of data

ii. ii. Vertical distribution of data

Horizontal distribution when data is distributed across many sites with same attributes then the distribution is said to be horizontal distribution which is described. 1. Horizontal distribution of information can be connected just when some total capacities or tasks are to be connected on the information without really sharing the information. For instance, if a retail location needs to break down their deals crosswise over different branches, they may utilize some investigation which does calculations on total information. In the event that the information is disseminated crosswise over various destinations which have a place with various associations, at that point consequences of total capacities may help one gathering in identifying the information held with different gatherings.

I. If the data is distributed across different sites which belong to different organizations, then results of aggregate functions may help one party in detecting the data held with other parties. In such situations we expect all participating sites to be honest with each other [21]. Vertical distribution of data When Person specific information is distributed across different sites under custodian of different organizations, then the distribution is called vertical distribution as shown in Fig.1. For example, in crime investigations, the police officials would like to know details of a particular criminal which include health, wealth, hobbies, personal etc. All this information may not be available at one site. Such a distribution is called vertical distribution where each site holds few set of attributes of a person. When some analytics has to be done data has to be pooled in

from all these sites and there is a vulnerability of privacy breach. In order to perform data analytics on vertically distributed data, where the attributes are distributed across different sites under custodian of different parties, it is highly difficult to ensure privacy if the datasets are shared.

Cryptographic techniques

We describe here results of a body of cryptographic research that shows how separate parties can jointly compute any function of their inputs, without revealing any other information. As we argued above, these results achieve maximal privacy that hides all information except for the designated output of the function. This body of research attempts to model the world in a way which is both realistic and general. While there are some aspects of the "real world" that are not modeled by this research, the privacy guarantees and the generality of the results are quite remarkable. The information holder may encode the information before discharging the equivalent for investigation. Encoding vast scale information utilizing traditional encryption procedures is very troublesome and must be connected just amid information gathering time.

There are three analytics related parties in Big Data; these are the data owner, the service provider, and the user miner or analyzer. Based on the previously mentioned privacy methods, we introduce a Multidimensional Sensitivity-Based Anonymization framework (MDSBA) that implements a bottom-up technique of k-anonymity. The framework, also, adapts a discriminated multi-access level for users. The framework aims to implement a complete solution for MapReduce operations in big data. The solution basis mimics the parallel distributed processes over MapReduce nodes. This divides the single rigorous anonymization process into multi-tasks that can be distributed on more than one node. Accessing data for analytics is conducted by many users with multi- level access in the big data environment. This imposes a gradual level of the data access and view. Users with a low-level permission are less trusted by data owners. Therefore, more restrictions are applied to a data view.

Fig 1

Multidimensional Sensitivity Based Anonymization is an improved Anonymization technique such that it can be applied on large data sets with reduced loss of information and predefined quasi identifiers. As part of this technique Apache MAP REDUCE framework has been used to handle large data sets. In conventional Hadoop Distributed Files System, the data will be divided into blocks of either 64 MB or 128 MB each and distributed across different nodes without considering the data inside the blocks. As part of Multidimensional Sensitivity Based Anonymization 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. Data distribution was made effectively when compared to conventional method of blocks. Data Anonymization was done using four quasi identifiers using Apache Pig. Since the data is vertically partitioned into different groups, it can protect from background knowledge attack if the bag contains only few attributes. This method also makes it difficult to map the data with external sources to disclose any person specific information. Analysis various privacy preservation techniques have been studied with respect to features including, type of data, data utility, attribute preservation and complexity. The comparison of various privacy preservation techniques is shown in Table 5.

Table 5 Comparison of privacy preservation techniques

Features	**Privacy preservation techniques**				
	Anonymization techniques	Cryptographic techniques	Data distribution	Randomization	MDSBA
Suitability of unstructured data	No	No	No	No	Yes
Attribute preservation	No	No	No	Yes	Yes
Damage to data utility	No	No	Yes	No	Yes
Very complex to apply	No	Yes	Yes	Yes	Yes
Accuracy of results of data analytics	No	Yes	No	No	No

Results and discussions as part of systematic literature review; it has been observed that all existing mechanisms of privacy preservation are with respect to structured data. More than 80% of data being generated today is unstructured. As such, there is a need to address following challenges.

i. Develop concrete solution to protect privacy in both structured and unstructured data.

ii. Scalable and robust techniques to be developed to handle large scale heterogeneous data sets.

iii. Data should be allowed to stay in its native form without need for transformation and data analytics can be carried out while ensuring privacy preservation.

iv. New techniques apart from Anonymization must be developed to ensure protection against key privacy threats which include identity disclosure, discrimination, surveillance etc.

v. Maximizing data utility while ensuring data privacy.

