ISSN : 2347 - 8446 (Online)
ISSN : 2347 - 9817 (Print)

Security for Personal Credentials in Big Data: Through Microaggregation and TCloseness

[I]Prof. Sarita Lalchand Tanay, [II]Prof. Vivek Jaysing Nagargoje, [III]Sayali Avinash Inamdar
[I,II]PCET N.M.V.P.,Talegaon Dabhade, Pune, India
[III]Professor, Information Technology, PCET NMVPMs,
Nutan Maharashtra VidyaPolytechnic, TalegaonDabhade, Pune, India

Abstract
This survey paper is about personal security when using or publishing data online or offline for big data. The term big data has come into use recently, since, the world is becoming digitized and also the ever increasing of data in huge volumes from and of the organizations leads to storing, processing and analyzing these data, since these data are used by these organizations for business purpose. With this growing popularity and development of data mining technologies, lead to various problems and threat to the privacy of individuals sensitive and confidential information. Microaggregation is a technique of limiting the exposure of information, aiming for the privacy for released information, i.e., microdata of the subjects. It is the alternative technique for the generalization and suppression for generating k-anonymous data sets, in which the identity of each subject is made hidden within a group of k subjects in a cluster. In microaggregation, the data is perturbs and this masking allows improving data in many ways, like data granularity, reducing the impact of outliers, avoiding discretization of numerical data. K-Anonymity, alone cannot provide the protection for the data, as it provides protection against identity disclosure but prone to attribute disclosure. To solve this problem, many refinements of k-anonymity is being proposed, in which t-closeness is one providing the solution for personal privacy for information of the subjects.

Keywords
Data privacy, Microaggregation, k-anonymity, t-closeness,EMD

I. Introduction

IN As the data mining process and the knowledge discovered from it, are used by organisations, government agencies for different purposes, but with this, they are also [3]concern with the privacy and confidentiality of their information, which is easily prone to privacy threats caused by data mining process. Individual privacy may be affected by unauthorized access to personal data, which can lead a person into embarrassed situation on disclosures of this personal data. This data can be personal photos, personal information like contact number, address, email id, etc which may be used by other unauthorized person for some evil reasons. These data may be used in different way rather than the intended use which may cause threat to a person or organization. The term 'sensitive information' to refer to privileged or proprietary information that only certain people are allowed to see and that is therefore not accessible to everyone. If sensitive information is lost or used in any way other than intended, the result can be severe damage to the person or organization to which that information belongs. The term 'sensitive data' refers to data from which sensitive information can be extracted. As the data mining process and the knowledge discovered from it, are used by various organizations for their business purposes, but with this, they are also concern with the privacy and confidentiality of their information, which is easily prone to privacy threats caused by data mining process. Individual privacy may be affected by unauthorized access to personal data, which can lead a person into embarrassed situation on disclosures of this personal data. This data can be personal photos, personal information like contact number, address, email id, etc which may be used by other unauthorized person for some evil reasons. These data may be used in different way rather than the intended use which may cause threat to a person or organization. As the world is becoming digitized, the government agencies and other organizations keep on publishing data, like medical data, population, gender count, etc. for the research and other purposes like development or any other. This information is collected and saved in table format, in rows and columns, where columns heading are provided known as attributes and row contains the information related to subject according to the columns heading known as records. Each table has number of attributes, which can be divided into the following three categories:

1. Attributes that clearly identify individuals or distinguish. These are known as explicit identifiers. e.g., PAN no, account no, customer-id etc.
2. Attributes whose values when combined with other values can potentially identify an individual. These are known as quasi-identifiers. e.g., Zip code, Birth-date, and Gender.
3. Attributes that are considered sensitive for subjects, such as Disease and Salary.

II. Literature Survey

After the data mining process is performed, the microdata is released, it is necessary to prevent the sensitive information of the individuals from being disclosed or published. Two types of information disclosure have been identified in the literature

- Identity disclosure
- Attribute disclosure.

Identity disclosure occurs when an individuals identity is able to recognize the individual because of released data. Attribute disclosure occurs when new information about subjects is revealed, and this released data make it possible to recognized the characteristics of an individual more accurately than it would be possible before the data release. Identity disclosure often leads to attribute disclosure. Once there is identity disclosure, an individual is reidentified and the corresponding sensitive values are revealed. Attribute disclosure can occur with identity disclosure and also even if identity disclosure doesn't occurs. It has been recognized that even disclosure of false attribute information may cause harm and also can put an individual into bad situation. An observer of a released table may incorrectly perceive that an individual's

sensitive attribute takes a particular value and behaves accordingly based on the perception. This can harm the individual, even if the perception is incor-rect. While the released table gives useful in-formation to researchers, it presents disclosure risk to the individuals whose data are in the table, i.e, threats to personal data. Therefore, our objective is to limit the disclosure risk to an acceptable level while maximizing the benefit. This is done by performing the process of anonymizing of the data before releasing. The First step of anonymization is to remove ex-plicit identifiers. However, this is not enough, as an adversary may already know the quasi-identifier values of some individuals in the table. This knowledge gain can be either from personal knowledge (e.g., knowing a particular individual in person), or from other publicly available databases sources (e.g., a voter registration list) that include both explicit identifiers and quasi-identifiers. A common anonymization approach is generalization, which replaces quasi identifier values with values that are less-specific but semantically consistent. As a result, more records will have the same set of quasi-identifier values, which helps the intruder to identify the individual's characteristics more quickly.

