ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 3, Issue 5, May 2015

Survey on Data Privacy in Big Data with K- Anonymity

Salini . S, Sreetha . V. Kumar, Neevan .R

M.Tech Student, Dept of CSE, Marian Engineering College, Trivandrum, Kerala, India

Asst. Professor, Dept of CSE, Marian Engineering College,Trivandrum,Kerala, India

Asst. Professor, Dept of CSE, College of Engineering Kottarakara, Kollam, Kerala, India

ABSTRACT: Big data concerns of large-volume, complex, growing data sets with multiple, autonomous sources. With the fast development of networking, data storage, and the data collection capacity, big data are now rapidly expanding in all science and engineering domains, including physical, biological and biomedical sciences. Data mining have challenges with big data. From all the challenges, existing systems focused on data privacy and data security. Data privacy can protect using K- Anonymity technique and data security is implemented using authentication method. K- Anonymity is the method that annonymized data fields such that sensitive information cannot be pinpointed to an individual record. But leakage of sensitive data are still there, so there must need a better privacy preserving technique. In this paper we are proposing an alternate method using Alpha K Anonymity in order to obtain better privacy.

KEYWORDS:- Randomization; Secure multiparty computation; K- Anonymity; Alpha K Anonymity

I. INTRODUCTION

Big Data concern large-volume, complex, growing data sets with multiple, autonomous sources. With the fast development of networking, data storage, and the data collection capacity, Big data are now rapidly expanding in all science and engineering domains, including physical, biological and biomedical sciences. Big data starts with large-volume, heterogeneous, autonomous sources with distributed and decentralized control, and seeks to explore complex and evolving relationships among data. Information sharing is an ultimate goal for all systems involving multiple parties, so data privacy is an important factor for data mining with big data. It cannot manage them with our current methodologies or data mining software tools. Big data mining is the capability of extracting useful information from these large datasets or streams of data, that due to its volume, variability, and velocity, it was not possible before to do it. The Big data challenge is becoming one of the most exciting opportunities for the next years. K- Anonymity is the method that anonymized data fields such that sensitive information cannot be pinpointed to an individual record. One of the major benefits of the data anonymization based information sharing approaches is that, once anonymized, data can be freely shared across different parties without involving restrictive access controls.

II. PROBLEM STATEMENT

Big data characteristics are HACE theorem. This theorem states that Big Data starts with large-volume, Heterogeneous; Autonomous sources with distributed and decentralized control, and seeks to explore Complex and Evolving relationships among data. Then the challenges of big data are classified in to three tiers and they are big data mining platform, big data semantics and inside big data semantics there are two issues information sharing and

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 3, Issue 5, May 2015

data privacy and domain and application knowledge and the tier III is big data mining algorithms. In tier III there are three steps local learning and model fusion for multiple information sources, mining from sparse uncertain, and incomplete data, mining complex and dynamic data. The challenges at [1] Tier I focus on data accessing and arithmetic computing procedures. Because Big Data are often stored at different locations and data volumes may continuously grow, an effective computing platform will have to take distributed large-scale data storage into consideration for computing. The challenges at Tier II center on semantics and domain knowledge for different Big Data applications. Such information can provide additional benefits to the mining process, as well as add technical barriers to the Big Data access (Tier I) and mining algorithms (Tier III) [1][7].

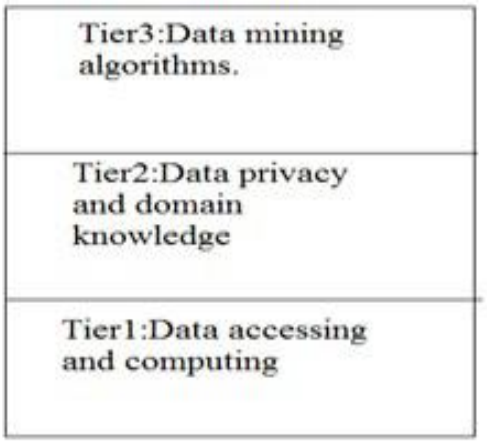

Figure 1:- Big Data Processing Framework

But implementing these all tiers are not feasible so concentrate in data privacy and also in security related to big data mining. Privacy preservation has become a major issue in many data mining applications. When a data set is released to other parties for data mining, some privacy-preserving technique is often required to reduce the possibility of identifying sensitive information about individuals. This is called the disclosure-control problem [1][7]. While the motivation for sharing is clear, a real world concern is that Big Data applications are related to sensitive information, such as banking transactions and medical records. Simple data exchanges or transmissions do not resolve privacy concerns. But public disclosure of an individual's information can have serious consequences for privacy. To protect privacy, two common approaches are to 1) Restrict access to the data, such as adding certification or access control to the data entries, so sensitive information is accessible by a limited group of users only, and 2) Anonymized data fields such that sensitive information cannot be pinpointed to an individual record.