Conclusion

No concrete solution for unstructured data has been developed yet. Conventional data mining 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 appropriate solution to privacy problems which include identity disclosure that can lead to personal embarrassment and abuse. Apart from technological solutions, there is a strong need to create awareness among the people regarding privacy hazards to safeguard themselves form privacy breaches. One of the serious privacy threats is smart phone. Lot of personal information in the form of contacts, messages, chats and files are being accessed by many apps running in our smart phone without our knowledge. Most of the time people do not even read the privacy statement before installing any app. Hence there is a strong need to educate people on the various vulnerabilities which can contribute to leakage of private information. We propose a novel privacy preservation model based on Data Lake concept to hold variety of data from diverse sources. Data analytics is done on the data collected from various sources. If an ecommerce site would like to perform data analytics, they need transactional data, website logs and customers opinion through social media pages. A Data lake is used to collect data from different sources. Apache Flume is used to ingest data from social media sites, website logs into Hadoop Distributed File System(HDFS). Using SQOOP relational data can be loaded into HDFS. A Hadoop map reduce job using machine learning can be executed on the data to classify the sensitive attributes. The data can be vertically distributed to separate the sensitive attributes from rest of the data and apply tokenization to map the vertically distributed data. The data without any sensitive attributes can be published for data analytics.

References

1. 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.

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

3. Yang D, Bingqing Q, Cudre-Mauroux P. Privacy-preserving social media data publishing for personalized rankingbased recommendation. IEEE Trans Knowl Data Eng. 2018. ISSN (Print):1041-4347, ISSN (Electronic):1558-2191.

4. Liu Y et al. A practical privacy-preserving data aggregation (3PDA) scheme for smart grid. IEEE Trans Ind Inf. 2018.

5. Duncan GT et al. Disclosure limitation methods and information loss for tabular data. In: Confidentiality, disclosure and data access: theory and practical applications for statistical agencies. 2001. p. 135–166.

6. Duncan GT, Diane L. Disclosure-limited data dissemination. J Am Stat Assoc. 1986;81(393):10–8.

7. Lambert Diane. Measures of disclosure risk and harm. J Off Stat. 1993;9(2):313.

8. Spiller K, et al. Data privacy: users' thoughts on quantified self personal data. Self-Tracking. Cham: Palgrave Macmillan; 2018. p. 111–24.

9. Hettig M, Kiss E, Kassel J-F, Weber S, Harbach M. Visualizing risk by example: demonstrating threats arising from android apps. In: Smith M,

editor. Symposium on usable privacy and security (SOUPS), Newcastle, UK, July 24–26, 2013.

10. Bayardo RJ, Agrawal A. Data privacy through optimal k-anonymization. In: Proceedings 21st international conference on data engineering, 2005 (ICDE 2005). Piscataway: IEEE; 2005.

11. 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.

12. 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.

13. 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.

14. Samarati, Pierangela, and Latanya Sweeney. In: Protecting privacy when disclosing information: k-anonymity and its enforcement through generalization and suppression. Technical report, SRI International, 1998.

15. Sweeney Latanya. Achieving k-anonymity privacy protection using generalization and suppression. In J Uncertain Fuzziness Knowl Based Syst. 2002;10(05):571–88.

16. Sweeney Latanya. k-Anonymity: a model for protecting privacy. Int J Uncertain, Fuzziness Knowl Based Syst. 2002;10(05):557–70.

17. Williams R. On the complexity of optimal k-anonymity. In: Proc. 23rd ACM SIGMOD-SIGACT-SIGART symp. principles of database systems (PODS). New York: ACM; 2004.

18. Machanavajjhala A et al. L-diversity: privacy beyond k-anonymity. In: Proceedings of the 22nd international conference on data engineering (ICDE'06), 2006. Piscataway: IEEE; 2006.

19. Xiao X, Yufei T. Personalized privacy preservation. In: Proceedings of the 2006 ACM SIGMOD international conference on Management of data. New York: ACM; 2006.

20. Rubner Y, Tomasi T, Guibas LJ. The earth mover's distance as a metric for image retrieval. Int J Comput Vision. 2000;40(2):99–121.

21. Aggarwal CC, Philip SY. A general survey of privacy-preserving data mining models and algorithms. Privacy-preserving data mining. Springer: US; 2008. p. 11–52.

22. 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.