III. Acheiving The Security For Informations

This is done by performing the process of anonymizing of the data before releasing. The First step of anonymization is to remove explicit identifiers. However, this is not enough, as an adversary may already know the quasi-identifier values of some individuals in the table. This knowledge gain can be either from personal knowledge (e.g., knowing a particular individual in person), or from other publicly available databases sources (e.g., a voter registration list) that include both explicit identifiers and quasi-identifiers. A common anonymization approach is generalization, which replaces quasi identifier values with values that are less-specific but semantically consistent. As a result, more records will have the same set of quasi-identifier values. Here, we define an equivalence class of an anonymized table to be a set of records that have the same val-ues for the quasidentifiers. To effectively limit disclosure, we need to measure the disclosure risk of an anonymized table. In other words, k-anonymity requires that each equivalence class contains at least k records. While k-anonymity protects against identity disclosure, it is insufficient to prevent attribute disclosure. To provide solution to this limitation of k-anonymity, t-closeness is alternative approach. T-Closeness requires that the distribution of the confidential attribute values within each group i.e, clus-ters. That is, the indistinguishable records to be distributed equally into each cluster, the distribution of the confidential attribute values in the entire data set.

IV. K-ANONYMITY

An intruder re-identifies a record from an anonymized data set is said to happen when he is able to determine the identity of the subject to whom the record corresponds. In case of re-identification, the intruder can associate the values of the QIT attributes in the reidentified record to the identity of the subject, thereby violating the subject as privacy. K-Anonymity, aims to limit ability of reidentification of the individuals record by the intruders. K-anonymity all alone is not able to provide complete security, two attacks were identi-fied the homogeneity attack and the back-ground knowledge attack[1]. Definition 1: (k-anonymity): Let T be a data set and QIT be the set of quasi-identifier attributes in it. T is said to satisfy k-anonymity if, for each combination of values of the quasi identifiers in QIT, at least k records in T share that combination. In a k-anonymous data set, no subject as identity can be linked (based on the quasi-identifiers) to less than k records. Hence, the probability of correct re-identification is, at most, 1=k. Here, we use the terms k-anonymous group or equivalence class to refer to a set of records that share the quasi-identifier values. Its limitation, even though k-anonymity protects against identity disclosure, it is already known fact that k-anonymous data sets are prone to attribute dis-closure. Attribute disclosure occurs when the variability of a confidential attribute within an equivalence class is too low i.e, the distribution of the confidential attribute is not done equally.

V. Microaggregation

Microaggregation is a family of perturbative methods for statistical disclosure control of microdata releases. The microaggregation uses dividing of informations into clusters. [1] Each cluster has equal K records from data sets. It consists of the following two steps:

1. Partition: The records in the original data set are partitioned into several clusters, each of them containing at least k records. To minimize the information loss, QIT records must be distributed equally in each cluster.
2. Aggregation: An aggregation operator is used to summarize the data in each cluster and the original records are replaced by the aggregated output.

The goal of microaggregation is to minimize the information loss.

VI. T-CLOSENESS

Even though k-anonymity protects against identity disclosure, it is known fact that k-anonymous data sets are prone to attribute disclosure. Attribute disclosure occurs when the variability of a confidential attribute within an equivalence class is too low i.e, the distribution of the confidential attribute is not done equally [1]. In this case, being able to determine the equivalence class of a subject may reveal too much information about the confidential attribute value of that subject. Several refinements of k-anonymity have been proposed to deal with attribute disclosure. For example, p-sensitive k-anonymity, l-diversity, t-closeness, and t-closeness. In this paper the focus is on t-closeness because of its strict privacy guarantee. T-Closeness seeks to limit the amount of in-formation that an intruder can obtain about the confidential attribute of any specific subject. To this end, t-closeness requires the distribution of the confidential attributes within each of the equivalence classes to be similar to their distribution in the entire data set. Definition 2: An equivalence class is said to satisfy t-closeness if the distance between the distribution of the confidential attribute in the class and the distribution of the attribute in the whole data set is no more than a threshold t. A data set (usually a k-anonymous data set) is said to satisfy t-closeness if all equivalence classes in it satisfies t-closeness.