For the first approach, common challenges are to design secured certification or access control mechanisms, such that no sensitive information can be misconduct by unauthorized individuals. For data anonymization, the main objective is to inject randomness into the data to ensure a number of privacy goals. For example, the most common k-anonymity privacy measure is to ensure that each individual in the database must be indistinguishable from k-1 others. Common anonymization approaches are to use suppression, generalization, perturbation, and permutation to generate an altered version of the data, which is, in fact, some uncertain data. Privacy relates to what data can be safely disclosed without leaking sensitive information regarding the legitimate owner. Thus, if one asks whether confidentiality is still required once data have been annonymized, the reply is yes if the anonymous data have a business value for the party owning them or the unauthorized disclosure of such anonymous data may damage the party owning the data or other parties [5].

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 3, Issue 5, May 2015

III. PRIVACY PRESERVING TECHNIQUES

A number of techniques have been developed to provide privacy to the databases such as, randomization, inference, secure multiparty computation, distributed privacy preservation, personalized privacy preservation and anonymization.

A. Randomization

The randomization method provides effective way of preventing the user from learning sensitive data which can be easily implemented because the noise added to the given record is independent from the other records. The amount of noise is large enough to smear original values, so individual record cannot be recovered. The randomization method is simple as compare to other methods because it does not require to knowledge of other records. That is why randomization can be used without the use of server that contains other records also. Large randomization increases the uncertainty and the personal privacy of the users. However, at the same time, larger randomizations can cause loss in the accuracy perturbation and randomization based approaches [4].

B. Distributed privacy preservation

The key goal in most distributed methods for privacy-preserving data mining is to allow computation of useful aggregate statistics over the entire data set without compromising the privacy of the individual data sets within the different participants. Thus, the participants may wish to collaborate in obtaining aggregate results, but may not fully trust each other in terms of the distribution of their own data sets. For this purpose, the data sets may either be horizontally partitioned or be vertically partitioned. In horizontally partitioned data sets, the individual records are spread out across multiple entities, each of which have the same set of attributes. In vertical partitioning, the individual entities may have different attributes (or views) of the same set of records. Both kinds of partitioning pose different challenges to the problem of distributed privacy preserving data mining. The problem of distributed privacy-preserving data mining overlaps closely with a field in cryptography for determining secure multi-party computations.

C. Secure Multi Party Computation

Goal of secure party computation is to compute function when each party has some input. It generally deals with problems of function computation with distributed inputs. In this protocol, parties have security properties e.g. privacy and correctness. Regarding privacy a secure protocol must not reveal any information other than output of the function. Example of such a computation is the millionaires' problem", in which two millionaires want to find out who is richer, without revealing their actual worth. Though there is difference between SMC and k-anonymity model. In k -anonymity model result can be out during the process but this is not the case in SMC, k-anonymity model protect actual value. K-anonymity and SMC are used in privacy-preserving data mining, but they are quite different in terms of efficiency, accuracy, security and privacy.

D. Inference Method

Inference is a method to subvert access control in database systems. An inference occurs when a user is able to infer some data without directly accessing them. A simple way to monitor user accesses is to examine each user query, and reject any query that accesses sensitive data. However, it is possible for a user to use a series of unsuspicious queries to infer data in the database. The interaction between the k-anonymity principle and the inference detection rules deems types of inferences harmless. Because of the static anonymity applied on the

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 3, Issue 5, May 2015

sensitive data via k-anonymity, no longer will any tuple have unique characteristics, so that unique characteristic inferences are impossible. For example if there was only one manager older than 50, his age will be statically modified, leading to one of two outcomes: either the query that requested the pair (age, salary) for all employees that are older than 50 will return at least k tuples, either it will not return any tuples.

E. Data Swapping

A related method is that of data swapping, in which the values across different records are swapped in order to perform the privacy-preservation. One advantage of this technique is that the lower order marginal totals of the data are completely preserved and are not perturbed at all. Therefore certain kinds of aggregate computations can be exactly performed without violating the privacy of the data. This technique does not follow the general principle in randomization which allows the value of a record to be perturbed independent; y of the other records. Therefore, this technique can be used in combination with other frameworks such as *k*-anonymity, as long as the swapping process is designed to preserve the definitions of privacy for that model

F. Personalized Privacy Preservation

Not all individuals or entities are equally concerned about their privacy. For example, a corporation may have very different constraints on the privacy of its records as compared to an individual. This leads to the natural problem that, they wish to treat the records in a given data set very differently for anonymization purposes. From a technical point of view, this means that the value of *k* for anonymization is not fixed but may vary with the record. A condensation based approach has been proposed for privacy-preserving data mining in the presence of variable constraints on the privacy of the data records. This technique constructs groups of non-homogeneous size from the data, such that it is guaranteed that each record lies in a group whose size is at least equal to its anonymity level. Subsequently, pseudo-data is generated from each group so as to create a synthetic data set with the same aggregate distribution as the original data. This technique assumes that an individual can specify a node of the domain generalization hierarchy in order to decide the level of anonymity that he can work with.

IV. K-ANONYMITY

A large number of privacy models were developed most of which are based on the k-anonymity property. The K-anonymity model was proposed to deal with the possibility of indirect identification of records form public databases, k-anonymity means each released record has at least (k-1) other records in the release whose values are indistinct. For example, Hospital contains large database in such a way that identity of individual cannot be revealed. It helps to reveal public databases without compromising privacy. Thus, it prevents database linkages. In k-annonymity the granularity of data representation is reduced by using techniques such as generalization and suppression. The granularity is reduced to such a level that any given record maps onto a least K other records in the dataset. A general method widely used for masking initial micro data to conform to the k-anonymity model is the generalization of the quasi identifier attributes.

One way to enable effective data mining while preserving privacy is to anonymize the data set that includes private information about subjects before being released for data mining. One way to anonymized data set is to manipulate its content so that the records adhere to k-anonymity. Two common manipulation techniques used to achieve k-anonymity of a data set are generalization and suppression. Generalization refers to replacing a value with a less specific but semantically consistent value, while suppression refers to not releasing a value at all. Generalization is more commonly applied in many domains since suppression may dramatically reduce the quality of the data mining results if not properly used.

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 3, Issue 5, May 2015

One of the major benefits of the data anonymization based information sharing approaches is that, once anonymized, data can be freely shared across different parties without involving restrictive access controls. This naturally leads to another research area namely privacy preserving data mining where multiple parties, each holding some sensitive data, are trying to achieve a common data mining goal without sharing any sensitive information inside the data[5].Similar to all other fields of security, database security uses authentication, authorization, and accounting to ensure that only authenticated users perform authorized activities at authorized points in time. Database security includes many layers of security, which can be classified in topics such as physical security, network security, encryption, and authentication. K-anonymity is one of the most important concepts in data anonymity through re-identification. Although there are many data sets available for linking to external attackers, k-anonymity does not make any assumptions regarding them. The re-identification algorithm is an implementation of k-annonymity, and is triggered after every change of the data set (insertion, deletion or update). The k factor is dynamically decided based on the number of records existing in the data ware house.

Name	Age	Job	Postcode	Cibil Value
Kashyab	30	IT-ENG	4351	630
Neerav	35	MED-DOC	4352	730
Andreia	28	MED-DOC	4353	700
Jacob	34	GOV-EMP	4354	710
Diya	25	GOV-EMP	4355	700

Table 1- Original Data Set

This table 1 contains five columns with five tuples, where name, job and postcode are quasi identifiers and cibil value is the sensitive attribute. Here a sample data set is given. So when we disclose this raw data set to the public, it is possible to identify the person and his cibil value. To avoid that identification K-anonymity is used, anonymized data set using k-annonymity is given below.

Name	Age	Job	Postcode	Cibil Value
Kashyab	30	IT-ENG	435*	630
Neerav	35	MED-DOC	435*	730
Andreia	28	MED-DOC	435*	700
Jacob	34	GOV-EMP	435*	710
Diya	25	GOV-EMP	435*	700

Table 2:- Anonymized Data Set Using K-Annonymity.