The closeness is calculated by using EMD [8], the earth's movers distance[1].The EMD uses bins for transferring data.Here two bins are used, P,Q. The data is moved from bin P to bin Q. EMD (P,Q) Calculates the cost of transforming one distribution P into another distribution Q by moving probability mass. EMD is computed as the minimum transportation cost from the bins of P to the bins of Q, so it depends on how much mass is moved and how far it is moved. For numerical attributes the distance between two bins is based on the number of bins between them. If the numerical attribute takes values If the numerical attribute takes values v1,v2,...,vm

ISSN : 2347 - 8446 (Online)
ISSN : 2347 - 9817 (Print)

,where vi¡ vj if i ¡ j, then ordered distance (vi,vj)=(i-j)/(m-1). Now, if P and Q are distributions over v1,v2,...,vm that, respectively, assign probability pi and qi to vi, then the EMD for the ordered distance can be computed as

EMD(P,Q)= 1m 1 mi=1 j ij=1 pj - qj j

01. Let (A1,....,Am) be a microdata set
02. Records r1,...., rn
03. Attributes A1,....,Am
04. the original data set T(A1,.....,An)
05. K-anonymity is performed and T'(A1,.....,An)is generated.
06. Let k be the size of cluster microaggregation is performed to form clusters
07. P,Q are bins of equivalence class
08. The records are moved from P to Q by calculating t-closeness to threshold value t. of the original data set is released. We use the term anonymized data set to refer to T'(A1,....,An)

VII. Conclusion & Future Scope

K-anonymity protects against identity disclosure, but fails to provide sufficient protection against attribute disclosure. There are many refinements done to K-anonymity, but doesn't provides complete security. Generalization also has some drawbacks, like recoding, no suppression level is given for suppressing the records, changes the values if number formats changes. Motivated by these limitations, proposed a privacy technique called closeness. The use of micro aggregation as a method to attain k-anonymous and then t-closeness. To incorporate semantic distance, we choose to use the Earth Mover Distance measure. PPDP and PPDM provide methods to explore the utility of data while preserving privacy. However, most current studies only manage to achieve privacy preserving in a statistical sense. Considering that the definition of privacy is essentially personalized, developing methods that can support personalized privacy preserving is an important direction for the study of PPDP and PPDM.

Privacy is done only on numerical attributes but categorical attributes requires more security as compared to numbers, as it contains images, addresses, descriptions, etc.

VIII. Acknowledgment

We would like to grab the opportunity to thank to Prof. Lomesh K. Ahire , Prof. Dhaneshree Patil, for their valuable guidance.We would also like to thank our Honorable principal Prof.S.N.More sir for encouraging us. Also we are grateful to our parents and friends for all their support and encouragement.

References

[1] *Jordi Soria-Comas, Josep Domingo-Ferrer, Fellow, IEEE, David Sanchez, and Sergio Martinez, "t-Closeness through Microaggregation: Strict Privacy with Enhanced Utility Preservation, VOL. 27, and NO. 11, NOVEMBER 2015.*

[2] *L. Gayathri,R. Ranjitha, S. Thiruchadai Pandeeswari,P.T. Kanmani, "Preserving Data Privacy in Third Party Cloud Audit", Volume 4, Issue 6, December 2015.*

[3] *LEI XU, CHUNXIAO JIANG, (Member, IEEE), JIAN WANG, (Member, IEEE), JIAN YUAN, (Member,IEEE), AND YONG REN, (Member, IEEE),"Information Security in Big Data:Privacy and Data Mining",version October 20, 2014.*

[4] *Ninghui Li, Member, IEEE, Tiancheng Li, and Suresh Venkatasubramanian, "Closeness: A New Privacy Measure for Data Publishing", VOL. 22, NO. 7, JULY 2010 943.*

[5] *Jun Zhou, Zhenfu Cao, Senior Member, IEEE, Xiaolei Dong, and Xiaodong Lin, Senior Member, IEEE, "PPDM: A Privacy-Preserving Protocol for Cloud-Assisted e-Healthcare Systems,VOL. 9, NO. 7, OCTOBER 2015.*

[6] *N. Li, T. Li, and S. Venkatasubramanian,"t-Closeness: Privacy beyond k-Anonymity and -Diversity," Proc. Intl Conf. Data Eng.(ICDE), pp. 106-115, 2007.*

[7] *X. Xiao and Y. Tao, "Personalized Privacy Preservation",Proc.ACM SIGMOD, pp. 229-240, 2006.*

[8] *J. Cao, P. Karras, P. Kalnis, and K.-L. Tan, "SABRE: A sensitive attribute Bucketization and Redistribution framework for tcloseness," VLDB J., vol. 20, no. 1, pp. 5981, 2011.*

Author's Profile

Mrs.Sarita L. Tanay (Professor), Information Technology, PCET NMVPMs, Nutan Maharashtra Vidya Polytechnic, Talegaon Dabhade, Pune, India.-410507. Email: tanaysarita@gmail.com

Mr.Vivek Jaysing Nagargoje (Professor), Information Technology, PCET NMVPMs, Nutan Maharashtra Vidya Polytechnic, Talegaon Dabhade, Pune, India.-410507. Email: viv7799@gmail.com

Ms. Sayali Avinash Inamdar, (Professor), Information Technology, PCET NMVPMs, Nutan Maharashtra Vidya Polytechnic, TalegaonDabhade, Pune, India. 410507. Email: say2351995@gmail.com