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 3, Issue 5, May 2015

In table 2 the post codes last value are anonymized, so no one can distinguish the post codes and cannot pin pointed to an individual record. K-anonymity doesn't concentrate on the relationship between the sensitive attributes so there is still leakage of sensitive data from these anonymized data set. To overcome the disadvantage we need a better anonymization technique to improve the privacy aspects.

V. CONCLUSION AND FUTURE WORK

Big Data mining is the ability of extracting constructive information from huge streams of data or datasets, that due to its variability, volume, and velocity. Data mining includes exploring and analyzing big quantity of data to locate different molds for big data. So while doing data mining in big data, data privacy is an important issue. So in this paper different types of privacy preserving techniques are discussed. But K-anonymization is not as much efficient in the area of big data mining so more sophisticated model is necessary to protect the association of individuals to sensitive information. For that Alpha K anonymity technique will be more sophisticated, trying to implement a personalized privacy preserving parallel Alpha K anonymity model in the place of K Anonymity.

REFERENCES

[1]. 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.
[2]. Providing Data Anonymity for a Secure Database Infrastructure Traian Popeea, Anca Constantinescu, Razvan Rughinis Politehnica" University of Bucharest Bucharest, Romania, 2012.
[3]. (∞, k)- Anonymity: An Enhanced k-Anonymity Model for Privacy Preserving Data Publishing" Raymond Chi Wing Wong ,Department of Computer Science and Engineering, The Chinese University of Hong Kong , Jiuyong Li and Ada WaiChee Fu, Departtment of Mathematics and Computing , The University of Southern Queensland and Ke Wang Department of Computer Science, Simon Fraser University, Canada. 2006
[4]. Generalization Based Approach to Confidential Database Updates Neha Gosai, S.H.Patil, International Journal of Engineering Research and Applications, June 2012.
[5]. A General Survey of Privacy-Preserving Data Mining Models and Algorithms, Charu C. Aggarwal Philip S. Yu, University of Illinois at Chicago.
[6]. Privacy-Preserving Updates to Confidential and Anonymous Databases" Alberto Trombetta, Wei Jiang, Elisa Bertino and Lorenzo Bossi, Purdue University 2007.
[7]. Review on Data Mining with Big Data" Vitthal Yenkar, Prof.Mahip Bartere, IJCSMC, Vol. 3, Issue. 4, April 2014.
[8]. k-Anonymization with Minimal Loss of Information", Aristides Gionis and Tamir Tassa, IEEE Transactions on knowledge and data engineering, February 2009.
[9]." K-Anonymity in the Presence of External Databases" Dimitris Sacharidis, Kyriakos Mouratidis, and Dimitris Papadias, IEEE transactions on knowledge and data engineering, March 2010.
[10]."Anonymizing Classification Data for Privacy Preservation",Benjamin C.M. Fung, Ke Wang, and Philip S. Yu, Fellow, IEEE, may 2007.
[11].A Scalable Two-Phase Top Down Specialization Approach for Data Anonymization Using Map Reduce on Cloud", Xuyun Zhang, Laurence T. Yang, Senior Member, IEEE, Chang Liu, and Jinjun Chen, Member, IEEE , February 2014.
[12].Anonymous Publication of Sensitive Transactional Data" Gabriel Ghinita, Member, IEEE,
Panos Kalnis, and Yufei Tao, IEEE transaction February 2011.
[13]. Accuracy-Constrained Privacy-Preserving Access Control Mechanism for Relational Dat
a", Za-hid Pervaiz, Walid G. Aref, Senior Member,IEEE, Arif Ghafoor, Fellow, IEEE, and Nagabhushana Prabhu, IEEE Transaction April 2014.
[14]. k-Anonymization with Minimal Loss of Information", Aristides Gionis and Tamir Tassa, IEEE Transactions on knowledge and data engineering, February 2009.

BIOGRAPHY

Ms.SALINI.S is a Mtech computer science student in Marian Engineering college Kazhakootam Trivandrum. She completed her Btech with first class in Information Technology from Kerala Universiy.

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 3, Issue 5, May 2015

Ms. SREETHA V KUMAR is working as an Assistant professor at Computer Science and Engineering department of Marian Engineering college from 2007 onwards.

Mr. NEEVAN R is working as an Assistant Professor at Computer Science and Engineering department of College of Engineering , Kottarakara under the IHRD. He completed his Mtech in Digital Image Processing from Cusat University. He Completed his Btech in Computer science and Engineering from Kerala University.